



Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado em Engenharia Informática
2007/2008

Gesture Based Interface for Image Annotation

Duarte Nuno de Jesus Gonçalves

Orientador
Professor Doutor Nuno Manuel Robalo Correia

Lisboa

2008

Gesture Based Interface for Image Annotation

Dissertação apresentada para obtenção do Grau de Mestre em Engenharia Informática pela Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia.

Nome : Duarte Nuno de Jesus Gonçalves nº26181

Orientador : Professor Doutor Nuno Manuel Robalo Correia

Lisboa

2008

Acknowledgments

Completing this master's degree as well as this life cycle is truly a marathon event, and I would not have been able to complete this journey “across the oceans” without the aid and support of countless people over the past years.

I must first express my gratitude towards my supervisor, Professor Nuno Correia for his support and guidance throughout this past year. His attitude and commitment towards research is an example to follow. Secondly, I would like to thank Rui Jesus, my “co-supervisor” in my thesis for the huge help and support in my work, in all the papers I have submitted as well as the daily life in the office. It was a privilege to work with him, as his friendship and all his advices surely will not be forgotten.

I would also like to thank all the Interactive Multimedia Group colleagues who helped me in this year, for all their good advices and friendship, I am truly proud of being an “IGMer”. Filipe “recognition master” for all the help in this project as well as all the friendship and support; Nóbrega “OGRE master” for all the help in my early days and countless hours playing RON!; Sabino for all the friendship and brilliant discussions about science, life and everything else!; Cabral for putting up with my constant “hand waves”, as well for all the good advices and friendship; Rossana and Carmen, the nicest girls at IMG - I'm sorry for my Benfica discussions every single day, thank you for all your support and smiles; Guida, Daniel, Rute and Tiago, you guys (and girls) rule !

As this is the end of nine years in Faculdade de Ciência e Tecnologia, I would like to shout out a big thank you to my fellow friends and colleagues who made my life so much easier - you guys I will never forget you! For all the nights spent in “calabouço” and “anexo” I would like to thank the following: Cen (Mr. Universe) and João Ozzy – thank you so much for your friendship, you guys are the best! Miguel, Marco, Tiago, Sansão, Shelly “Malibu”, Catia, Dagmar, João “muletas”, Flip Robert. For the early days I would like to thank Rui, Roque, Orlando, Candeias, Mauro, Botelho, Evaristo, Viegas, Carlão, Hugão, Rita, Alex and my friend Xixa for all the guidance, patience, support, friendship and good times! Obviously, there were many people

whose life crossed with mine, and helped me to be a better person. Their names aren't written down, but I will never forget them.

A special thank you for all the professors who during these years made me what I am. My gratitude is with you, for all the good advices, guidance, support and friendship; I will truly miss your classes!

Needless to say, this entire thesis and all the previously work would be at most a dream were it not for my loving family and friends. To my parents and sister: Luisa, Adriano and Inês, whose love and understanding was unconditional, you are the real reason for all of this. They have given their support, even knowing that doing so contributed greatly to my absence these last years. A special thanks and love to Marta Trigueiro, who is one of the most wonderful human beings I ever met – half of this work is yours! Thank you for all the love and invaluable emotional support, understanding and missed hours – you will always make me smile. For all family and friends who carry me throughout these past years, for all the love, friendship and innumerable laughs, thank you so much, you brought me peace of mind and happiness in my life.

A humble “thank you” - it is been said that the way to the light is through the darkness, and I just thank You that light shines in darkness, and darkness can't stop it.

At last but surely not least, I would like to dedicate this work to my grandmother Leopoldina de Jesus, who recently past way. With her smile, nurture and dedication, she made my life better in every way, everyday, as I am truly a proud grandson. May your peace and love guide my life.

Resumo

Dada a complexidade da informação visual, a pesquisa de imagens em bases de dados multimédia apresenta maiores dificuldades do que a pesquisa de informação textual. Esta complexidade está relacionada com a dificuldade em anotar automaticamente uma imagem ou um vídeo com palavras-chave que descrevam o seu conteúdo. Em geral, esta anotação é realizada manualmente (e.g., Google Image) e a pesquisa é baseada em palavras anotadas. Contudo, esta tarefa requer tempo disponível e é aborrecida.

Esta dissertação propõe-se definir e implementar um jogo para anotar fotografias pessoais em formato digital de forma semi-automática. O motor do jogo classifica imagens de forma automática, sendo o papel do jogador a correcção destes erros de anotação. A aplicação é constituída pelos seguintes módulos principais: um módulo de anotação automática de imagens, um módulo destinado à interface gráfica do jogo (mostra imagens e palavras ao utilizador), um módulo destinado ao motor de jogo e um módulo para a interacção. A interacção é feita usando um conjunto pré-definido de gestos para uma câmara. Estes gestos são reconhecidos usando técnicas de processamento de imagem e vídeo e interpretados como jogadas do utilizador. Esta dissertação apresenta uma análise detalhada da aplicação, módulos computacionais e design, assim como uma série de testes de usabilidade.

Palavras-chave : Anotação semi-automática de imagens; Recuperação de imagens; Interface gestual; Computação humana; Interacção pessoa-máquina

Abstract

Given the complexity of visual information, multimedia content search presents more problems than textual search. This level of complexity is related with the difficulty of doing automatic image and video tagging, using a set of keywords to describe the content. Generally, this annotation is performed manually (e.g., Google Image) and the search is based on pre-defined keywords. However, this task takes time and can be dull.

In this dissertation project the objective is to define and implement a game to annotate personal digital photos with a semi-automatic system. The game engine tags images automatically and the player role is to contribute with correct annotations. The application is composed by the following main modules: a module for automatic image annotation, a module that manages the game graphical interface (showing images and tags), a module for the game engine and a module for human interaction. The interaction is made with a pre-defined set of gestures, using a web camera. These gestures will be detected using computer vision techniques interpreted as the user actions. The dissertation also presents a detailed analysis of this application, computational modules and design, as well as a series of usability tests.

Keywords: Semi-Automatic image annotation; Image retrieval; Gesture interface; Human computation; Human computer-interaction

Acronyms

This section contains the list of acronyms used in the thesis.

MIR	Multimedia Information Retrieval
CBIR	Content Based Image Retrieval
URL	Uniform Resource Locator
MSM	Manhattan Story Mashup
HCI	Human Computer Interaction
OGRE	Object-oriented Graphics Rendering Engine
OpenCV	Open Source Computer Vision
XML	Extensible Markup Language
DOM	Document Object Model
SAX	Simple API for XML

Table of contents

Chapter 1	Introduction.....	1
1.1	Image annotation and CBIR systems.....	2
1.2	Automatic annotation vs. manual annotation.....	3
1.2.1	Human computation.....	4
1.3	Solution presented.....	5
1.4	Main contributions and objectives.....	6
1.4.1	Publications.....	7
1.5	Organization.....	8
Chapter 2	Related work.....	9
2.1	Automatic image annotation.....	10
2.1.1	ALIPR - Automatic photo tagging and virtual image search.....	10
2.1.2	Hierarchical classification for automatic image annotation.....	12
2.1.3	AnnoSearch - Image auto-annotation by search.....	14
2.2	Interfaces for image annotation.....	17
2.2.1	Games with a purpose - ESP game and peekaboom.....	17
2.2.2	Label it with LabelMe.....	20
2.2.3	Manhattan story mashup.....	21
2.2.4	Flickr – A public image-sharing tool.....	23
2.3	Semi-automatic image annotation.....	24
2.3.1	Semi-Automatic image annotation using relevance feedback.....	24
2.3.2	A Semi-Automatic image annotation using frequent keyword mining.....	25
2.4	Gestural interfaces.....	27
2.4.1	Having fun while using gestures – Eye Toy experience.....	27
2.4.2	Developing games with Magic Playground.....	28
Chapter 3	Concept and architecture.....	31
3.1	Main concepts.....	31
3.2	System overview.....	32
Chapter 4	Tag-Around.....	35
4.1	Human interaction and interface.....	35
4.1.1	Gesture user interface.....	35

4.1.2	Perceptual user interface	36
4.2	Game application.....	36
4.2.1	Game interface	37
4.2.2	Game engine.....	40
4.2.3	Motion detection	43
4.2.4	Face recognition	44
4.3	Automatic image annotation	45
4.3.1	Automatic Image Annotation	45
4.3.2	Updating the parameters.....	45
4.4	Implementation.....	46
4.4.1	Technology.....	46
4.4.2	Application modules	47
Chapter 5	Interface design	49
5.1	Paper prototype	50
5.2	Usability tests	51
5.2.1	Participants.....	51
5.2.2	Setup and methodology.....	51
5.2.3	Questionnaire	52
5.2.4	Results	52
Chapter 6	Conclusions and future work.....	57
6.1	Alternative human interaction and design interfaces	58
6.2	Future work scenarios	58
Chapter 7	References	61

List of figures

Figure 2.1 - ALIPR results on several different photos	11
Figure 2.2 - Some of ALIPR common mistakes using uploaded photos	11
Figure 2.3 - The flowchart for bridging the semantic gap hierarchically.....	13
Figure 2.4 - Two different views of hyperbolic visualization of large-scale concept ontology.....	14
Figure 2.5 - AnnoSearch system’s framework.....	15
Figure 2.6 - Output examples from the AnnoSearch system	16
Figure 2.7 - Typical image search results presented in [14]	16
Figure 2.8 - The online ESP game	18
Figure 2.9 - Peekaboom online game.....	19
Figure 2.10 - LabelMe website tool	20
Figure 2.11 - The SMS web tool	22
Figure 2.12 - The SMS mobile client application	22
Figure 2.13 - Flickr website	23
Figure 2.14 - User interface framework scenario.....	25
Figure 2.15 - Experimental results using CorelDraw images	26
Figure 2.16 – Users playing EyeToy.....	28
Figure 2.17 – Magic Tetris.....	29
Figure 3.1 – System overview	32
Figure 4.1 – Player login interface	36
Figure 4.2 – Game application main modules.....	37
Figure 4.3 – Highscores layout.....	38
Figure 4.4 – Initial menu layout	39
Figure 4.5 – Game interface	39
Figure 4.6 - Score and player confidence evolution.....	43
Figure 4.7 - Tag Around main modules	47
Figure 5.1 - Paper prototype.....	50
Figure 5.2 – “It was easy to learn how to use the application” question.....	53
Figure 5.3 – “would you use the application to have fun with family and friends” question	55
Figure 5.4 – General comments about Tag Around	56

Chapter 1

Introduction

The Internet is growing at a pace never seen before. Currently, millions of people are exchanging information across networks spread around the world. Blogs, websites and portals are being created, visited and changed every second. As a collaborative society, we have the urge to organize, index and share our information, for example using Multimedia Information Retrieval (MIR) systems. Since the last century, we changed our habits and almost stopped visiting libraries and other physical places for retrieval purposes, and started to use MIR systems. This is becoming an essential channel for research and entertainment, as there are portals like Google, Wikipedia, and even the internet-based virtual world Second Life, and more and more of our professional (as well as social) lives depend on its existence.

With this need for information sharing and organization came the concept of folksonomy, which is part of the Web 2.0 proposals. This is a [43] “*trend in web design and development — a perceived second generation of web-based communities and hosted services (such as social-networking sites, wikis, blogs, (...)) which aim to facilitate creativity, collaboration, and sharing between users.*”. This is a concept derived from the fact that regular people organize content when using and sharing their documents using keywords to describe what is in their images, videos, or text files. Web-based social tagging systems like Flickr [37] and Del.icio.us [34] allow users to annotate a resource, such as a web page or an image, with a chosen set of keywords

commonly known as “tags”. This “tagging” system can bring a set of advantages to users. A link to the resource is saved in the user’s account, and can be retrieved from any web-connected device by using any of the tags used to describe the resource. This can improve the social aspect of the web - bringing together people with similar interests, as well as improve media search and browsing systems that are still behind in terms of accuracy.

1.1 Image annotation and CBIR systems

The digital camera era has given everyone the opportunity to capture the world in pictures, and therefore the possibility to share them with others. Today we can easily generate thousands of images with content as diverse as family reunions and holiday visits. The low-cost storage and easy Web hosting has triggered the changes from passive consumers of photography to active producers. Today, searchable image data exists with extremely diverse visual and semantic content, spreading geographically throughout different locations and swiftly growing in size. All these factors have created a huge amount of possibilities and thus opportunities for real-world image search system designers and engineers.

For this matter, CBIR systems assume an important part in handling this problem. CBIR stands for Content-Based Image Retrieval and it is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases or in this case, large image servers spread throughout the Web. Generally speaking, multimedia information retrieval refers to a set of proposals, algorithms and systems that aim at extracting pertinent descriptors or metadata related to multimedia content and allowing search, retrieval, and other user level functions. Image retrieval can be based on different levels, regarding several different features. We can separate these levels in low-level visual features (such as color[18], texture [2], and shape [20]), high-level semantics [4], or both [33].

For low-level features, CBIR systems can perform image search with good accuracy while for high-level features, CBIR systems have proven to be unsatisfactory. For many years, CBIR systems used pre-annotated sets of images, from image repositories like COREL, with fixed size and orientation, and with few objects per image, which facilitated the search. This accuracy drops

exponentially if we address this problem in terms of Internet repositories, because images appear in different sizes, orientations and with multiple objects. There are even cases of those images that are blurred or missing some parts. As result of these problems, automatic image retrieval for high-level features is still challenging due to the difficulty in object recognition and image understanding. There is an urgent need to build image retrieval systems, which support high-level (semantics-based) querying and browsing of images.

To explain the distance between low-level features and high-level features, we can summarily introduce the concept of the semantic gap. The semantic gap is the lack of correlation between the information that one can extract from the visual data and the interpretation that the same data has for a user in a given situation. To somehow overcome this semantic gap, studies turned to the interacting user. Interaction of different users with a data set has been studied most systematically in categorical information retrieval [16]. The techniques reported in [16] need rethinking when used for image retrieval as the meaning of an image, due to the semantic gap, can only be defined in context. Image retrieval requires active participation of the user to a much higher degree than required by categorized querying. In content-based image retrieval, interaction is a complex interplay between the user, the images, and their semantic interpretations.

In terms of information access, retrieval systems use several different approaches, like query by example (QBE) [15], retrieval through semantic indexing, interactive retrieval and personalized and adaptive content delivery. A more detailed discussion and analysis can be found in [8].

More studies and approaches are being made in different fields, in an attempt to solve the fundamental open problem of image comprehension, such as computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics, and psychology contributing and becoming a part of the CBIR community, as described in [30].

1.2 Automatic annotation vs. manual annotation

As discussed before, the traditional way to assign metadata to a digital image is through CBIR systems. One of the major disadvantages of automatic image annotation versus manual image

annotation is that in manual annotation users can more naturally specify queries. These CBIR systems usually use previous tags assigned by users or even the text surrounding the images to perform automatic annotation.

The systems that use computer vision techniques to organize content in a server or in a database are called automatic annotation systems. They usually do not depend on humans to perform annotations, but, as mentioned, they are not always correct in terms of accuracy, because they lack the human perception and intuition (semantic gap).

Manual annotation is becoming more a subject of interest and is another approach regarding effective image annotation, as many systems and interfaces are being developed to provide humans a more effective way to perform their annotations.

1.2.1 Human computation

As pointed earlier, one of the solutions to overcome the semantic gap is to develop systems and applications that use manual annotation, therefore providing a more accurate set of results in terms of semantic significance. These results are of great importance to MIR systems in general and in CBIR systems in particular. For that reason, research is focusing more and more on the human side of this open problem.

Human computation can be described as a technique when a computational process performs its function by outsourcing certain steps to humans [13]. This approach leverages differences in abilities and alternative costs between humans and computer agents to achieve symbiotic human-computer interaction. As stated in [5, 13, 28], combining human computation skills to annotate image is a way for improving CBIR systems. Each year, people all over the world spend billions of hours playing computer games or visiting social networks around the web. To channel that energy and time to help solving large-scale problems (problems that computers are still unable to resolve) some ideas have been developed [28]. If we address the problem on image annotation in terms of human computation, we can infer that humans can annotate images with much more accuracy, bringing the low-level and the high-level (semantics) close together.

Using humans to annotate images is, as stated, one of the solutions for image annotation, but it brings some constraints. It is relatively enjoyable to take pictures but sitting at home or at the office writing tags to describe them is a tedious activity [31]. There is a lack of direct motivation in manual image tagging and the entertainment aspect is missing. The concept of Games with a Purpose [27] changed this status by introducing the idea of using the human computational ability to perform image labeling with a computer game. There are some other applications (see chapter 2) that introduce this concept - using web applications for users to upload, share and organize visual content with the possibility of tagging. There are systems that combine the web, mobile phones, and even public displays for manual annotation purposes. These kind of approaches brought new proposals to the image annotation process, but constrain the experience in the following ways:

- There are limits to the type of audience - considerable technological skill is required.
- They do not explore the situations where people have idle time (e.g., airports, bus stops or hospitals).
- They usually do not use automatic image annotation mechanisms, based on content, to help the manual annotation.

After analyzing the constraints above, this thesis presents a solution to overcome several issues regarding automatic as well as manual image annotation. The next section describes the proposed solution.

1.3 Solution presented

There are two major approaches regarding image annotation - the automatic and the manual, both of them with positive and negative aspects (automatic annotation lacks accuracy when focusing on the high-level features retrieval, while manual annotation lacks human direct motivation).

In the attempt to overcome these negative aspects, a proposal for image annotation is presented. Tag Around is a gesture based image annotation game that addresses the constraints above. It consists of a combination of manual and automatic image annotation, with interaction by means of gesture signs in front of a camera. It is a three dimensional game, where people move and

match tags and images, using a motion detection algorithm applied to the captured (user) image from a camera. A face recognition module for user login was also integrated in the Tag Around application.

In Tag Around, a user is in front of a camera and interacts with the interface using gestures. The user image is displayed in the screen, along with a set of images and tags. When playing the game and using hand movements, the user can rotate images and tags in order to pair them up, to receive a set of points. If the user matches a tag-image pair, the game engine will verify (based on the automatic algorithm, the user confidence and the group feedback) if it is a good or a bad annotation. After this, a score is attributed to the user, and if he performs a good annotation the user confidence is also incremented. The main goal is to the user match as many images and tags as he can, given a time window. After the player energy (that is related with his good and bad annotations as well as the game time) runs out, the game ends.

This game is an application of the Memoria project [11]. The main goal of this project is to design applications for accessing and retrieving personal media (images and videos). Currently the project includes a mobile user interface, a PC user interface, the Tag Around application and a multimedia retrieval system that supports all the interfaces. The idea of designing a game for image tagging was motivated by the experiments and discussions during the development of the ongoing Memoria project [11] and inspired by [27].

The next section describes in detail all the main contributions and objectives as well as the papers already submitted and accepted during the course of this dissertation.

1.4 Main contributions and objectives

After analyzing the issues regarding MIR systems, as well as CBIR systems, automatic, manual and semi-automatic annotation systems, the following set of contributions and objectives were identified:

- Understand the motivation as well as problems, constraints and current development status regarding CBIR systems.

- Gain theoretical knowledge regarding manual and automatic image annotation systems.
- Understand the importance of human computation in image annotation and CBIR systems.
- Understand the contributions of psychology to games, human motivation and entertainment.
- Study computer-human interaction methods regarding gesture-based interfaces, including its technologies and approaches.
- Develop a conceptual game model applied to image annotation.

Based on the previous knowledge and background, build a game based system that involves:

- Developing a novel interface for image annotation based on tasks vs. scoring.
- Creating a bridge between technology used on automatic annotation systems and manual systems, to overcome the semantic gap.
- Integrating the interface with motion detection and gesture detection and recognition algorithms.

1.4.1 Publications

During the course of this dissertation, efforts were made to validate the results and the work made in this project. It is understood that publications in the most important conferences of this field would help to validate ideas and disseminate the results.

The following were the papers published so far:

Gonçalves, D., Jesus, R., Grangeiro, F. e Correia, N. 2008. Tag Around – Interface Gestual para Anotação de Imagens. *Interacção 2008 – 3ª Conferência Interacção Pessoa-Máquina.*

Jesus, R., Gonçalves, D., Abrantes, A., Correia, N., *Playing Games as a Way to Improve Automatic Image Annotation*, Proceedings of IEEE International Workshop on Semantic Learning Applications in Multimedia (SLAM08), in conjunction with CVPR08 (2008).

Gonçalves, D., Jesus, R., and Correia, N. 2008. *A gesture based game for image tagging*. In CHI '08 Extended Abstracts on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 2685-2690.

Another paper has been submitted and is pending for approval :

International Conference on Advances in Computer Entertainment Technology (ACE 2008)

1.5 Organization

This document is organized in the following chapters:

Chapter 1 – Introduction, motivations as well as an overview of the literature regarding image annotation. Project presentation and summary. Introduction to the problem and its context. Objectives and practical contributions.

Chapter 2 - State of the art, a survey on the most important concepts and projects in the field of this project.

Chapter 3 – Describes the proposal for the developed game. It includes a system overview as well as of the main components.

Chapter 4 – Specifications of the application. It includes a description of the interface, the game engine as well as the automatic annotation system. An overview of all the technology used is also presented in this chapter.

Chapter 5 – A complete description of the interface design and evaluation. A paper prototype is presented as well as usability tests and results.

Chapter 6 – A brief reflection regarding the work done in the scope of the thesis and ideas for future work.

Chapter 2

Related work

This chapter includes some of the most important research topics in the fields of human computation, CBIR systems, automatic tagging systems and general development regarding multimedia information retrieval (MIR), that are related with the work done in this thesis.

Content-based image retrieval (CBIR) is a technology that helps humans to organize digital image collections by their visual content. Searching in digital repositories began years ago when humans started to move from printed documents to server databases and using Internet based systems for primary content storing. Many contributions in the field of MIR have roots in areas such as artificial intelligence, optimization theory, computational vision and psychology. Character and face recognition were some of the first case studies in this area, and researchers used essentially pure image similarity algorithms to perform searches. These concepts were experimented in the Internet, as several systems used it including Webseer (1996) [22] and Webseek (1997) [21]. The next step in this area has to do with the need to understand the semantics of a query, and not just its lower level computational representation. Content comparison techniques based purely on shape, texture and color were not enough in terms of what an image means semantically. Some projects are now being designed to help image classification by combining computers and humans to perform correct image annotation.

Tagging can be described as attributing semantic properties to an image, using human innate ability to associate images and thoughts, by means of keywords. For example, a computer algorithm can produce similarities between dog images, but if those same dogs appeared on a movie, it is difficult for a computer to make that resemblance based purely on algorithms.

With that in mind, we can consider three major approaches regarding image annotation techniques: automatic annotation systems, semi-automatic annotation systems, and manual annotation systems, all of them extremely relevant in this field of research. While automatic tagging algorithms depend almost entirely on CBIR algorithms to perform annotations using pre-annotated sets of images for comparison, semi automatic systems tend to use user feedback interaction to improve annotations, and minimize the gap between visual features and semantic content. Manual tagging systems are now trying to involve all human abilities in image annotation, to overcome the lack of annotated image databases.

The next section presents the most important projects and their objectives considering the techniques described above.

2.1 Automatic image annotation

This subsection presents some of the work involving automatic image annotation. It presents applications that use CBIR algorithms to perform annotations with different approaches.

2.1.1 ALIPR - Automatic photo tagging and virtual image search

Using advanced statistical modeling and optimization techniques, ALIPR [14] presents a system that can be trained for hundreds of semantic concepts using example pictures from each concept. It provides an automatic tagging system, using labels to describe image content and it is a solution for people that do not want to manually tag their images. Although their vocabulary is somehow limited (they only use about 322 words out of the English dictionary), results using over 5400 general-purpose photographs show that the system can automatically annotate images in real-time and provide more than 98% images with at least one correct annotation out of the top 15 selected words. The highest ranked annotation word for each image is accurate with a rate above 51%.

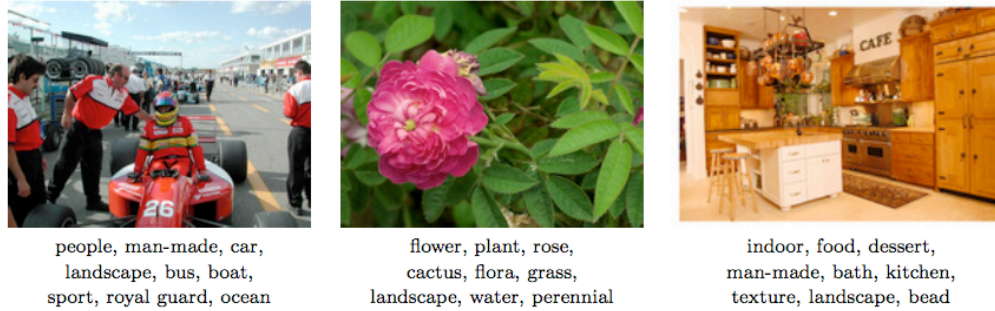


Figure 2.1 - ALIPR results on several different photos

An online demonstration is available at <http://alipr.com> and users can upload and get their images annotated by the system, if they provide an URL.

This is a system that relies on previously annotated images to perform future correct annotations, establishing a probabilistic set of associations between images and words. For that, this system is designed to achieve real-time annotation results as well as optimization properties while preserving the architectural advantages from other general modeling approaches. Real-time results are possible because when an image with new concepts is added, ALIPR only needs to learn from the new images, while previous concepts are stored in the form of profiling models.

This is a sound approach regarding automatic annotation systems. However, there are some issues like the few labels available - only 322 English words - and the errors regarding the 15 labels suggested by the system (in spite of one or two generally correct annotations).

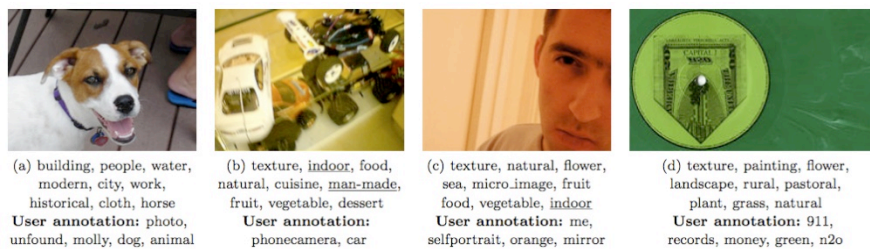


Figure 2.2 - Some of ALIPR common mistakes using uploaded photos

2.1.2 Hierarchical classification for automatic image annotation

One of the major problems regarding automatic image classification systems, as discussed before, is the semantic gap between low-level computable visual features and the user information. To minimize this gap a novel algorithm for automatic multi-level image annotation using hierarchical classification [10] was developed.

The goal of this project is to simultaneously learn a set of classifiers for large amount of image concepts with huge within-concept visual diversities and inter-concept visual similarities. To do that a structure was developed to provide automatically multi-level image annotation, reducing the semantic gap presented before to four smaller gaps. To accomplish the proposed objective, three different approaches were considered in this work:

Multi-modal boosting algorithm - This algorithm is used to understand relations between atomic image concepts and co-appearances of salient objects in those images. This was used to handle a huge diversity of within-concept visual properties, and to select the most significant features and the most suitable kernel functions for each atomic image concept.

Hierarchical boosting algorithm - This scheme is used to perform hierarchical image classification and to avoid inter-level error transmission (with automatic error recovery), outperforming the traditional techniques such as multi-class and multi-task boosting.

Hyperbolic visualization framework - This framework is used to smoothly bridge the gap between computable image concepts and the user real information needs. This framework also ensures a new approach on enabling intuitive query specification and similarity-based evaluation of large amounts of returned images.

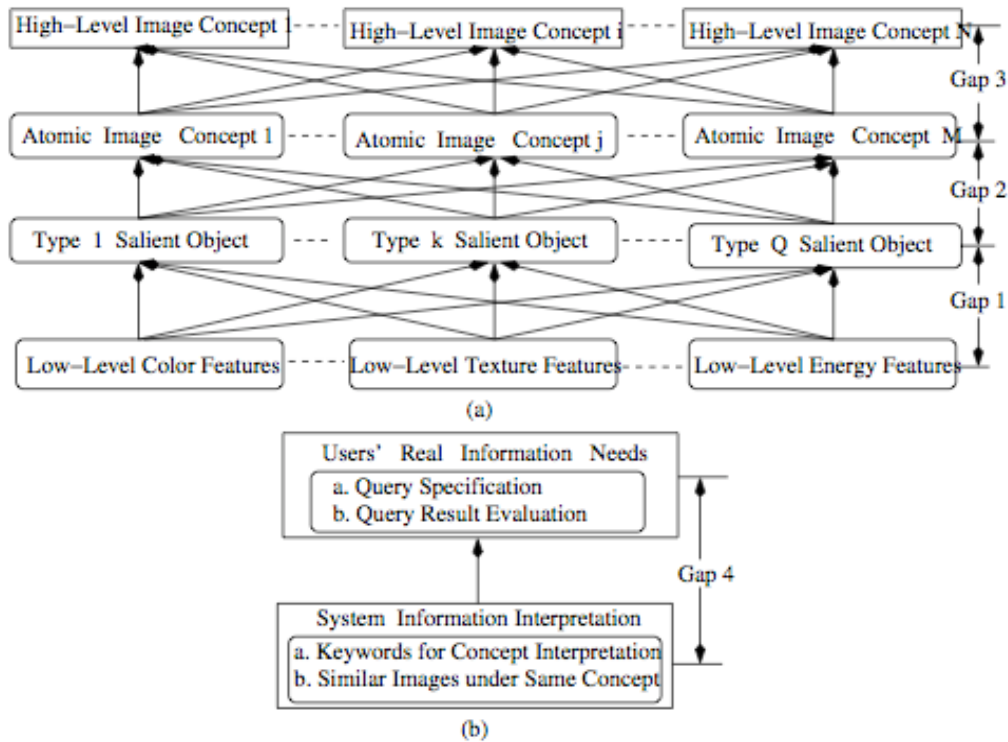


Figure 2.3 - The flowchart for bridging the semantic gap hierarchically

To ensure that an image is classified using relevant concepts at different semantic levels, the authors propose a new scheme by introducing an architectural ontology for image concept organization as well as for hierarchical image classifier training and visualization of large-scale image collections. This ontology is built using a hierarchical network, where each node (defined as concept node) represents either a concept from an image or a specific salient object class.

To illustrate this architecture, the LabelMe (an interface for concrete image object labeling) Web tool is used. LabelMe annotates words regarding particular objects in an image, but lacks explicit labels at the image concept levels (an image containing a car could not be car related, being the

car merely an object in corner of that image), which leads to a lower level of connectivity between semantic related images. This information is filtered by removing uninformative words, and using LSA (Latent Semantic Analysis) to group (and extract) the most important words regarding an image. This is used to further integrate both contextual and logical relationships concepts to perform new measurement. Using the results to construct an ontology grid, the keywords for interpreting the relevant image concepts at the higher semantic levels can be propagated automatically, reducing the hand-labeling cost significantly.

With this novel approach and using a novel hyperbolic visualization framework, an intuitive query specification and similarity-based evaluation of large amounts of returned images is provided.



Figure 2.4 - Two different views of hyperbolic visualization of large-scale concept ontology

2.1.3 AnnoSearch - Image auto-annotation by search

Despite all of the research made in this field, image annotation is still far from practical everyday use. AnnoSearch [32] brings a novel approach to this problem, because it uses data mining technologies to improve automatic image annotation.

The approach is to resolve several issues regarding traditional computer vision approaches, such as supervised learning process and the few existent presented on the previous system (ALIPR). Those systems use the Corel Stock-Style database that has well-organized images, with clean descriptions regarding the semantic concepts.

AnnoSearch methodology can be divided in two important steps. Firstly, at least one reasonably accurate keyword is required to enable text-based search for a group of semantically similar images, which can be a problem. This also happens with desktop photo search, where users normally provide the location for the images, or with web image search tools, where users can choose an image and use one of the surrounding keywords as the query keyword.

The second step is accomplished by mining the annotations from the image descriptions like titles, URLs and surrounding text - this is also the way Google image search is performed. To do this, high dimensional visual features are mapped to hash codes, which significantly speed up the content-based search process.

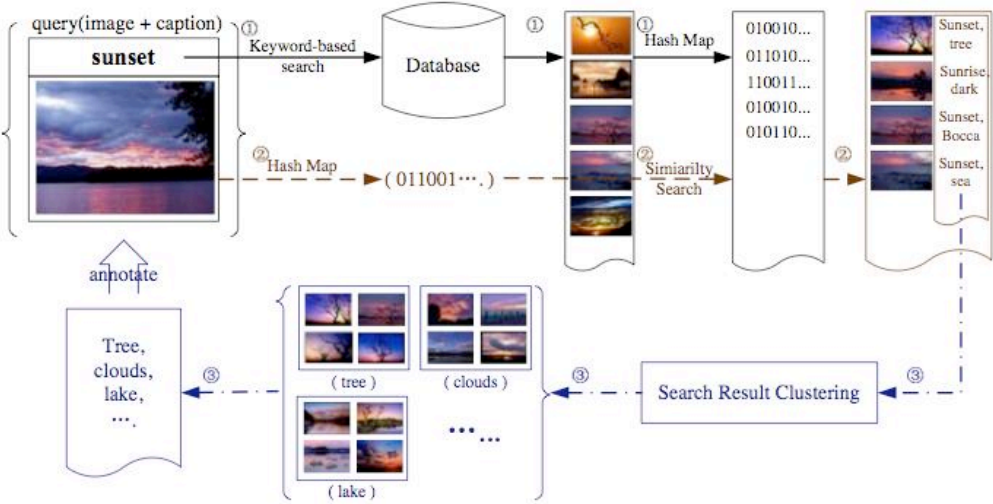


Figure 2.5 - AnnoSearch system's framework

With these two steps (that are not as complex as ALIPR or [10]), the AnnoSearch system seems to avoid all the stated disadvantages. It handles highly scalable vocabulary and is entirely unsupervised. Their future work resides on resolving the problem of how to annotate query images without any associated keywords (removing the first step of the process).














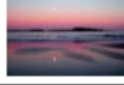






	Paris Las vegas, effel tower, love paris		Paris Sacre coeur, paris building, effel tower		Paris Eiffel tower, france, sky, paris nights		Clouds Dark clouds, sun, sky, sunrise, morn
	Sunset Lake, tree, mountain, sky, beautiful, water		Tiger Whiter tiger, usa, zoo		Tree House, flower, snow, sky, tree trunk		Clouds National park, europe, south america, blue sky
	Apple Studio, kitchen, fruit, color		Apple Fruit, apple tree		Butterfly Flower, butterfly house, beautiful butterfly		Beach South america, beautiful beach, beach house
	Clownfish Anemone, reef, red sea		Beach Sky, island, sun beach, sunrise, beach island		Butterfly Yellow butterfly, swallowtail, nature		Liberty York, liberty statue, sun
	Campus college, campus life, center, tree		Football stadium, school football, football game, football player		Iran mashhad, kish island, esfahan		Cannon Beach haystack rock

Figure 2.6 - Output examples from the AnnoSearch system

In this section, different automatic annotation systems were described. These systems were chosen because they use different methods and algorithms, while having the same objective. It can be argued that this kind of technology can still be improved significantly, as the computer capacity to recognize human semantic concepts is difficult. For example, the next two examples, are typical cases of semantic concepts that are hard to identify:

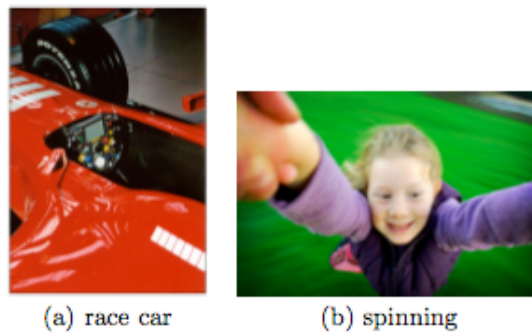


Figure 2.7 - Typical image search results presented in [14]

Automatic annotation of images with a large number of concepts is extremely challenging, because humans use a lot of intuitive background and subjectivity when they interpret an image. The two pictures could be easily identified by humans as being a race car (given the shape and

color of the model), and a girl spinning (taking in account the fuzzy background and the position of her arms).

2.2 Interfaces for image annotation

CBIR systems usually depend on pre-annotated image databases. As mentioned, databases such as CorelDraw or freefoto.com are used to test CBIR algorithms. One of the problems associated with this type of databases is the kind of images they include, because they have good quality, meaning that they are a bit different from pictures taking by casual users. People usually do not care much about centering objects or building, or if there is more than one object in the image.

To increase the amount of digital images annotated with keywords, rather than have the same pre-treated set of images, manual annotation is becoming more and more a subject of interest. This is a concept that can help researchers in the field of image-retrieval to improve results. The real issue regarding manual annotation is that it is a dull job. No one likes tagging thousands of images for hours, and so the quest for new ways to make people annotate pictures began. Using appealing interfaces that provide entertainment or simply a place to share photos (with the tagging feature included) is a way to help overcome this issue. These next applications are typical examples of that approach.

2.2.1 Games with a purpose - ESP game and peekaboom

People spend millions of hours playing computer games each year. To channel this energy and time into helping computers to tag images and detect objects is the approach presented in [28]. People, without really knowing, can help computers to solve large-scale problems – using their innate abilities to associate images and concepts.

One of the problems associated with using these human skills is the lack of human motivation, because unlike computers we require incentive to perform any task. With the increasing amount of people playing online games, they are actually a good way for people to participate in the process. The concept of Games with a Purpose [28], brought humans and computers together in a problem solving architecture – the human brain is the real processor while computers are merely the tools to get the job done.

These human computational abilities are useful because it has been noticed that there is a considered lack of accuracy when performing image queries. The reason for this is that for engines to track down, let us say “Dog” images, they perform textual searches in websites that have pictures. People do not always describe or even label their pictures while posting them in blogs or websites. While a search engine could miss a dog picture in a website, just because that image is labeled as *image1.jpg*, a person looking at that picture could instantly describe it as being, at least, dog related.

The ESP Game (www.espgame.org) handles these issues in a way that users can have fun while tagging images. The ESP Game is an online game that pairs up two unrelated players in cooperation while tagging images. They have a time limit, and within that limit, images appear in the screen and they have to describe it. If they agree in a word an amount of points will be attributed to them and the next image will appear. The objective is for both players to agree on as many tags as they can, and therefore obtain a maximum amount of points.



Figure 2.8 - The online ESP game

Because both users do not know who they are playing with, when a word is “agreed”, the chances that the word is semantically attached to the image increases considerably and therefore that word becomes attached to the image. There are some extra features to this game such as taboo words - words that have been previously agreed by other players, and cannot be typed. This method prevents images to be tagged with only a few labels, but rather with dozens of different words. The results became obvious [28] “The ESP Game is extremely popular, with many people

playing more than 40 hours per week. Within a few months of initial deployment on 25 October 2003, the game collected more than 10 million image labels; if hosted on a major site like MSN Games or Yahoo! Games, all images on the Web could be labeled in a matter of weeks.”

Peekaboom [29] (<http://www.peekaboom.org/>) is another game included in the Games with a Purpose category, and its goal is data to improve data collection on specific objects. While the ESP Game tags words in images, those words are attached to the entire image and not to objects inside that specific image.

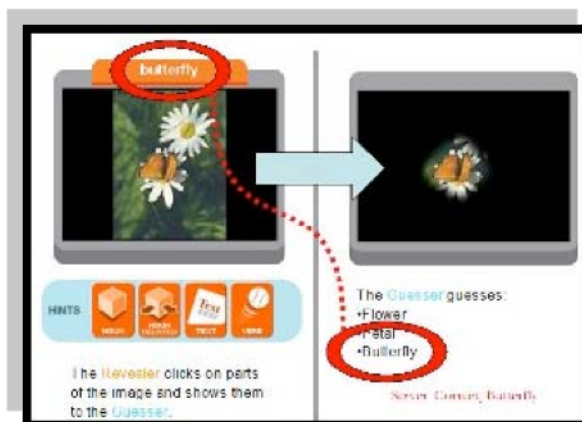


Figure 2.9 - Peekaboom online game

This is also a two player cooperative game, where one of the players is “peeking” and the other “booming”. The booming player (Boom) receives an image along with a word related to that image, and the peeking player (Peek) gets no image. Booming consists of clicking parts of the image and when Boom clicks a part of the image, it is revealed to Peek. The object of the game is for Peek to type the word associated to that part of the image. There are also a couple of extra features like hints – the booming player can tell the peeking player if he is hot or cold, depending on the word he wrote to describe the portion of the image displayed on his screen.

Other recent games have been developed with this concept such as Phetch [40], which annotates images with descriptive paragraphs, and Verbosity [42], which collects common sense facts to train reasoning algorithms.

2.2.2 Label it with LabelMe

Computer vision researchers need huge amounts of image information and content. For a long time, researchers used constrained data for automatic annotation training. Specific images for specific data analysis were good for some areas, but for new concepts and algorithms it was necessary to collect additional data to solve the problems. LabelMe [19] (<http://labelme.csail.mit.edu/>) was created to solve the difficulty of not having enough data available.

LabelMe is a web tool that allows anonymous and registered users to “discover” objects inside images, with more accuracy than Peekaboom. The reason for that to happen is because the users actually redraw the image objects by inserting bounding boxes around them. They have a toolbox in the screen that allows users to point those objects, using polygon boxes. Some of the new aspects introduced with LabelMe were that firstly LabelMe was design for the recognition of a class of objects instead of single instances of an object. This helps because traditional datasets can contain images of cars, each of the same dimensions and orientation, where LabelMe contains images of cars in multiple sizes and orientations. Secondly, it was designed for random image scenes rather than cropped and resized ones that contain one single object. Another improvement came with specific object oriented labels in a single image, using bounding boxes containing the objects. LabelMe also ensures non-copyrighted images for most cases and allows extra additions to the annotations.

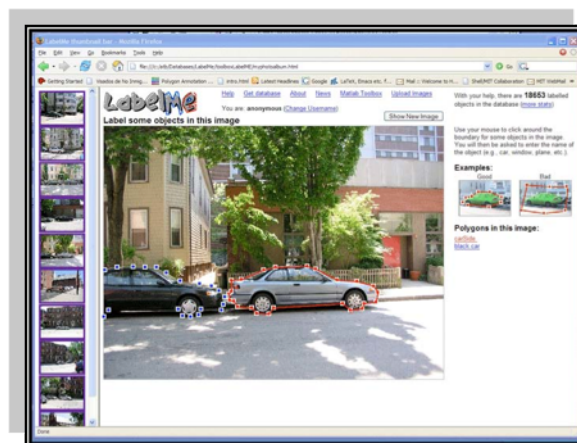


Figure 2.10 - LabelMe website tool

To ensure dataset manipulation and content viewing, a Matlab (a numerical computing environment) toolbox [38] has been developed. Functionalities that are implemented in the toolbox can be used to perform queries, online tool communication, image manipulation and other dataset extensions.

2.2.3 Manhattan story mashup

The SensorPlanet project at Nokia Research Center developed the Manhattan Story Mashup [24] with the aim of collecting image information. This is a game that combines the web, camera phones, and a large public display. It is an interactive game with online users and street users, and provides a new kind of storytelling. Over 150 players played it over the web, while in the outdoor side of it registered an amount of 184 players in Midtown Manhattan. This game was played on September 23rd 2006 between noon and 1:30 pm in Midtown Manhattan, and it was included as one of the featured games in Come Out and Play Street games festival.

This application also uses human computation to perform annotations, as this urban photo hunt brings close together the virtual and the real world, in a real time game. In this case, illustrating stories by taking photos provides new images to train computer vision algorithms, with different sizes, perspectives and orientations, all of them with labels describing its content.

The MSM game works in the following way: a web player uses the MSM web tool to mash up stories, writing sentences or reusing already illustrated ones. Afterwards, a noun from the sentence is sent to the street player's mobile, and they have to take a photo that describes the given word in less than 90 seconds. The photo taken by the street users was then sent to other two street players that had to pick up a noun (from a set of four, including the correct one) for that photo. If the photo-noun was picked up correctly, the original sentence was showed with the photo and turned into a valid piece of information for new stories.

The application's global objective was to extract valid content information, and as described, it tried to contribute to both sides of the tagging paradigm. In one hand describe photos with sentences, and in the other hand create new sets of photos to describe labels, ensuring a diverse image repository for further studies.

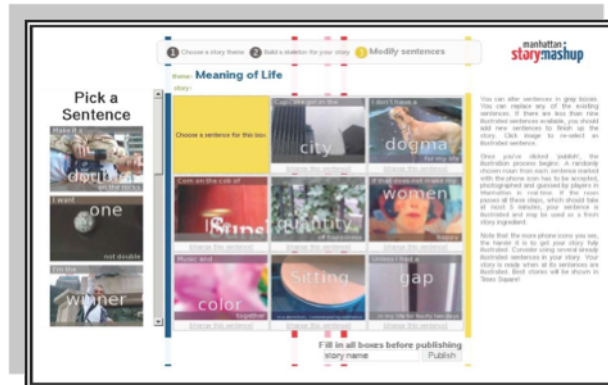


Figure 2.11 - The SMS web tool

There were two main technologies used in this project, to support the street players and the web players, both rather different in their objectives and principles. With no registration required, a user can pick up a previously contributed story and use it as a prelude for her story, or just use previous sentences to mash up or remix a personal story. The street players used the Nokia N80 mobile using S60 3rd edition software platform, Wi-Fi support and a 3 Mega pixel digital camera. The software was built with Python that is frequently used by programmers for its rapid prototyping and extendibility.



Figure 2.12 - The SMS mobile client application

For this event, a huge display was used to show the stories as they were built. The Reuters Sign in Times Square was the main stage, where people could look at the screen while the stories appeared within a time gap of 1 to 5 minutes. This leveled up the game's interest because all the street players could see their photos been showed on a huge display. As to results, a total of 184

players played the game and a total of 3142 photos were taken. In this processes 4529 guesses were made, 2194 (48.4%) of which were correct.

2.2.4 Flickr – A public image-sharing tool

One of the most popular and effective ways for people to organize, share and catalog images over the Web is Flickr [25]. This was a project that begun as a tool for Ludicorp game Neverending, a web-based multiplayer online game. This application popularity exceeded all expectations and Neverending Game became obfuscated, bringing Flickr to an independent context. After a period of constant (and still ongoing) mutation, Flickr became one of the most important realities worldwide, so that Yahoo in 2005 acquired Ludicorp and consequently Flickr, migrating their previously Yahoo!Photos [44] (Yahoo's photo sharing service) to the Flickr database in 2007.



Figure 2.13 - Flickr website

Flickr is basically a web photo-sharing tool that allows users to share and organize their photos. It allows public and private image storage – private meaning that a user can restrict the image access to others, by means of control lists. The public images can be categorized in large groups, helping others when in need of a specific query regarding an image category. Flickr also uses tags to describe photo content, so when a user uploads a photo she can search other images that fit tag parameters such as places, events or animals. In terms of technology, this application uses the following tools: PHP for core application logic; smarty template engine; PEAR for XML & Email ;Perl for "controlling"; ImageMagick; MySQL 4.0; Java for the node service; Apache Web Server 2; Adobe Flash and Fotonotes for photo annotation.

Among others, there is also a desktop tool for uploading photos called FlickrUploadr that is available in all popular operating systems, and also mobile connectivity for instant photo upload.

In spite of its popularity, Flickr has some disadvantages compared to other systems: in one hand, there is an excessive social weight and freedom in the tagging system and so it cannot be categorized as a game with a purpose and rather a powerful tool designed to upload and tag images. This is a tool for photo sharing and search that has the possibility (among others) of image tagging – not a tagging oriented technology. However, it cannot be denied the positive aspect of this kind of web applications because they are popular and encourage users to organize metadata for search and analysis.

2.3 Semi-automatic image annotation

Analyzing both approaches reviewed before - automatic and manual annotations, it can be said that both of them have problems and strengths. We can furthermore determine that if automatic annotation systems lack in accuracy, manual annotation systems lack in efficiency. To overcome these issues, projects were developed using the better of the two approaches, combining automatic techniques and human feedback or interactivity.

2.3.1 Semi-Automatic image annotation using relevance feedback

Based on the knowledge that humans can perform accurate image annotation (but tend not to enjoy it), and automatic systems can be designed to help on this task, researchers at Microsoft Research, developed a progressive annotation process [31] using content-based image retrieval and user feedback. When the user inputs a keyword query and then gives her feedback, the search keywords are automatically attached to the images that received positive feedback and can then facilitate keyword-based image retrieval in future searches. As expected, the more the system evolves, the more accurate will be the results.

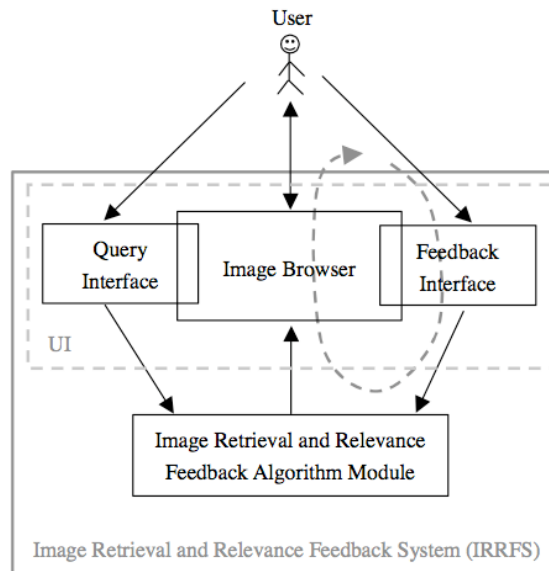


Figure 2.14 - User interface framework scenario

This concept involves common automatic techniques for image retrieval, because that is not the strength of their system, but rather the user feedback aspect. The user interface consists in a framework that provides image search using keywords. If the system cannot find any images tagged by that keywords, it will present a random set of images (which can be confusing for the user), and if there is a group of images in the database annotated by that keywords, the top relevant images appear, as well as other images containing visual similarities (using automatic algorithms).

This system was tested in the MiAlbum [39] prototype, using a desktop application to understand user feedback on such a system. The results were encouraging, but there were some unsolved issues about some parts of this proposal. For instance, when there is no user feedback, this system becomes a simple automatic annotating system, and it relies only on good CBIR algorithms. In that case, there is need for manual annotation of the images, and that can be a difficult task. When there is a good percentage of user feedback, this system relies almost entirely on good relevance feedback strategies.

2.3.2 A Semi-Automatic image annotation using frequent keyword mining

The use of CBIR systems that combine query by visual example and text to overcome perceptual features description is common. Usually these systems provide automatic extraction of most of

perceptual information such as color, texture, shape, structure and spatial relationship. This low-level information processing has the advantage that the applications can be classified as domain independent. Some other kinds of systems tend to use higher level of information such as semantic primitives and related semantic information. These systems tend to be domain-specific, as well as user dependent.

Researchers at the University of London propose a semi-automatic image annotation process using frequent data mining and Fuzzy Color Signature (FCS) [9] to select keywords for new image annotation. FCS is a compact color descriptor scheme and an efficient metric to compare and retrieve images, used in this process to extract the most similar images from an annotated database. With this, researchers hope to establish a bridge between visual data and their interpretation using a weak semantic approach.

This process considers different stages. Firstly, a group of images is hand labeled, using Smeulders [1] notation. Candidate keywords are then extracted from its most similar images (using Earth Mover’s Distance metric) after a frequent pattern mining process. As seen before, this kind of process (manual image annotation) is in most cases a time consuming task, but to overcome the semantic gap problem it is necessary to have user supervision.

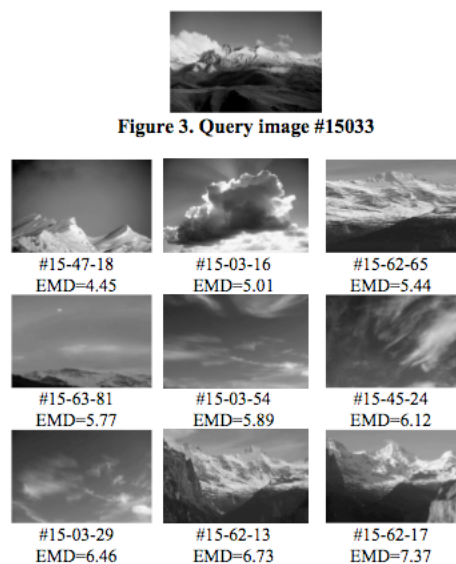


Figure 2.15 - Experimental results using CorelDraw images

This new process was experimented using a set of 2K non-annotated images taken from CorelDraw image CDs. Because this system depends on a pre-set of annotated images, 371 images were downloaded from www.freefoto.com, with their annotations corresponding to the headings grouping photographs by category.

2.4 Gestural interfaces

In our daily lives we interact with other people and objects to perform a variety of actions that are important to us. Computers and computerized machines have become a new element of our society, as they increasingly influence many aspects of our lives. Human-computer interaction is an area concerned with the design, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them. The use of hand gestures and movements provides an attractive alternative to cumbersome interface devices for human-computer interaction applications. Human hand gestures are a mean of nonverbal interactions among people and they range from simple actions of pointing at objects and moving them around to the more complex ones that express our feelings or allow us to communicate with others.

2.4.1 Having fun while using gestures – Eye Toy experience

Involving humans in a game from start to finish was the kind of entertainment that Sony Computer Entertainment group proposes when developing EyeToy® [36]. EyeToy® is a digital camera device, similar to a webcam, for the PlayStation 2 and PlayStation Portable. The device technology uses computer vision algorithms to process images. This allows multiple players to interact with games using motion, color detection and also sound, through its in-built microphone. The camera is mainly used for playing EyeToy® games developed by Sony and other companies, as it is not intended for use as a normal PC camera, although some people have developed unofficial drivers for it. This kind of interaction definitely brought a fresh approach to the game industry, as it involves a computer game and a gesture based interface.



Figure 2.16 – Users playing EyeToy

2.4.2 Developing games with Magic Playground

Another example of games using a gesture recognition system is Magic Playground. This project [3] consists in a game engine that enables the development of entertainment applications with real-time gesture-based Human-Computer Interaction (HCI). The main components of this system are a Video Capture Module, a Statistical Image Processing Unit, a Motion Analyzer, an Image Segmentation Unit and a Rendering Module. The process is the following: The Video Capture Module is responsible for retrieving the image data from the input video device (webcam) and for delivering it to the Statistical Image Processing Unit. After creating a statistical model of the input real scene (by pixel luminous energy examination) and delivered to the Motion Analyzer, this unit performs a YUV transformation and a convolution with a Gaussian Filter, in order to reduce image noise. After composing a motion mask binary image, the next step is Image Segmentation, where all foreground contours are retrieved and the movement ratio of all foreground image blobs is computed. Finally, the Rendering Module is responsible for blending fixed or moving virtual backgrounds with the foreground segmented image, so that background substitution can be performed.

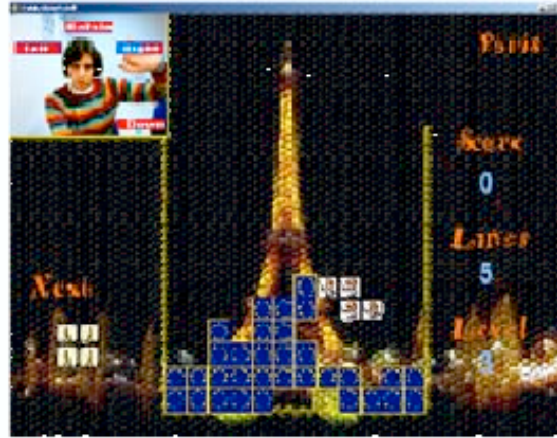


Figure 2.17 – Magic Tetris

Magic Playground was evaluated using MagicTetris, which emulates the Tetris game, where the users play the game by generating moving hand gestures in the appropriated screen control areas (Fig. 2.17).

Chapter 3

Concept and architecture

This chapter describes the main concepts and architecture regarding the Tag Around application. It presents an extensive analysis on the main components and how they are integrated.

3.1 Main concepts

Tag-Around is an application for image tagging that uses human skills to overcome the semantic gap between low-level features and semantic concepts. This application uses an automatic tagging system that previously annotates all the images in a database. After a set of users have played the game (using their feedback) the system corrects those annotations and improves the automatic system. The application also tries to engage the user into having fun, for manual image annotation lacks entertainment as previously explained (see section 1.2.1). This game has a 3D interface and a motion detection module that provides an interactive way for people to annotate images. For motion detection, a camera is used for detecting the user movements and interaction. There is also a face detection module that provides login for a user, so the system will remember that user next time she logs in. The next section will present an overview of the system and all of their components.

3.2 System overview

This section presents an overview of all the main components of the application. Tag Around is a game composed by three main blocks (see fig. 3.1): the application (game), the human interaction and the automatic annotation system. These three modules constitute a semi-automatic annotation system. The system was separated into these main modules with the purpose of modularity. For instance, a different kind of interaction (using hand signs, joystick or even a multi-touch screen) or even other automatic annotation systems could also be easily integrated.

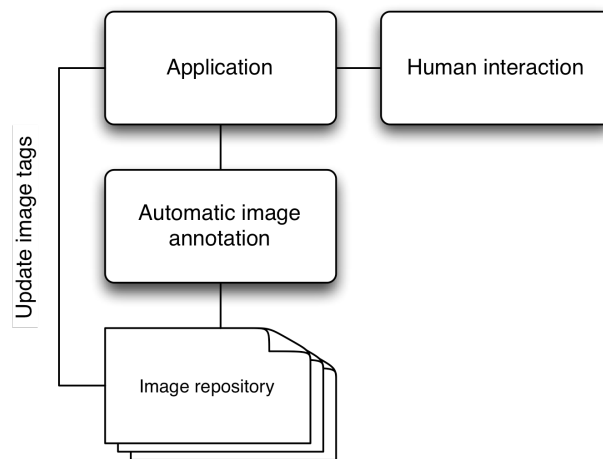


Figure 3.1 – System overview

The application module is the main component of the game; it is composed by a 3D interface and a game engine. Images, tags and the user image are integrated in this interface. These components allow the user to interact with the system. The game engine is responsible for the game score, as well as all the modules involving 3D interaction and motion detection. There is also a module dedicated to perform face recognition. The human interaction block deals with the user interaction; gesture and perceptual inputs are used. Finally, the automatic image annotation module is used for semantic image annotation using the low-level features (e.g., color, texture and shape) automatically extracted.

With these blocks it is defined an algorithm for image annotation [12]. Initially, a set of previously annotated images using the automatic annotation algorithm is presented to the user. Subsequently, for each new move, the user matches a tag with an image, and a set of points is

calculated using the scoring algorithm (see section 4.2.2). If a concept has been annotated with more than N images, the tag model is again trained (with the automatic algorithm). This will improve the automatic algorithm for future annotations.

The next section describes in detail the Tag Around application in terms of specific modules and their objectives. It is described the user interaction (in terms of motion and face detection), the game engine (including the scoring algorithm) and the 3D interface. The section will also describe all the technologies and present a general class diagram of the system.

Chapter 4

Tag-Around

This chapter describes the main components of the Tag Around application. It presents all the main modules that compose the interface, as well as the game engine and a scoring formula. An overview perspective of the automatic annotation system as well as all the technologies used in this project will also be presented in the next sections.

4.1 Human interaction and interface

To provide a different and interesting game interaction, this application uses a gesture-based interface. The objective is to use hands instead of sitting down using a keyboard, pads or joystick. In this case, a user stands in front of a camera at home or in a public place, and using gestures rotates images and tags with the objective of pairing them up.

4.1.1 Gesture user interface

To play Tag Around the user has to perform hand movements in pre-defined hotspots. These actions are captured by a camera and then processed by motion detection algorithms. This type of interactive navigation is made in some areas of the screen (designated by hotspots), which give access to the game as well as to a highscores screen. In a second stage, the player uses the

hotspots to rotate annotations and images and pair them up, to achieve a maximum set of points. These points are saved with the player profile and shown in the highscores area.

4.1.2 Perceptual user interface

In Tag Around it is difficult to maintain and update the information about each user, because the interaction is not made using the usual techniques like the mouse or the keyboard. The proposed solution to register the users in this system is to make the player login based on the recognition of their faces.

To use this interface, the player should place his or her face on an area limited by a square for ten seconds. During that period, the system proceeds to the face recognition (see section 4.2.4) showing the progress of the recognition. The next figure illustrates the player's login interface of this system.

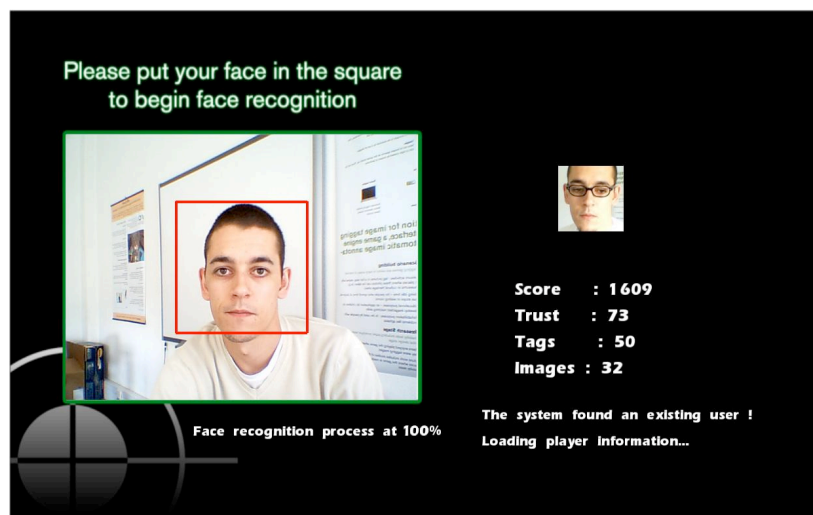


Figure 4.1 – Player login interface

4.2 Game application

This section describes the game application module. This application is divided in different modules: the interface, the game engine, the motion detection module and the facial recognition module. These modules have been created to ensure easy adaptation to the different work scenarios that can be tested with this application. Using Tag Around in a school involves a

different interface than using the application in a hospital or even in an airport. The interface can be altered according to the social requirements of the scenario and even the motion detection can be modified pending the different settings (light conditions are an aspect that has to be checked prior to the experiments, as the application is using cameras and face recognition algorithms as well as motion detection).

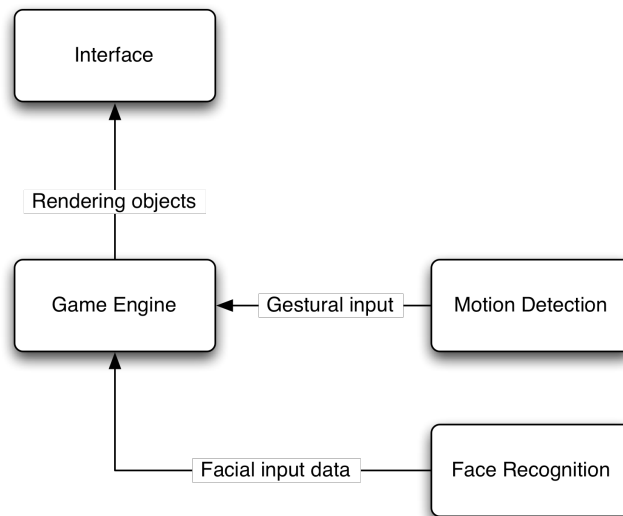
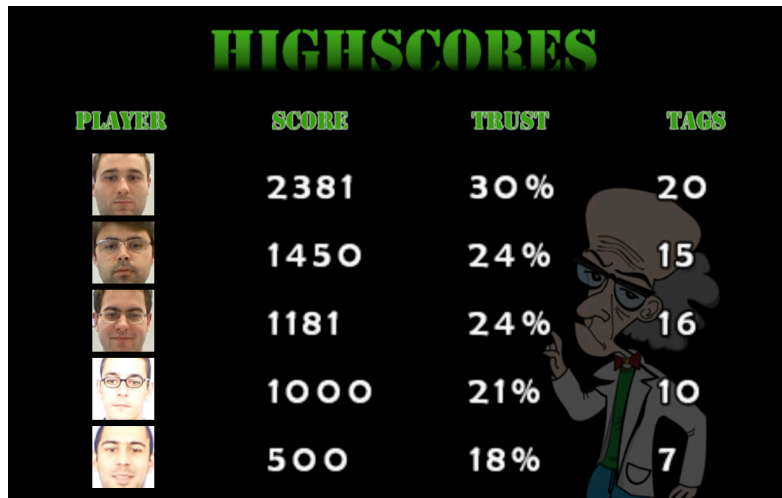


Figure 4.2 – Game application main modules

4.2.1 Game interface

The game interface was implemented using OGRE (Object-oriented Graphics Rendering Engine), a scene-oriented flexible 3D engine written in C++. The goal here was to present the users with a 3D interface scenario, where people could interact with images and tags, and at the same time understand if they were making good or bad annotations. The fact that during some prior tests and discussions people would often ask if they had made a reasonable amount of points and if they were the best players so far brought us the idea of keeping the best players in a highscores interface (see fig. 4.3). The initial layout is composed by two different options - Play Game and Highscores (see fig. 4.4). The user then can pick one up by moving her hand in front of the designated hotspots. If the user chooses the highscores option, she can then visualize the top 5 players which have made the best score while performing correct annotations. When a user plays the game for the first time his or her picture is taken by the system. This image identifies the user

in the highscores list. This was built so that the users do not need any kind of keyboard or input devices (traditional games use a nickname to identify players).

A screenshot of a highscores list on a black background. The title 'HIGHSCORES' is at the top in large, green, blocky letters. Below it is a table with four columns: 'PLAYER', 'SCORE', 'TRUST', and 'TAGS'. The 'PLAYER' column contains five small portrait photos of men. The 'SCORE' column shows values 2381, 1450, 1181, 1000, and 500. The 'TRUST' column shows percentages 30%, 24%, 24%, 21%, and 18%. The 'TAGS' column shows values 20, 15, 16, 10, and 7. To the right of the table is a cartoon character of a man with a large nose, wearing a suit and tie, pointing upwards.






PLAYER	SCORE	TRUST	TAGS
	2381	30 %	20
	1450	24 %	15
	1181	24 %	16
	1000	21 %	10
	500	18 %	7

Figure 4.3 – Highscores layout

Once the user enters the play game mode, it is presented with a facial recognition layout (see fig. 4.1) for login purposes, and then the user starts to play the game (see fig. 4.5). The game interface is composed by several elements displayed in the screen: the user image with different hotspots, a set of tags placed in a rotational platform, a set of images in the bottom part of the screen, an energy bar (that allows the user to perceive when the game ends), the score (that changes depending if the user performs good or bad annotations) and a list of tags that have been already paired up with the image located in the center of the screen.

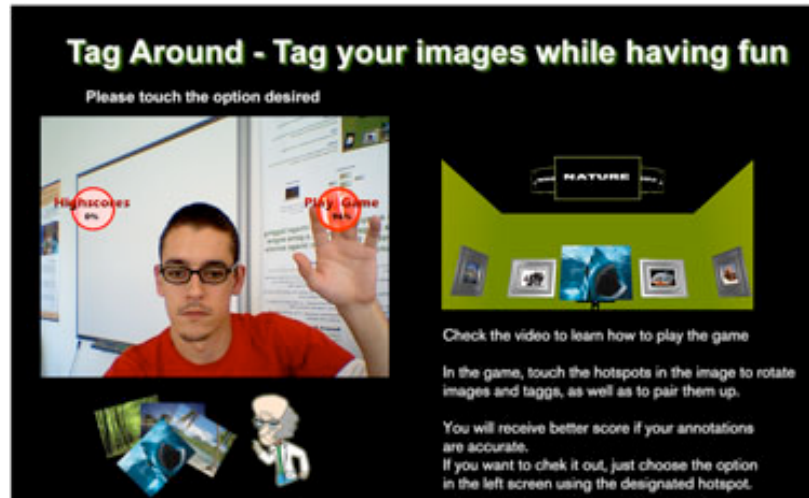


Figure 4.4 – Initial menu layout

When the game ends (the energy bar disappears from the screen), the score, the number of annotations made by the user, as well as the confidence that the user has earned is shown in the screen and the player profile (with that information) is saved in disk for further games played by that user.



Figure 4.5 – Game interface

4.2.2 Game engine

The game dynamics is the following: when the game begins a timer is activated and a set of images (randomly selected) is presented to the user; in the interface there will also be a set of tags which the player has to rotate to annotate the images; The player, using the designated hotspots has to pair up as many images and tags as she can to receive more points and also more energy; the game ends when the user has no more energy left.

While playing the game, new images will appear on the screen (depending on the level of the game) and the user has the possibility of tagging the new set of images with the same concepts. In the beginning of the game, the user has approximately 3 minutes to tag 5 images, and by furthering advancing in the game, less time the user has to tag the images. It is important to notice that a good move improves the user score in the sense that the more energy the user gets (by performing good annotations), more time the user will get to tag other images. The timer is always decrementing, but it is incremented with the user good moves. On the contrary, bad annotations will penalize the user with even less energy (and therefore less time).

One of the main issues behind this game engine was the concept of “good annotations”, because good annotations mean better scores and the goal is to obtain higher scores. Therefore a robust scoring algorithm had to be implemented. After analyzing several cooperating and non-cooperating games and interviewing users to understand the expected game dynamics, the score formulas were developed and then tested. An annotation made by the player (commonly named “move”) is analyzed by 3 distinctive factors: (1) the automatic image annotation algorithm output (see section 4.3.1); (2) the confidence that the system has on the player (that is obtained by previous annotations); (3) the feedback from previous players that placed that same tag on that particular image.

When a player chooses the first annotation in an image, the score will depend exclusively on the automatic image algorithm and the player’s confidence level. A set of points is given to the player and the increase or decrease of the confidence in the player depends of the percentage of success given by the automatic algorithm to that specific tag-image pair.

When a group of players made several annotations in an image, the player score is influenced mostly by the group feedback, becoming a social (and also manual) annotation system. These results will then be matched with the automatic annotation system original output, in an effort to improve the results and efficiency.

As pointed before, the score plays an important role in the game, because it measures the quality of the moves (annotations) performed by a particular player. For good annotations the score should be high and it should be low for bad moves. However, sometimes it is difficult to classify the annotation, especially when the image does not have previous annotations. Assuming the user has to annotate a set of images $L = \{I_1 \dots I_N\}$ with a set of labels $V_w = \{w_1, \dots, w_M\}$, when the player annotates the concept w in a new image I (without previous annotations) the score is given by,

$$S_{new}(I, w, n) = C_{player}(n) + [1 - C_{player}(n)]p(w/I), \quad (1)$$

where $p(w/I)$ is obtained by the automatic algorithm (see section 4.3) and $C_{player}(n)$ is the player confidence that expresses the quality of the previous annotations provided by the player,

$$C_{player}(n) = \begin{cases} k_p n & n < K_{moves} \\ k_{conf} & n \geq K_{moves} \end{cases} \quad (2)$$

K_{moves} is a constant with the number of good moves to reach the player confidence maximum value k_{conf} , n is the number of good moves and k_p is a constant that is used to increment the player confidence.

When the player annotates with the concept w an image I that already has this annotation provided by previous users, the score is calculated using,

$$S_{total}(I, w, n, m) = C_{group}(m) + [1 - C_{group}(m)]S_{new}(I, w, n) \quad (3)$$

The $C_{group}(m)$ represents the group confidence and m is the annotations number of the concept w on image I ,

$$C_{group}(m) = 1 - e^{-\left(\frac{m}{kg}\right)} \quad (4)$$

The number of good moves n increases when the group confidence is different from zero or the score is greater than a defined threshold. It decreases when the score is above another threshold. These thresholds were obtained empirically.

In order to evaluate the model used to compute the score several simulations were performed. Figure 4.6 shows the score evolution for 200 moves using the final training set (more 40 images in the training set). This test was conducted for a player with 5% of wrong annotations (Player 2 in Figure 4.6) and for a player that makes 50% of mistakes (Player 1 in Figure 4.6). It is estimated that 5% of mistakes should represent the behavior of a regular player and 50% is the behavior of a bad player. It is also presented the evolution of the player confidence (equation (4)) for both players. As it can be seen, the score is higher for player 2 (blue curve) than for player 1 (green curve) and the confidence increases for player 2 (black curve) and decreases for player 1 (red curve).

Table 1 presents the mean value of the final score obtained after 200 moves by 10 regular players (5% of errors) and 10 bad players (50% of errors) using the initial training set and the final training set (more 40 images). As expected, the score obtained by the set of regular players is higher than the bad players score. When the training set increases, the regular players score increases and the bad players score decreases. Good annotations improve the semantic models accuracy and since the semantic concepts are an important part of the score computation, this also makes the game more interesting.

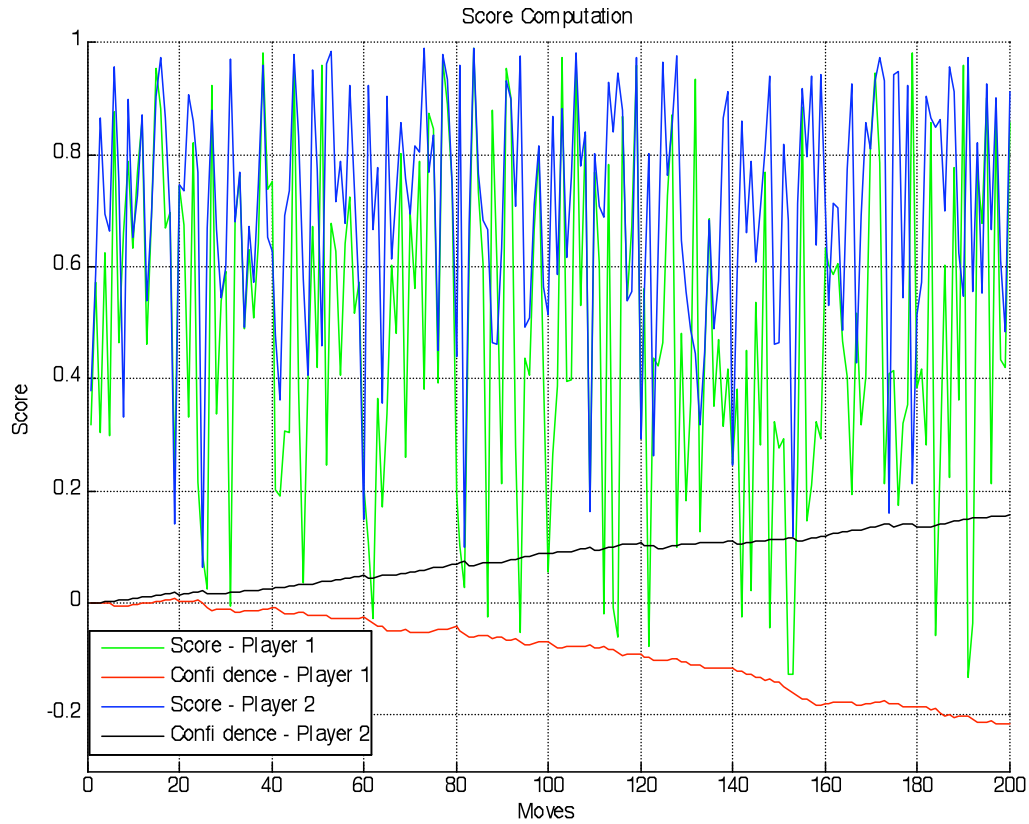


Figure 4.6 - Score and player confidence evolution.

Players	Score	Score
	Initial Training Set	Final Training Set
10 (5%)	14156	14445
10 (50%)	10984	10652

Table 4.1 - Mean of the final score.

4.2.3 Motion detection

As mentioned previously, the Tag Around application has a gesture based interface. For this kind of interaction, OpenCV was used for detecting movement. OpenCV (Open Source Computer

Vision) is a library of programming functions mainly aimed at real time computer vision. One of the issues regarding any gesture-based interface is the kind of interaction to have. After the paper prototype testing (see section 5.1), the results showed that the users would have better success in coordinating their interaction with the interface if they simply had to make simple gestures in order to rotate images and tags. Several techniques including hand signs and flow motion gestures were tested, but the most successful interaction was made with simple hand motion. Therefore, in the final version, the motion detection algorithm was used.

4.2.4 Face recognition

This module uses image-processing algorithms to detect and recognize the user's face. This module was provided by Filipe Grangeiro [6] for testing his own work and developing a new approach for the Tag Around login interface.

This module handles three tasks: detection, normalization and recognition of faces. The first step detects the presence of a face on an image captured by the camera. The method is based on the system described in [26] complemented with a skin detection algorithm to confirm the detected face. In this step, the player's face is also extracted from the captured image to be used in the next two steps.

Once a face is detected and extracted from the captured image, it can further be normalized. The second step normalizes the detected facial image. The goal of this step is to transform the face image into a standard format that attenuates variations that can reduce the performance of the face recognition algorithm.

In the final step, facial images were represented with a technique used in [23] to reduce the dimensionality of the face data. Then, the faces were classified using a machine learning algorithm called Support Vector Machines [17] in a binary tree structure strategy proposed in [7] to classify more than two individuals. To complement this process, a facial pose estimation technique was also implemented using the method described in [26] to compare only facial images having the same facial pose.

4.3 Automatic image annotation

Tag Around is a semi-automatic interface for image annotation. It was conceived to help bridging the semantic gap between low-level features and semantic concepts. Some of the systems use exclusively automatic systems to tag images, but this application tries to engage manual annotation as a way to improve such systems. This application, as pointed before, depends exclusively on the automatic system when no manual annotation has been made in a particular image. The annotations performed by the players will be used in the automatic system to improve future results. This section will describe the automatic image annotation system.

4.3.1 Automatic Image Annotation

Given a training set, previously annotated with a pre-defined set of tags, a probabilistic model is estimated for each concept that gives the probability of a tag (object or scene) being present or absent in a given image. These models are trained using the low level features automatically extracted from the training images. New images are classified according to these models. This automatic algorithm lacks accuracy as expected, which motivate the use of this application. It uses the Regularized Least Squares Classifier (RLSC) [12] to perform a binary classification over the database and the sigmoid function give a probabilistic sense to the classifier output. The models were evaluated in [12]. Initially all the database is classified with the estimated models but when a concept is annotated in more than N images these pictures are included in the training set and the model is estimated again. More annotations (player moves) will improve the models and consequently the score will reflect with better precision the quality of a move. This work was developed by Rui Jesus and was the starting point of this application. A more detailed analysis can be found in [12].

4.3.2 Updating the parameters

Automatic annotation systems that use semantic concepts employ training sets with pre-annotated images to engage image classification. The Tag Around automatic system uses Flickr [37] images that have been previously annotated to train the RLS Classifier. With the images that have been annotated with the application, improved probabilistic models will be created and the accuracy increases as stated in [12].

4.4 Implementation

This section describes all the main technologies used in the implementation of the Tag Around application. It will describe the programming environment as well as other applications that supported the prototype and the final game.

4.4.1 Technology

The use of correct technologies to make a useful and effective application is an important issue. A technology should adapt to the requirements and be easy to use. There were three different technologies involved in the making of the application. There were many different technologies that could work for this, but after discussing and analyzing the possibilities, the choice became somewhat obvious. The application consisted, as presented before, in a set of modules that uses a 3D environment platform, as well as a motion detection system. Additional technology was used to extract data from XML files.

OGRE3D - OGRE (Object-Oriented Graphics Rendering Engine) is a scene-oriented, adaptable 3D rendering engine. It is written in C++ and is designed to make it easier and intuitive for developers to produce applications (in this case, it is a game oriented application) using hardware-accelerated 3D graphics. The class library abstracts the details of using the underlying system libraries like Direct3D and OpenGL and provides an interface based on world objects and other high level classes. All the interface was designed using Photoshop, and then implemented in OGRE using 3D Studio Max for all the 3D objects in the Tag Around interface.

XML parsing - To extract keywords and the images correspondent probabilities, as well as image paths, a XML parser was used. In this case, the Xerces-C++ [45] makes the application ready to read and write XML data, using a shared library that parses, generates, manipulates and validates XML documents. This library uses DOM [35], SAX [41] and SAX2 APIs.

OpenCV - OpenCV (Open Source Computer Vision) is a library of programming functions mainly aimed at real time computer vision. Some of uses of the OpenCV library are Human-Computer Interaction (HCI); Object Identification, Segmentation and Recognition; Face Recognition; Gesture Recognition; Motion Tracking, etc. In this case, it was used to perform Motion Detection and Face Recognition.

4.4.2 Application modules

The Tag Around application is divided in several different modules. The next figure describes an overview of the main modules.

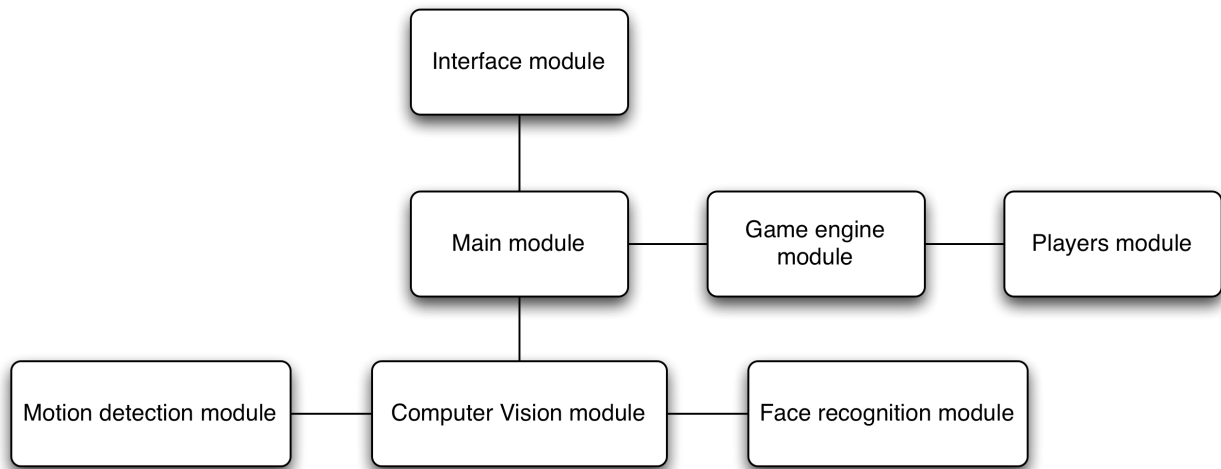


Figure 4.7 - Tag Around main modules

Interface module – It is responsible for all the objects in the interface, as well as all the motion in the game (images rotation, tags rotation, animations, etc.). It was developed using OGRE3D.

Main module – It is the core class of the application. It handles the time manager, the main loop and connects all the modules that correspond to the game engine and the computer vision modules.

Computer vision module – It is responsible for capturing the frames from the camera and analyze the users motion in the designated hotspots. It was developed with OpenCV classes and algorithms.

Game engine module – It is the core of the game, as it computes all the scoring, confidence levels as well as all the annotations made by a player during the game.

A more extensive description on the project classes can be found in the Appendix section.

Chapter 5

Interface design

To design the application, different scenarios and opportunities for playing the game were considered, e.g., while waiting or when visiting a place, such as a museum. The time that people spend waiting for an event or simply doing nothing was understood to be a frame window for applications that: (1) help people to spend their time; (2) help the community to create folksonomies. The idea of people playing a game without a (visible) computer also influenced this work. As a result of brainstorming, several opportunities and scenarios for playing the game were proposed:

- Leisure activities - for people who want to have fun tagging photos, especially in the places where these photos can be taken;
- Using idle time – for people who spend time at airports, bus stops, or waiting rooms;
- Educational purposes – for children that could use this application to develop image/text matching skills;
- Rehabilitation purposes - for people who have problems like aphasia;

After defining what scenarios and therefore objectives to achieve, a paper prototype was built to test Tag Around in terms of functionality.

5.1 Paper prototype

One of the relevant features of the application is the use of a video camera, such as a web camera, and the interaction with human gestures for image tagging. The paper prototype (see figure 5.1) included a series of tasks presented to the users, which had prior knowledge of the main concept but did not know about the gesture interactivity. Paper prototype tests were done with five users all of them college students with experience in working with computers.

To start, users were asked to interact with the application with no prior knowledge of the objectives. In the next stage, users were told how to perform annotations, using the tags in the top part of the screen and the images in the bottom part of the screen. They had to perform correct tagging, without knowing the time and score restrictions. Finally, users were asked to perform correct annotations, knowing now that they had a time limit and a score associated with every annotation they made. After these tests the interface was refined and the Tag Around application was developed.



Figure 5.1 - Paper prototype

5.2 Usability tests

The Tag Around was subjected to usability testing, aiming to evaluate the interface complexity, usefulness and aesthetic aspects, to understand how easy it is to learn and use and to analyze the fun component of the game. The usability tests are described below. The questionnaire can be analyzed in detail in the appendix section.

5.2.1 Participants

15 voluntary participants, 8 of them female, tested the application. The participants in this experiment ranged in age from 18 to 31 years old with a mean age of 24. Ten of the participants work in the field of information technologies. All participants had their first contact with the application during the test and used it under similar conditions. All participants frequently use the Internet to search for images and they all claim they use their computers to manage personal images, but only about 50% do it frequently. The participants also declared they only catalogue around half of their image collection in average. When they need to search for a particular digital image in their computers, they all (except 3 who made no comments on this issue) search folder by folder until they find it. One of the participants also declared to use IPhoto. They use their images mainly for work purposes or for future memory of life experiences.

5.2.2 Setup and methodology

The tests were conducted by two researchers in a university office and were accomplished individually by each user. Participants were first briefed about the objectives of the test. After a short description of the application and an explanation of the goals to be achieved, users were encouraged to explore the Tag Around game, with no objective goal associated. After that, users were asked to play the game with the objective of performing a maximum set of points, annotating a several number of images according to the labels available on the system.

During the initial test, participants were persuaded to “think aloud” and were allowed to ask for help if they really did not know what to do. All users’ comments were recorded for future analysis. When they finished the game (the time/energy bar disappeared) users were asked to fill in a questionnaire and express their opinions regarding the application they had just tested. The main objective in this questionnaire was to perceive if the application was: easy to learn, easy to

use, useful, enjoyable, engaging and intuitive and also to analyze new approaches to enhance dynamic interaction and aesthetic aspects of the user interface. Each test lasts for a maximum of 30 minutes, depending on the users' performance, since each wrong label selected causes a loss of energy (time) and the game stops when there is no energy left. All the information collected was then analyzed with the ultimate goal of refining the Tag Around as described below.

5.2.3 Questionnaire

The questionnaire captured user's personal data and experimental feedback. It was composed by five sections: personal data, motivation, game dynamic, interaction and aesthetic aspects. It also raised several open-answer questions. Personal data included age, gender and digital image usage. The experimental feedback was measured by a total of 24 questions distributed by 4 sections (motivation, game dynamic, interaction and aesthetic aspects). Three of the questions were open answer questions, aiming to collect suggestions concerning changes and improvements that could be made in the interface. The remaining ones were answered on a 5-point Likert-type scale, where 1 = totally disagree, and 5 = totally agree.

5.2.4 Results

This section describes the most important preliminary conclusions and observations made during the tests. The options selected by the different participants for each question were analyzed and the average scores were calculated to observe if there were general trends in disagreement or agreement with the corresponding statements (strong feelings one way or the other showing up as mean scores closer to 1 or 5) and the standard deviation of the mean score to evaluate how broad the consensus about the issue was.

Easy to learn

Several participants had some trouble to find out what they should do to initialize the application and start playing. Some needed help from the researchers supervising the tests to carry on. This happened because the motion detector in these hotspots was calibrated to a different scenario (the camera distance to the user and brightness conditions). When users manipulated the application for the first time, they needed a short period of time to understand the interaction paradigm and to get used to the features available. However, they seemed to quickly realize what to do. Most

participants agreed that “it was easy to learn how to use the application” (Mean = 4.27, SD = 0.57) and that “it was easy to use the application” (Mean = 4.33, SD = 0.60).

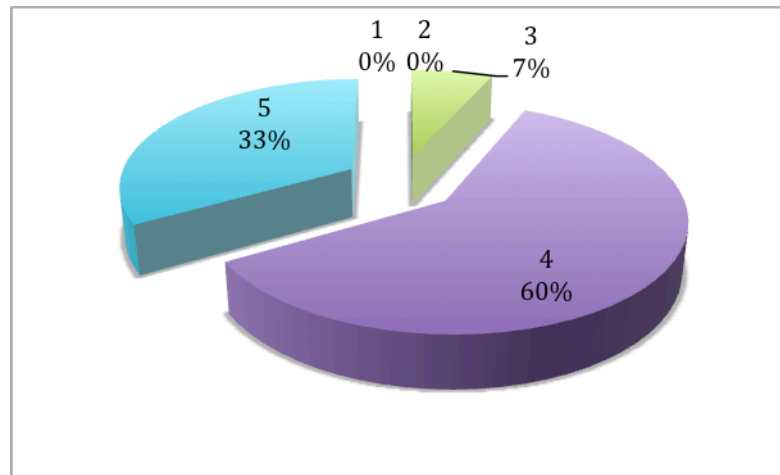


Figure 5.2 – “It was easy to learn how to use the application” question

Interaction

The participants tended to agree with the statement “it is easy to manipulate the hotspots used to rotate the images and the concepts”, though on average they have a neutral opinion (Mean = 3.33, SD = 1.10). If “the usage of this type of interaction is physically exhausting” was a controversial question: (Mean = 2.10, SD = 1.20), since most participants disagreed, but one totally agreed, another partially agreed and 3 had a neutral opinion. Identical results were obtained for the sentence “the usage of this type of interaction is mentally demanding”. In general participants disagreed with the sentence “the application would be more intuitive if I could use the keyboard and the mouse instead of the gesture input” (Mean = 2.53, SD = 1.31). However, it was a very controversial question: while 5 participants totally disagreed, 1 totally agreed and 3 partially agreed. Although the results are not as convincing as it was expected, there is a tendency to agree with the interaction technique proposed.

Application features

The visual interface includes several elements, such as the images and tags to be paired-up, the tags already associated with the current image, the score and the time elapsed (see figure 4.5).

The objective was to focus on the understanding of how users perceived these elements and all the application dynamics, as well as in identifying their preferences regarding the application features.

When asked if “they were able to understand how the score evolves during the game”, users gave very different answers, 2 participants totally disagreed while 2 others totally agreed (Mean = 3.20, SD = 1.22). The score calculation is based on many different issues and users were mostly concentrated in their actions, so they did not have the time to examine in detail how the score was processed during their short usage of the application. However, they all detected a correlation between the score and the correctness of the annotations they made.

Each participant had exactly the same opinion about the sentences “the images should stand still and only the annotations should rotate” and “the annotations should stand still and only the images should rotate”. The majority of the subjects totally disagreed with both sentences. This indicates that the opinion of allowing users to rotate both the images and the annotation seemed to be appropriate.

To the sentence “the application would work better if there were more images to annotate”, the majority of subjects (9 out of 15) had a neutral opinion (Mean = 3.07, SD = 0.77). Quite the same happened with the sentence “the application would work better if there were more available tags to select”. These questions were not very conclusive. To achieve further results additional comparative tests should be performed. The majority of the participants stated that they liked the interface (Mean = 4.20, SD = 0.54) as well as its aesthetics (Mean = 3.80, SD = 0.65).

Usefulness

Most of the subjects claimed, “it was fun to use the application” (Mean = 4.47, SD = 0.62). When asked if “they would use the application in a public place while waiting for any service”, most of them agreed totally or partially, only 3 kept a neutral position (Mean = 4.40, SD = 0.80). A similar attitude was detected when subjects were asked if “they would use the application to have fun with family and friends”. These are important results because the goal was to build an application to tag images in a fun way and in public spaces.

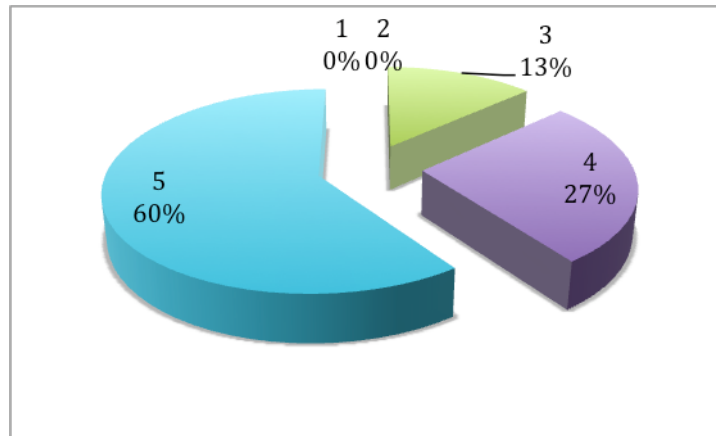


Figure 5.3 – “would you use the application to have fun with family and friends” question

Most subjects agreed that they would use the Tag Around game for their personal use. The majority of the participants agreed with the statement “it would be more fun to annotate my own images with my own annotations”. However, the answers to a related question, if “they would use the application to catalogue their own images” were not so consensual: 2 totally agreed, 7 partially agreed, 3 had a neutral position and the remaining 3 partially disagreed (Mean = 3.53, SD = 0.96).

Open questions

From the analysis of the open answer question and the comments made by the participants during the test, it was possible to collect some ideas that will help to improve the application tested. The collected data and resulting plans are summarized below. It could be noticed during the tests that all subjects were very engaged in doing a good score by correctly cataloguing as much images as they could. Some users were even interested in finding out how the score was calculated, how the energy bar works, how they can get to a higher level and tried to examine all the application mechanisms. Most of the participants memorized the concepts available for annotation and even the order in which they were shown and most of them knew how many concepts were available.

When starting to use the application many users did not know they have to move their hand when they try to select a hotspot for moving the images or concepts or to make an annotation. However, they quickly recognize that requirement. Participants who were familiar with EyeToy [36] intuitively move their hands when selecting a hotspot. They find Tag Around more useful,

since it has a practical usage and they can have their images annotated while playing with their friends. The participants noticed that the available concepts moved around in a circle, but some of them did not recognize at first that images have the same behavior. This can be explained by the different style in which they are presented and by the fact that the images change when the game level changes. This later behavior, which confused a few users, seems to be adequate for a game, but not so much when the user is only interested in cataloguing their images.

User suggestions

The participants were also encouraged to make comments that pointed out possible improvements to be made: “There should be a “cancel” button to undo annotations”; “During login, there should be a timer (count down) to indicate when I can start playing”; “The tags associated with the current images should be presented in a position closer to the corresponding image”; “Sound could be used to emphasize good and bad annotations”. Several users mentioned that score and level information should be highlighted. Concerning the interface aesthetic, the main remark was related to the colors used in the interface. Participants suggested the use of more appealing colors that highlight key information, such as level and selected tags for the current image. In general, participants described the application as useful, funny, easy and intuitive.

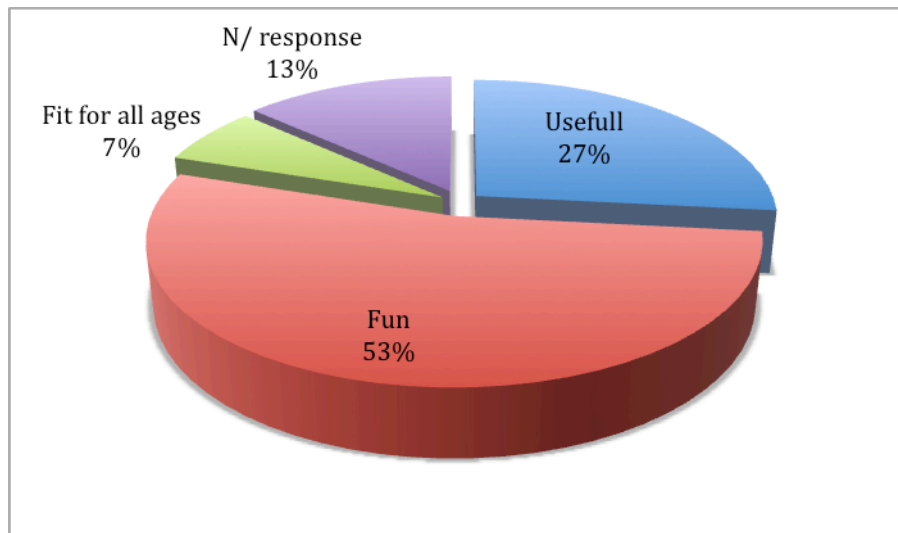


Figure 5.4 – General comments about Tag Around

Chapter 6

Conclusions and future work

As a collaborative society, the need for organization and sharing of multimedia contents is becoming more and more essential. The huge amounts of digital information (in particular images) spread around databases in the World Wide Web raises problems in terms of searching for specific content. Typically CBIR systems use low-level features for image searching that lack accuracy, because any image query is usually expressed semantically. For this matter, researchers propose new ways for image annotation that overcome the semantic gap. Manual annotation is a solution but it lacks motivation. Humans can use their computational skills to resolve this kind of problems but tend not to enjoy it.

Given this set of premises, this project proposes to overcome the negative aspects of automatic image annotation as well as the lack of motivation in manual annotation. It motivates users by presenting a fun game where people can play in different places, using nothing but their hands to interact with the interface. People play games all the time and they could use that energy and time to help doing image annotation. This application also tries to supply correct image annotations for

CBIR repositories, so future work can improve the automatic algorithms and reduce the semantic gap.

This project also addresses the dynamics of human-computer interaction, in terms of interface interaction and user feedback. The game dynamics as well as the game engine methodology was an interesting and demanding subject – build an algorithm that helped image annotation (manual and automatic) and at the same time provide a reasonable understanding of what happens to the users that are playing the game.

6.1 Alternative human interaction and design interfaces

Exploring alternative methods of user interaction is one of the novelties of this application if compared to other traditional systems (that are used for image annotation). To use the keyboard or mouse to interact with the system was understood to be a downside to the entertainment factor. As several approaches have been promoted in the game industry (e.g., EyeToy, Wii) with successful results, it is believed that users will benefit from different kinds of interaction modalities.

As this project continues, new kinds of interaction like hand signs, motion flow detection (which includes speed and interaction of motion) as well as sound or other human natural language expressions will be studied and tested in Tag Around.

6.2 Future work scenarios

This project was built upon the concept that manual image annotation is an effective way to overcome the semantic gap. To overcome the lack of human motivation on image annotation several scenarios were designed and projected. As a stable version of the Tag Around project is concluded and the usability tests have been performed and analyzed, there is opportunity to refine some aspects regarding the interface and user interaction. To do this, tests are planned with real case scenarios like Schools, Airports and other public spaces (hospitals, museums, etc). Multiple scenarios are important because unrelated social groups can bring new approaches as well as problems to the design methodology. There are also issues regarding the user image background

(people passing by while a user plays the game) as well as luminosity in the physical spaces. It is hoped that testing the application in different scenarios and analyzing results will bring improvements as well as new opportunities for development.

Chapter 7

References

- [1] A. W. M. Smeulders, M. W., S. Santini, A. Gupta, and R. Jain Content-Based Image Retrieval at the End of the Early Years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22, 12 2000), 1349--1380, 2000.
- [2] B. Manjunath , W. M. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*1996), 1996.
- [3] Cabral, C., Dehanov, Juana, Miguel, Jose , Dias, Salles , Bastos , Rafael. Developing games with Magic Playground: a gesture-based game engine. In *Proceedings of the ACE 2005 International Conference on Advances in computer entertainment technology* (Valencia, Spain, 2005), 2005.
- [4] D. Forsyth, J. M., M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. *Finding pictures of objects in large collections of images*, 1996.
- [5] Furnas, G. W., Fake, Caterina, von Ahn, Luis , Schachter, Joshua , Golder, Scott , Fox, Kevin , Davis, Marc , Marlow, Cameron , Naaman, Mor *Why do tagging systems work?* ACM Press, 2006.

- [6] Grangeiro, F. *Detecção e Reconhecimento de Faces num Contexto de Memórias Pessoais*. New University of Lisbon, Lisboa, 2008.
- [7] Guo G., L. S. Z., Chan K. *Face Recognition by Support Vector Machines*, 2000.
- [8] Hanjalic, A., Lienhart, R. , Ma, W.-Y. , Smith, J. R. *The Holy Grail of Multimedia Information Retrieval: So Close or Yet So Far Away?* Proceedings of the IEEE, 2008.
- [9] Izquierdo, E., Dorado, A. Fuzzy Color Signatures. In *Proceedings of the Proc. of the IEEE Int. Conf. on Image Processing* (Rochester, NY. USA, 2002), 2002.
- [10] J. Fan, Y. G., H. Luo. Hierarchical classification for automatic image annotation. In *Proceedings of the SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (New York, NY, USA, 2007). ACM Press, 2007.
- [11] Jesus, R., Dias, R., Frias, R., Abrantes, A., Correia, N. Sharing Personal Experiences while Navigating in Physical Spaces. In *Proceedings of the 5th Workshop on Multimedia Information Retrieval in 30th international ACM Information Retrieval Conf (SIGIR07)*, 2007.
- [12] Jesus, R., Gonçalves, D., Abrantes, A., Correia, N. Playing Games as a Way to Improve Automatic Image Annotation. In *Proceedings of the IEEE International Workshop on Semantic Learning Applications in Multimedia (SLAM08), in conjunction with CVPR08 (2008)*, 2008.
- [13] Kosorukoff, A. *Human based genetic algorithm*, 2001.
- [14] Li, J., Wang, J. Real-time computerized annotation of pictures. In *Proceedings of the ACM Intl. Conf. on Multimedia* (2006), 2006.
- [15] M. , Z. *VLDB '75: Proceedings of the 1st International Conference on Very Large Data Bases*. ACM, Framingham, Massachusetts, 1975.
- [16] M.L. Pao , M. L. *Concepts of Information Retrieval*. Libraries Unlimited, 1989.
- [17] Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, 122001, 181-201, 2001.

- [18] Pass, G., Zabih, R., Miller, J. *Comparing images using color coherence vectors*, 1996.
- [19] Russell, B. C., Torralba, A., Murphy, K. P. and Freeman, W. T. LabelMe: a Database and Web-Based Tool for Image Annotation. *MIT AI Lab Memo AIM-2005-0252005*, 2005.
- [20] Smith, J. R., Chang, S. *Transform features for texture classification and discrimination in large image databases*, 1994.
- [21] Smith, J. R. a. C., Shih-Fu Visually Searching the Web for Content. *IEEE Computer Society Press*, 4, 3 1997, 12--20, 1997.
- [22] Swain, M., C. Frankel, and V. Athitsos WebSeer: An image search engine for the world wide web1997, 1997.
- [23] Turk, M. A., Pentland, A. P. *Face recognition using eigenfaces*, 1991.
- [24] Tuulos, V., Scheible, J. and Nyholm, H. Combining Web, Mobile Phones and Public Displays in Large-Scale: Manhattan Story Mashup. In *Proceedings of the Proc. of Pervasive Computing 07* (London, 2007). Springer, 2007.
- [25] Van House, N. A. Flickr and public image-sharing: distant closeness and photo exhibition. In *Proceedings of* (New York, NY, USA, 2007). ACM Press, 2007.
- [26] Viola, P., Jones, Michael Robust Real-time Object Detection. *International Journal of Computer Vision*2004), 137-154, 2004.
- [27] von Ahn, L., Dabbish, L. Labeling images with a computer game. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems CHI '04*, 2004.
- [28] von Ahn, L., Dabbish, L. Games with a purpose. *Computer*, 392006, 92--94, 2006.
- [29] vonAhn, L., Liu, R and Blum, M. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the CHI 2006 Proceedings* , 2006.
- [30] Wang, J. Z., Boujemaa, N., Del Bombo, A., Geman, D., Hauotnabb, A., and Tesic, J. *Diversity in multimedia information retrieval research*, 2006.

- [31] Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B. Semi-Automatic Image Annotation. In *Proceedings of the Human-Computer Interaction--Interact '01*, 2001.
- [32] X.-J. Wang, L. Z., F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *Proceedings of the CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2006). IEEE Computer Society, 2006.
- [33] Zhao, R., Grosky, William I. *Bridging the semantic gap in image retrieval*. IGI Publishing, 2002.
- [34] Delicious <http://del.icio.us>, 2008
- [35] DOM <http://www.w3.org/DOM/>, 2008
- [36] Eye Toy <http://www.eyetoy.com>, 2008
- [37] Flickr <http://www.flickr.com>, 2008
- [38] LabelMeToolbox <http://labelme.csail.mit.edu/LabelMeToolbox>, 2008
- [39] MiAlbum <http://research.microsoft.com/research/pubs/view.aspx?pubid=915>, 2008
- [40] Phetch <http://www.peekaboom.org/phetch>, 2008
- [41] SAX <http://www.saxproject.org/>, 2008
- [42] Verbosity <http://www.peekaboom.org/cgi-bin/verbosity>, 2008
- [43] Wikipedia - Web 2.0 http://en.wikipedia.org/wiki/Web_2, 2008
- [44] Wikipedia-Flickr <http://en.wikipedia.org/wiki/Flickr>, 2008
- [45] XERCES <http://xerces.apache.org/xerces-c/>, 2008

Appendix A

Usability tests

I. Questionnaire

TAG AROUND

Aplicação 3D para a anotação de imagens

Conteúdos multimédia são trocados a todo o momento na Internet a um ritmo nunca visto. Vídeos e imagens enchem os nossos computadores, blogues e comunidades online espalhadas pela rede. É necessário organizar todo este conteúdo para uma melhor pesquisa e utilização do mesmo.

Tag Around é um projecto que propõe analisar mais profundamente a questão motivacional e lúdica da anotação manual de imagens, propondo uma solução em que os utilizadores se divertem enquanto anotam as suas imagens.

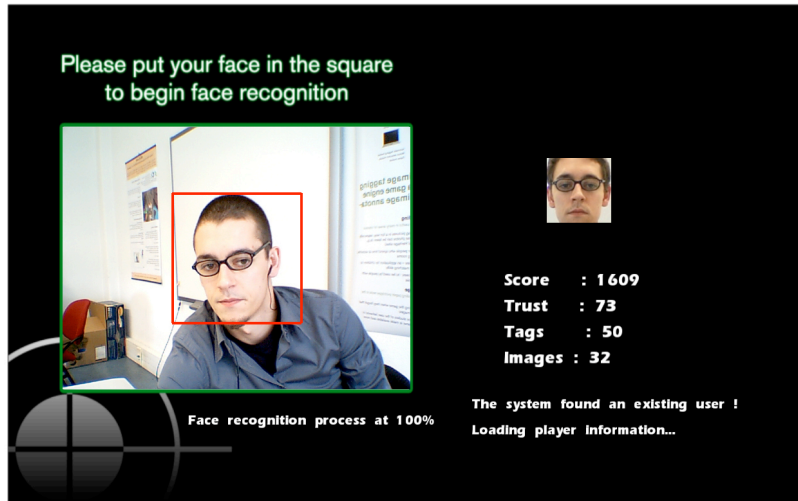
Durante esta sessão, pretendemos compreender a iteração dos utilizadores com a interface, em termos da sua complexidade, facilidade de aprendizagem, divertimento, compreensão dos objectivos propostos, e aspecto audiovisual da interface. Para isso propomos que experimente a aplicação, complete os objectivos propostos, e que acima de tudo, se divirta enquanto explora as suas potencialidades.

Aplicação

Esta aplicação consiste num jogo 3D cujo objectivo é anotar correctamente o máximo número de imagens no menor espaço de tempo possível. Como acreditamos que teclados e ratos são aborrecidos, vamos tentar interagir tanto com as imagens como com as anotações usando apenas

gestos. As imagens seguintes descrevem as várias etapas da aplicação, para se familiarizarem com a mesma.

Antes de começar a jogar este jogo, o sistema precisa de identificar o jogador no sistema. Para isso, terá de colocar a sua cara dentro do quadrado encarnado, enquanto o sistema o tenta identificar. No caso de ser um jogador novo, o sistema irá criar um novo perfil.



Logo após o login, o jogador irá começar a jogar. Antes porém, vamos fazer um *preview* do que irá acontecer.



Pontuação : A pontuação reflecte-se na tua perícia de anotar as imagens, associando os conceitos (em cima na imagem, ás imagens em baixo)

Energia : A energia vai aumentando com boas anotações, e diminuindo com o tempo e com as más anotações.

Imagem do Utilizador : O jogador irá ver-se na imagem, e terá 5 círculos vermelhos onde pode tocar. Cada um deles tem um objectivo distinto. Os círculos inferiores servem para rodar as imagens para a direita e para a esquerda, enquanto que os círculos de cima servem para rodar as anotações para a direita e para a esquerda. O círculo em cima do utilizador serve para anotar a palavra que está ao centro na imagem que também se encontra no centro.

4. De que modo utiliza as imagens guardadas no seu computador ?

Para pesquisa/trabalho

Para recordar com amigos

Para colocar em blogues

Outro(s) : _____

5. Quando pretende pesquisar as suas fotos pessoais em formato digital o que costuma fazer ?

Motivação - Assinale o número que corresponde melhor à sua resposta, sendo o mais objectivo possível.

1. É simples aprender a utilizar esta aplicação

1 2 3 4 5

Discordo totalmente

Concordo totalmente

2. É simples usar esta aplicação

1 2 3 4 5

Discordo totalmente

Concordo totalmente

3. É divertido utilizar esta aplicação

1 2 3 4 5

Discordo totalmente

Concordo totalmente

4. Utilizaria esta aplicação para anotar as minhas imagens

1 2 3 4 5

Discordo totalmente

Concordo totalmente

5. Usaria esta aplicação num sitio público para passar o tempo (aeroporto, cinema, hospital, etc.)

1 2 3 4 5

Discordo totalmente

Concordo totalmente

6. Utilizaria esta aplicação para me divertir com amigos/família

1 2 3 4 5

Discordo totalmente

Concordo totalmente

Dinâmica do jogo - Assinale o número que corresponde melhor à sua resposta, sendo o mais objectivo possível.

1. Consigo perceber como a pontuação vai mudando ao longo do tempo

1 2 3 4 5

Discordo totalmente

Concordo totalmente

2. Percebi que estava a fazer boas ou más anotações

1 2 3 4 5

Discordo totalmente

Concordo totalmente

3. As imagens deveriam estar paradas, apenas as anotações deveriam rodar

1 2 3 4 5

Discordo totalmente

Concordo totalmente

4. As anotações deveriam estar paradas, apenas as imagens deveriam rodar

1 2 3 4 5

Discordo totalmente

Concordo totalmente

5. Seria mais divertido usar imagens minhas com as minhas próprias anotações

1 2 3 4 5

Discordo totalmente

Concordo totalmente

7. A aplicação seria mais fácil/intuitiva se usasse teclado / rato

1 2 3 4 5

Discordo totalmente

Concordo totalmente

8. A aplicação funcionaria melhor com mais imagens

1 2 3 4 5

Discordo totalmente

Concordo totalmente

9. A aplicação funcionaria melhor com mais anotações

1 2 3 4 5

Discordo totalmente

Concordo totalmente

10. Quais as principais alterações que faria à interface em termos de dinâmica de jogo (objectos no jogo, pontuações, etc.) ?

Interacção - Assinale o número que corresponde melhor à sua resposta

1. É fácil manejar os “hotspots” que rodam imagens/conceitos

1 2 3 4 5

Discordo totalmente

Concordo totalmente

2. Usar este tipo de interacção é fisicamente desgastante

1 2 3 4 5

Discordo totalmente

Concordo totalmente

3. Usar este tipo de interacção é mentalmente desgastante

1 2 3 4 5

Discordo totalmente

Concordo totalmente

4. A imagem que mostra o utilizador/hotspots é pequena demais

1 2 3 4 5

Discordo totalmente

Concordo totalmente

Estética - Assinale o número que corresponde melhor à sua resposta

1. O aspecto estético da interface agrada-me

1 2 3 4 5

Discordo totalmente

Concordo totalmente

2. Considero, em termos gerais, uma interface agradável

1 2 3 4 5

Discordo totalmente

Concordo totalmente

3. Utilizaria esta interface para uso pessoal

1

2

3

4

5

Discordo totalmente

Concordo totalmente

4. Em termos estéticos, quais as principais alterações que faria à interface ?

5. Em termos gerais, qual a sua opinião desta interface ?

II. Results

Tester	Info			1 - General				2 - Motivational					
	Age	Sex	IT	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.5	2.6
1	18	F	N	4	4	3	1,2	4	5	4	3	4	4
2	25	M	S	5	4	4	2	4	4	5	4	5	5
3	19	F	N	5	3	2	1,2	3	4	4	3	3	4
4	25	F	S	4	5	4	1,2,3	5	4	5	4	5	5
5	24	F	N	5	5	5	1,2	5	5	5	5	5	5
6	24	F	S	5	5	4	1,2	5	5	5	5	5	5
7	24	F	N	5	5	5	1,2,3	4	5	5	4	5	5
8	26	F	S	5	3	3	1,2	4	4	5	4	4	5
9	27	M	S	4	2	2	1,2	5	4	3	2	5	5
10	31	F	F	4	5	1	1,2	4	4	4	2	5	3
11	25	M	S	5	2	1	1,2	4	3	5	4	5	5
12	24	M	S	4	4	4	1,2	4	4	4	2	4	4
13	26	M	S	5	5	3	2	5	5	4	3	3	3
14	23	M	S	4	2	2	1	4	5	5	4	3	4
15	19	M	S	5	3	2	2	4	4	4	4	5	5
Average	24			4.60	3.80	3.00	1.75	4.27	4.33	4.47	3.53	4.40	4.47
SD				0.51	1.21	1.31	0.50	0.59	0.62	0.64	0.99	0.83	0.74

Tester	3 - Game dynamics							
	3.1	3.2	3.3	3.4	3.5	3.7	3.8	3.9
1	5	4	2	2	4	3	3	3
2	3	5	1	1	4	3	3	3
3	4	4	2	2	3	2	3	4
4	1	5	1	1	5	4	3	3
5	3	4	1	1	5	1	3	3
6	4	4	3	3	4	1	3	3
7	5	4	1	1	4	1	4	2
8	3	4	1	1	3	1	2	3
9	4	2	1	1	2	5	2	2
10	1	5	1	1	5	3	2	2
11	2	5	1	1	5	3	3	3
12	2	5	1	1	5	4	3	4
13	4	3	1	1	4	2	3	5
14	4	3	2	2	2	4	4	3
15	3	4	3	3	4	1	5	2
Average	3.20	4.07	1.47	1.47	3.93	2.53	3.07	3.00
SD	1.26	0.88	0.74	0.74	1.03	1.36	0.80	0.85

Tester	4 - Interaction				5 - Aesthetic		
	4.1	4.2	4.3	4.4	5.1	5.2	5.3
1	4	3	3	2	4	4	3
2	4	1	1	2	4	4	4
3	2	4	3	2	4	4	4
4	2	2	2	4	3	4	5
5	3	1	1	3	4	5	5
6	5	1	1	1	5	5	5
7	4	2	2	2	4	4	5
8	3	1	2	2	3	4	5
9	2	1	1	3	3	3	2
10	4	5	3	1	3	4	2
11	2	2	1	2	4	5	5
12	4	1	1	3	3	4	4
13	4	3	1	2	4	4	4
14	5	2	1	2	5	5	5
15	2	3	1	5	4	4	5
Average	3.33	2.13	1.60	2.40	3.80	4.20	4.20
SD	1.11	1.25	0.83	1.06	0.68	0.56	1.08

III. Performance evaluation

Tester	Age	Score	Time	Level
1	18	360	3:20	4
2	25	1347	5:13	6
3	19	1045	4:25	6
4	25	263	2:18	3
5	24	960	4:13	6
6	24	1195	4:12	6
7	24	1385	4:39	6
8	26	1563	4:30	6
9	27	680	3:55	6
10	31	2159	5:12	6
11	25	1649	5:20	6
12	24	2511	6:00	6
13	26	2474	5:50	6
14	23	1647	5:00	6
15	19	898	3:40	4

IV. Usability tests – Pie charts

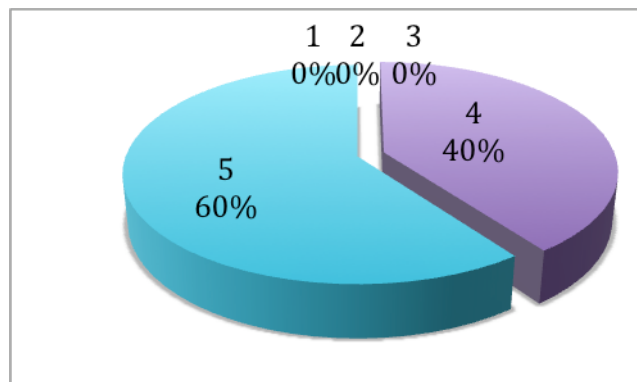
Geral

1. Costuma utilizar a internet para fazer pesquisas de imagens ?

1 2 3 4 5

Raramente

Muitas vezes

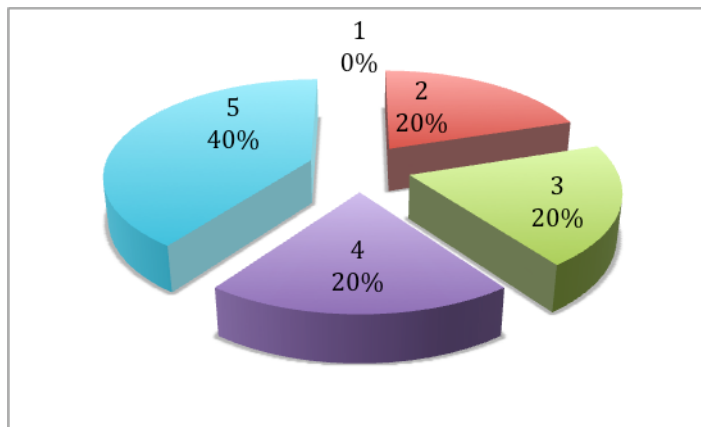


2. Costuma organizar imagens pessoais no seu computador ?

1 2 3 4 5

Raramente

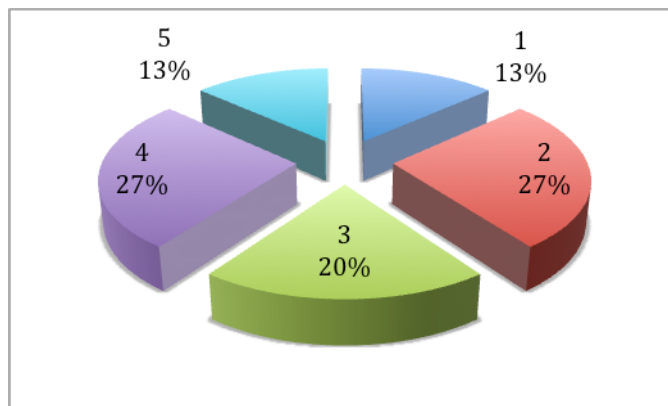
Muitas vezes



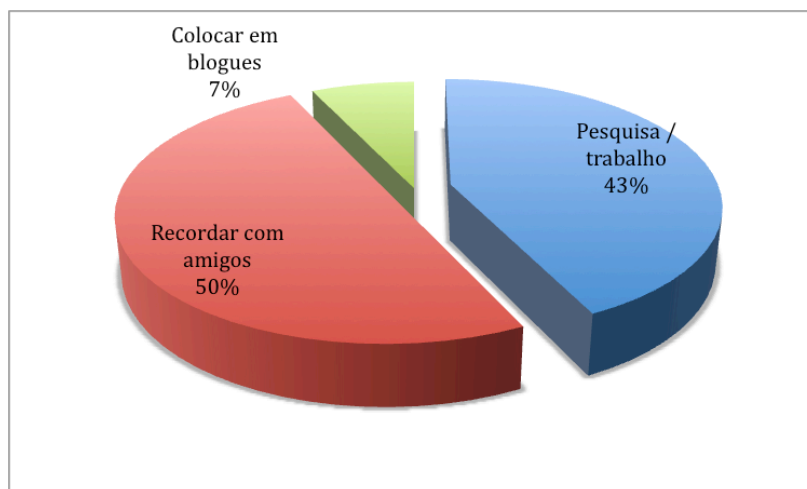
3.As suas imagens pessoais/pesquisadas estão catalogadas ?

1 2 3 4 5

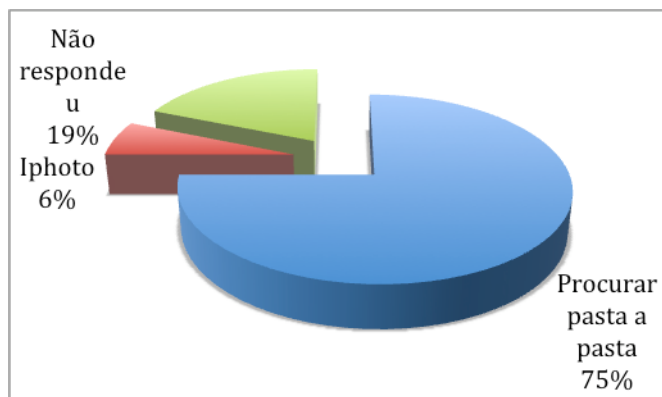
Nenhumas Todas



4.De que modo utiliza as imagens guardadas no seu computador ?



5. Quando pretende pesquisar as suas fotos pessoais em formato digital o que costuma fazer ?



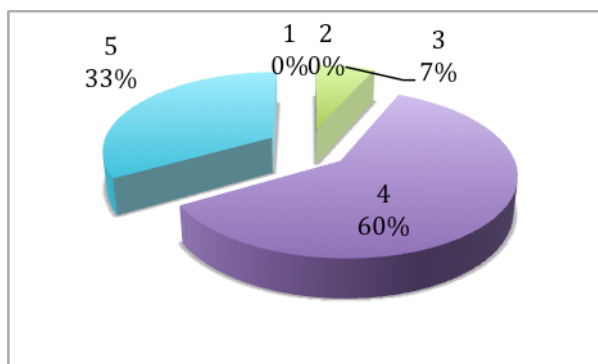
Motivação - Assinale o número que corresponde melhor à sua resposta, sendo o mais objectivo possível.

1. É simples aprender a utilizar esta aplicação

1 2 3 4 5

Discordo totalmente

Concordo totalmente



2. É simples usar esta aplicação

1

2

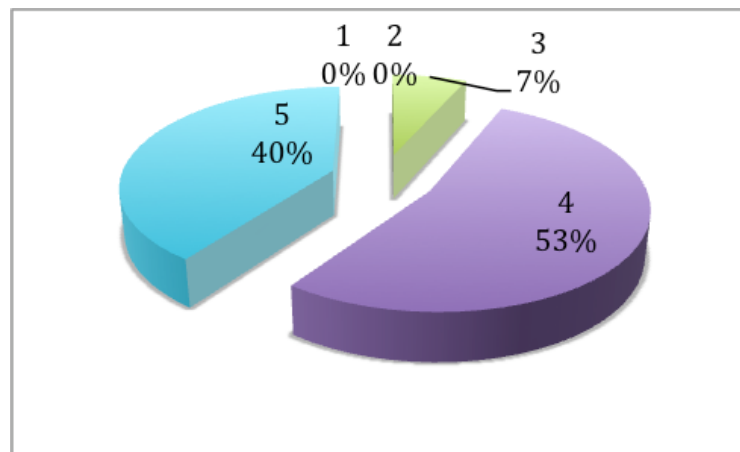
3

4

5

Discordo totalmente

Concordo totalmente



3. É divertido utilizar esta aplicação

1

2

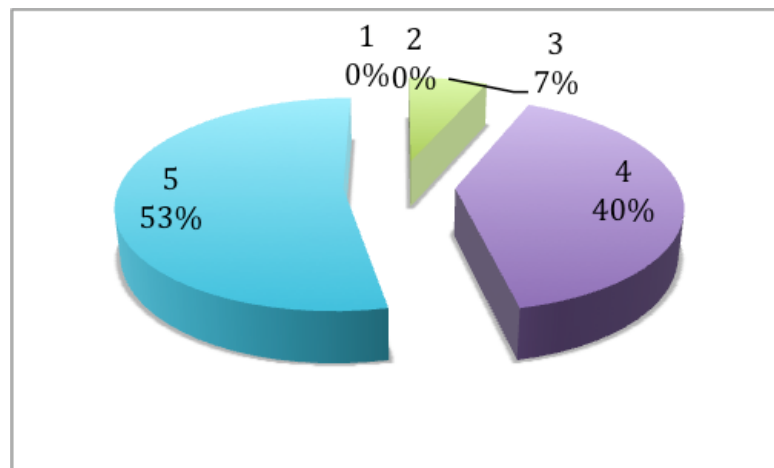
3

4

5

Discordo totalmente

Concordo totalmente

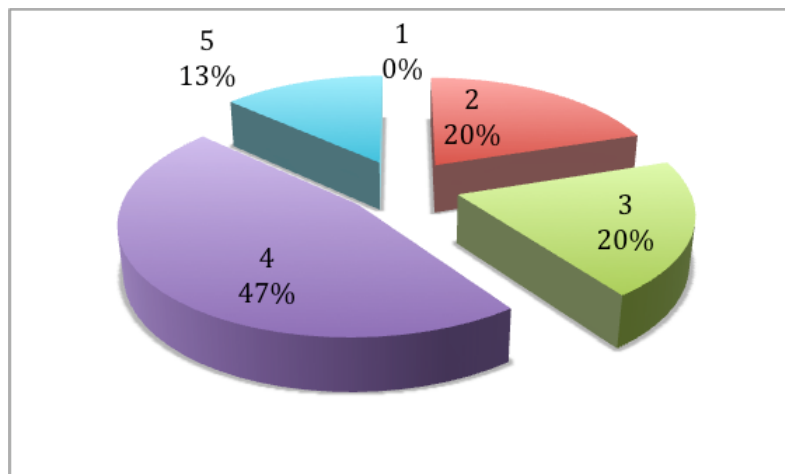


4. Utilizaria esta aplicação para anotar as minhas imagens

1 2 3 4 5

Discordo totalmente

Concordo totalmente

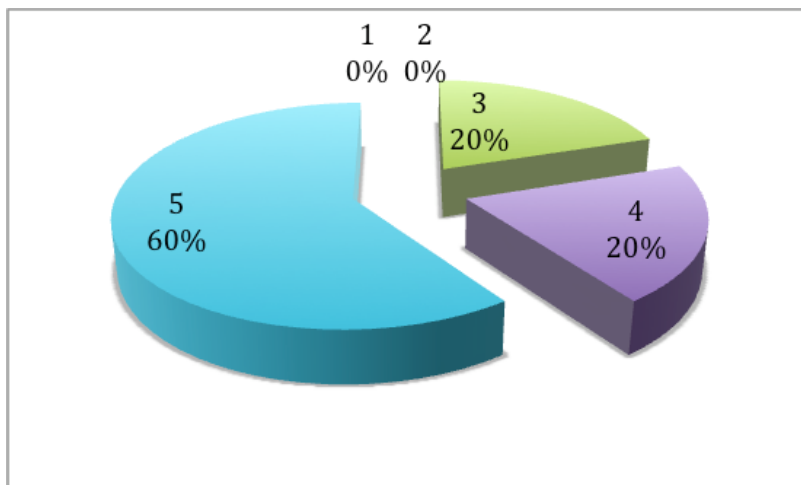


5. Usaria esta aplicação num sitio público para passar o tempo (aeroporto, cinema, hospital, etc.)

1 2 3 4 5

Discordo totalmente

Concordo totalmente

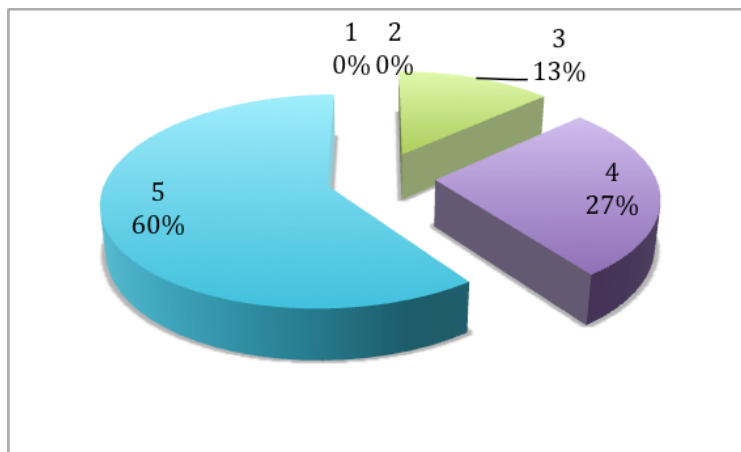


6. Utilizaria esta aplicação para me divertir com amigos/família

1 2 3 4 5

Discordo totalmente

Concordo totalmente



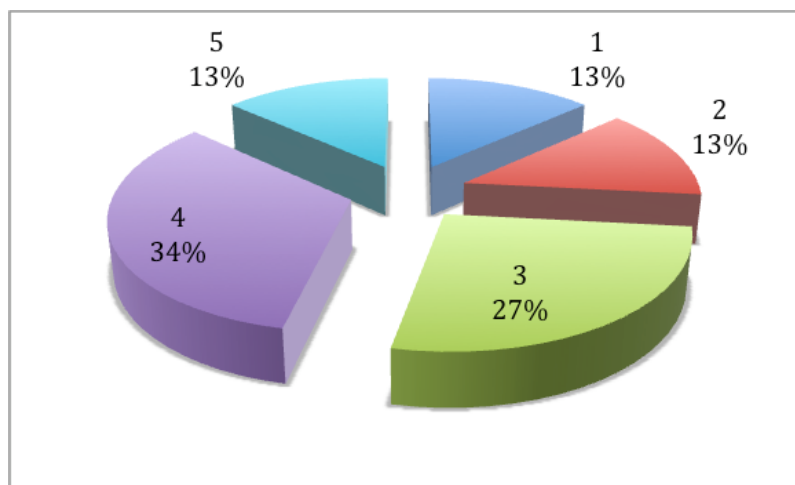
Dinâmica do jogo

1. Consigo perceber como a pontuação vai mudando ao longo do tempo

1 2 3 4 5

Discordo totalmente

Concordo totalmente



2. Percebi que estava a fazer boas ou más anotações

1

2

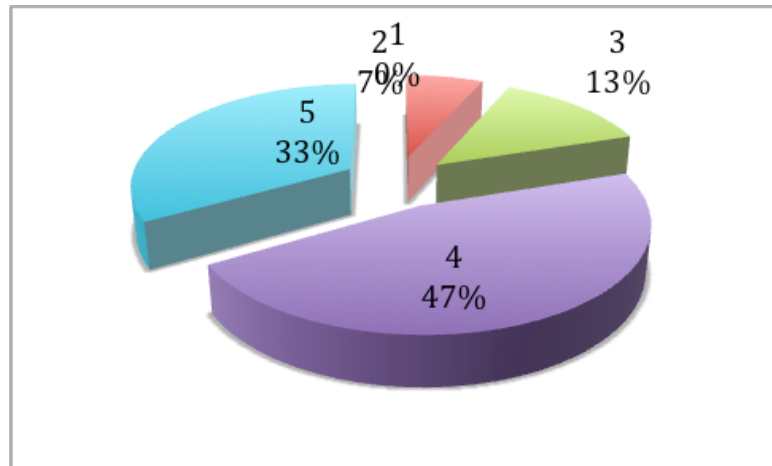
3

4

5

Discordo totalmente

Concordo totalmente



3. As imagens deveriam estar paradas, apenas as anotações deveriam rodar

1

2

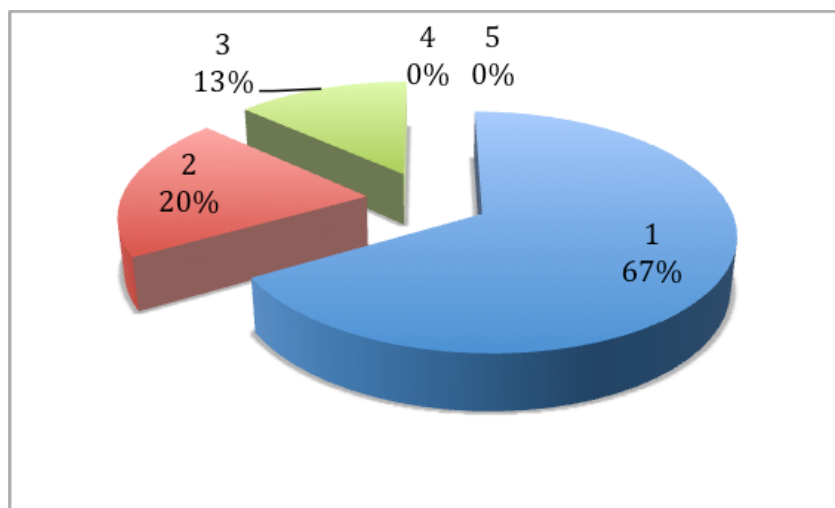
3

4

5

Discordo totalmente

Concordo totalmente

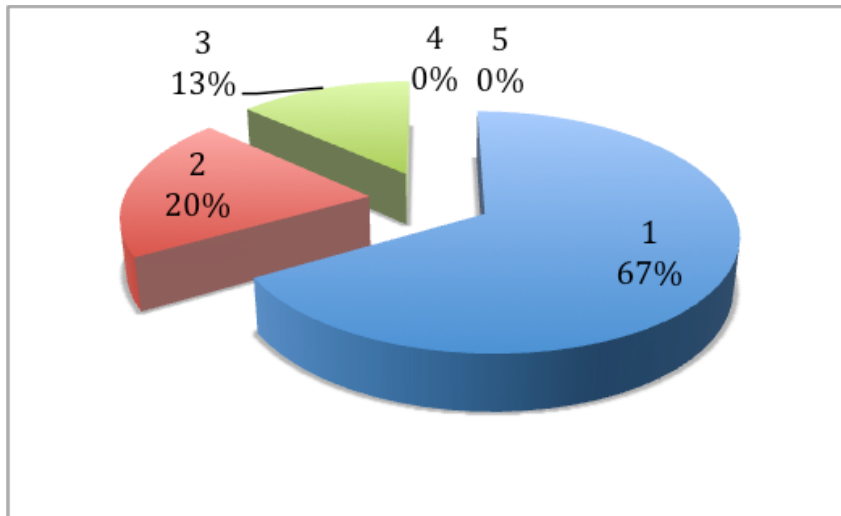


4. As anotações deveriam estar paradas, apenas as imagens deveriam rodar

1 2 3 4 5

Discordo totalmente

Concordo totalmente

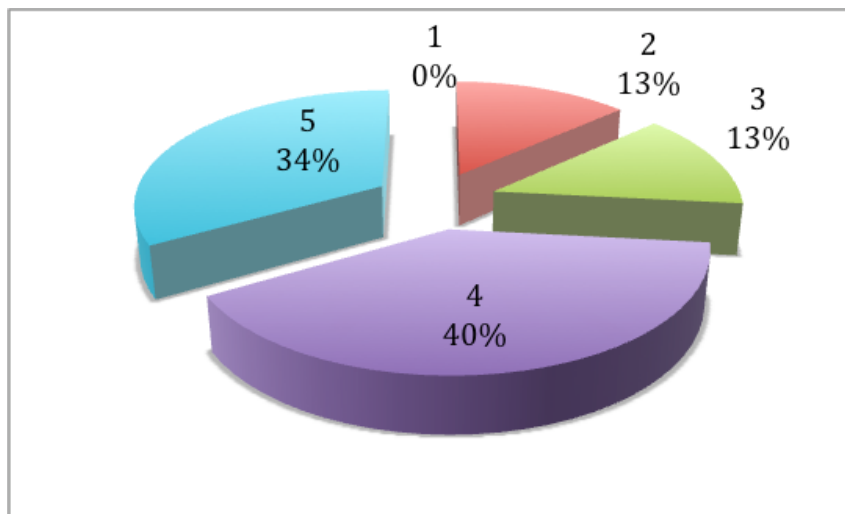


5. Seria mais divertido usar imagens minhas com as minhas próprias anotações

1 2 3 4 5

Discordo totalmente

Concordo totalmente

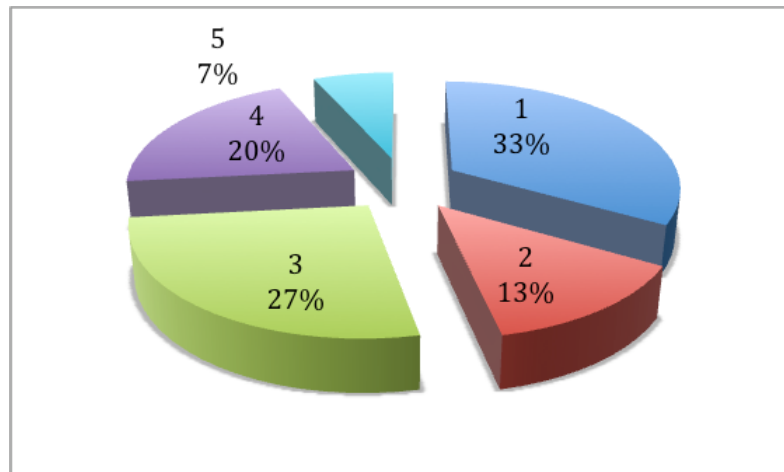


7. A aplicação seria mais fácil/intuitiva se usasse teclado / rato

1 2 3 4 5

Discordo totalmente

Concordo totalmente

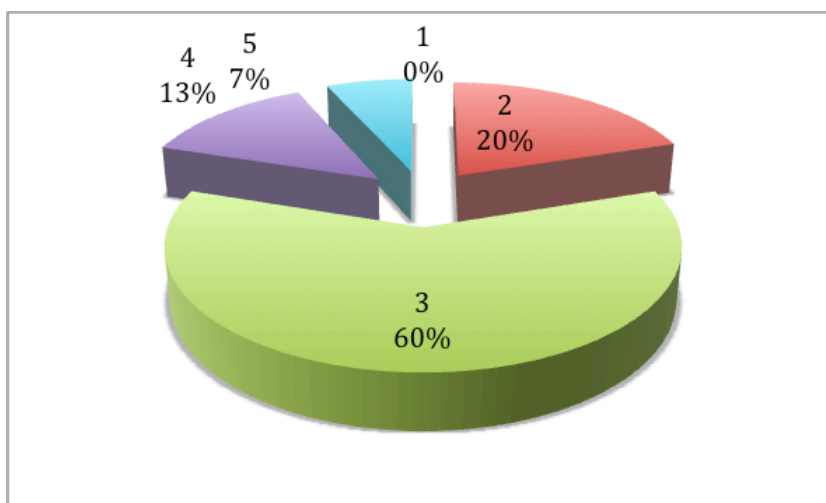


8. A aplicação funcionaria melhor com mais imagens

1 2 3 4 5

Discordo totalmente

Concordo totalmente



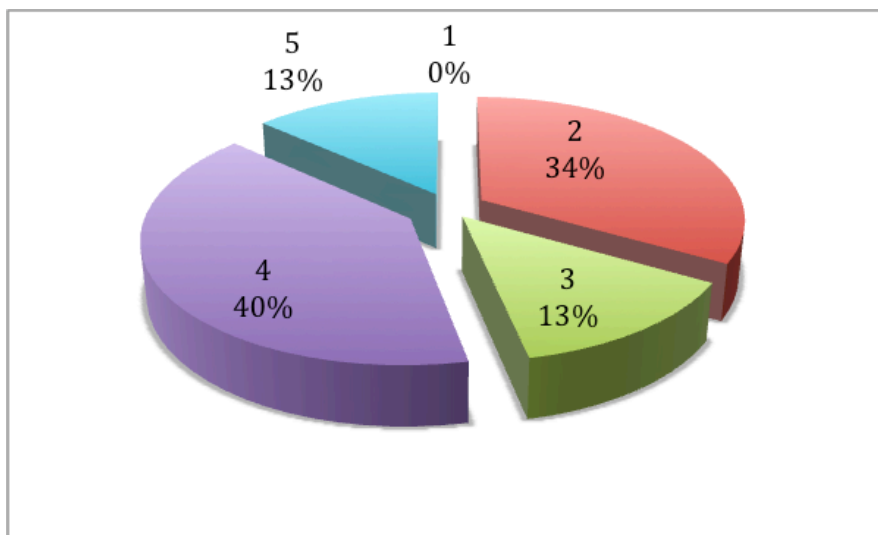
Interacção

1. É fácil manejar os “hotspots” que rodam imagens/conceitos

1 2 3 4 5

Discordo totalmente

Concordo totalmente

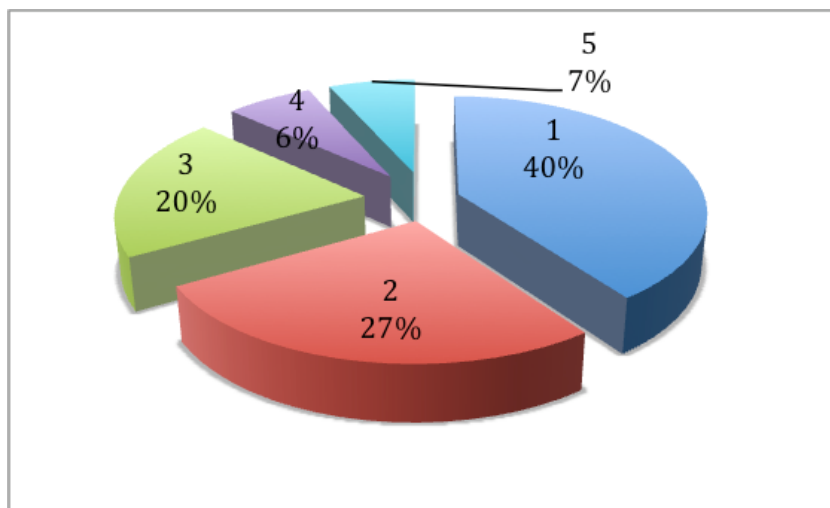


2. Usar este tipo de interacção é fisicamente desgastante

1 2 3 4 5

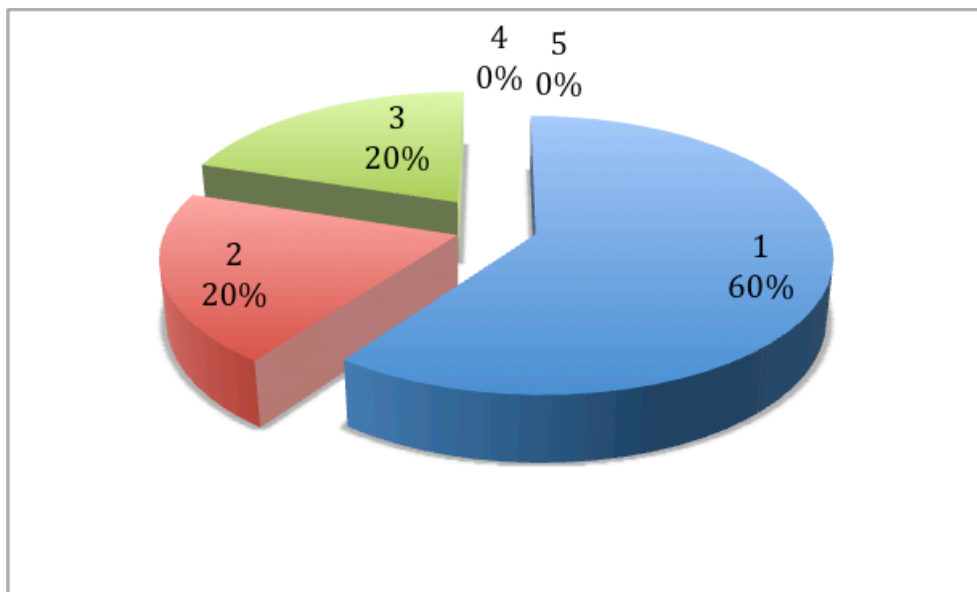
Discordo totalmente

Concordo totalmente



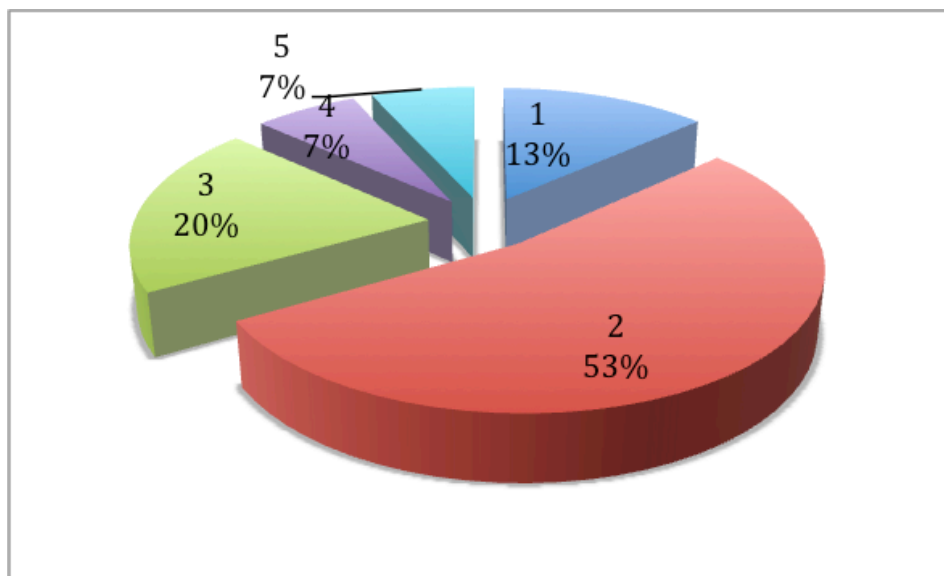
3. Usar este tipo de interacção é mentalmente desgastante

1 2 3 4 5
Discordo totalmente Concordo totalmente



4. A imagem que mostra o utilizador/hotspots é pequena demais

1 2 3 4 5
Discordo totalmente Concordo totalmente



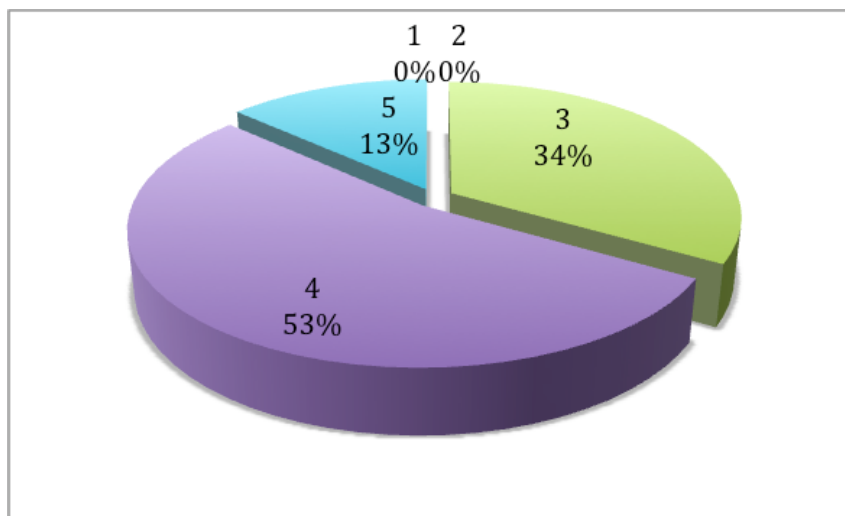
Estética

1. O aspecto estético da interface agrada-me

1 2 3 4 5

Discordo totalmente

Concordo totalmente

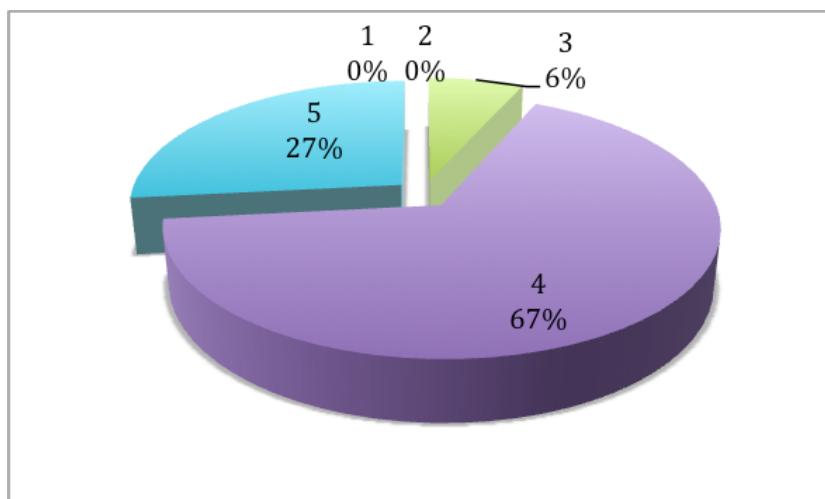


2. Considero, em termos gerais, uma interface agradável

1 2 3 4 5

Discordo totalmente

Concordo totalmente



3. Utilizaria esta interface para uso pessoal

1

2

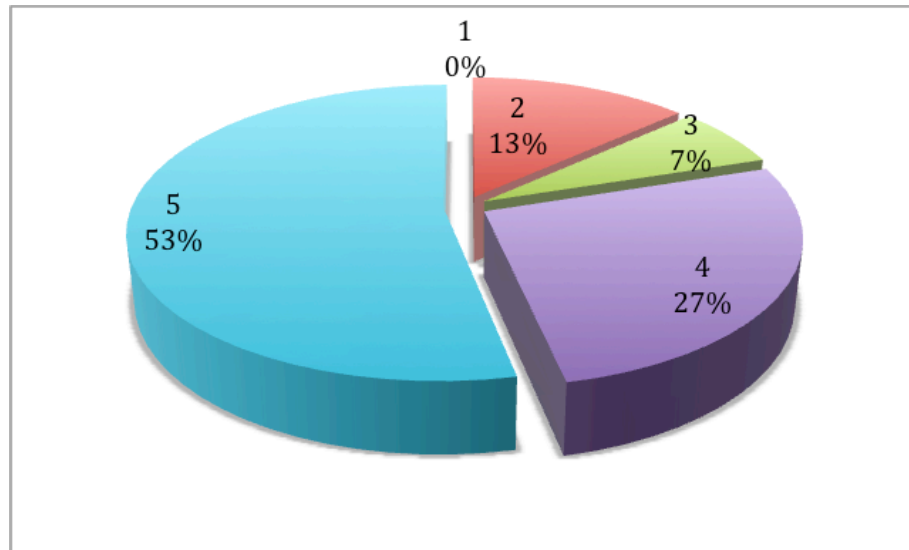
3

4

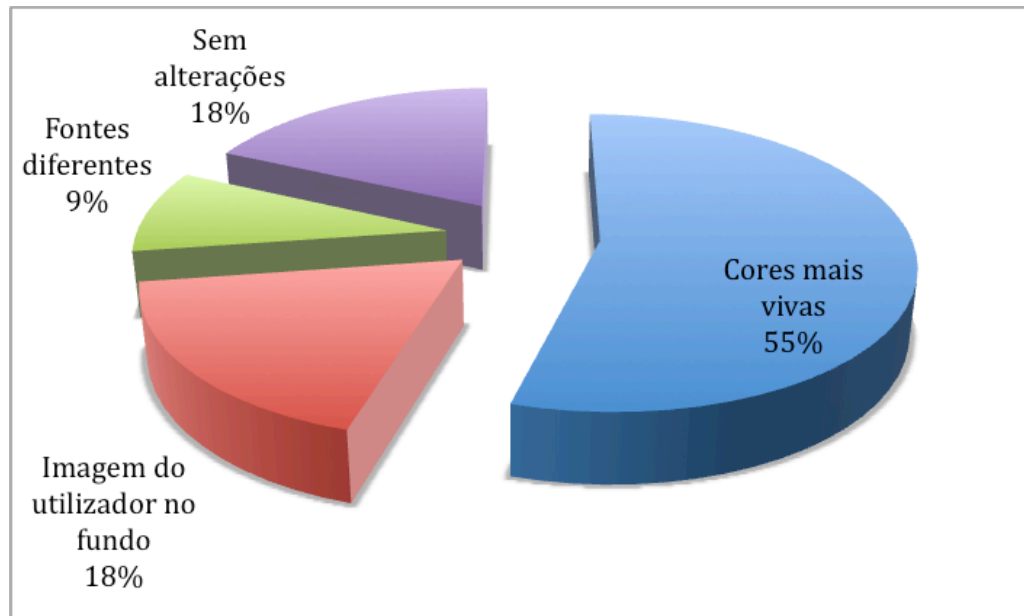
5

Discordo totalmente

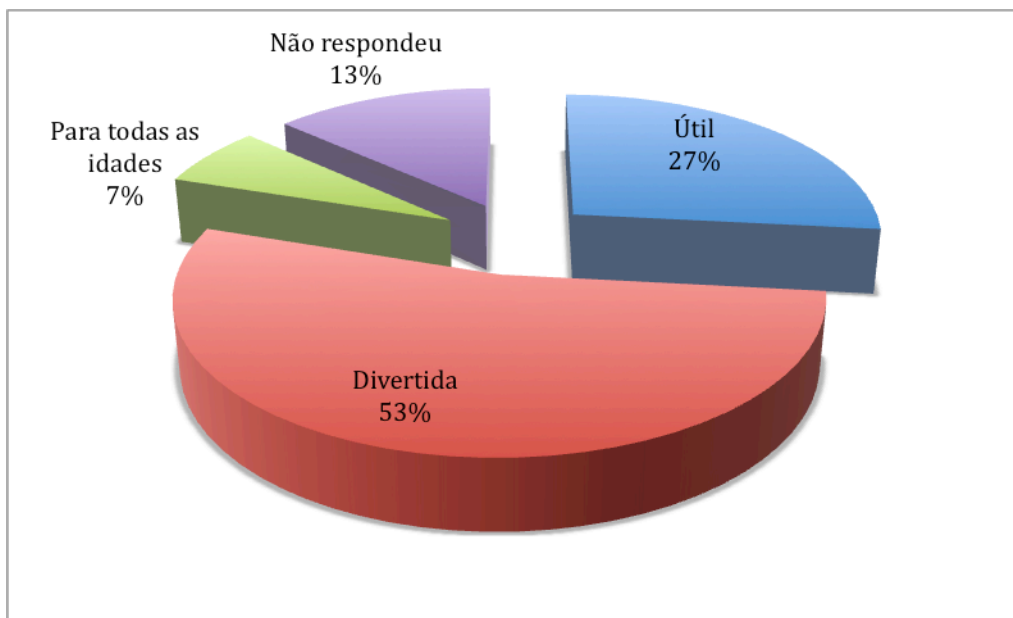
Concordo totalmente



4. Em termos estéticos, quais as principais alterações que faria à interface ?



5. Em termos gerais, qual a sua opinião desta interface ?



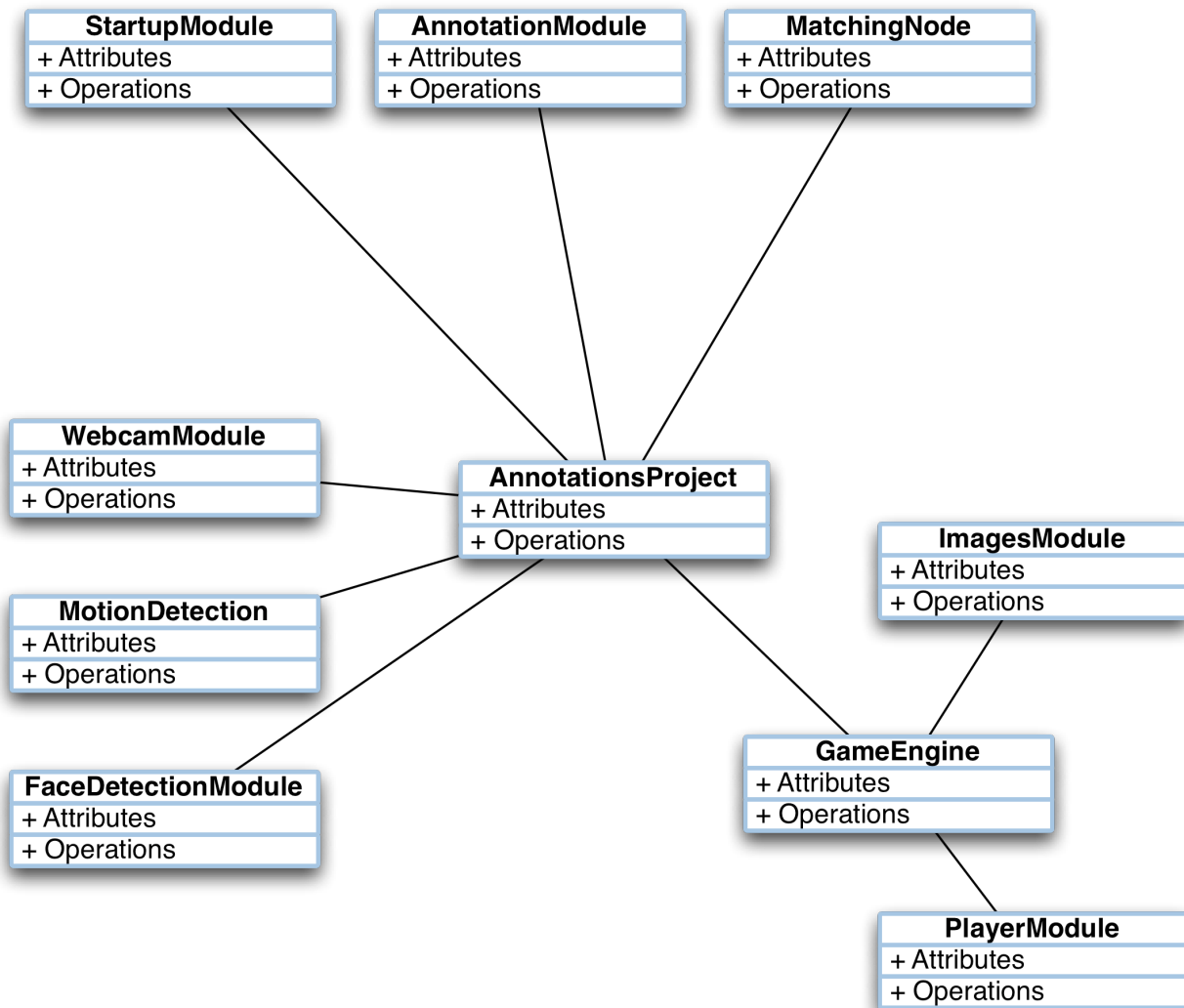
Appendix B

Class diagram

I. Class Diagram

Tag Around main classes

These are the main classes that define the Tag Around Application. It includes the game engine class, the motion detection classes and the interface components classes.



Other Tag Around classes

These classes represent the secondary classes that compose Tag Around Application. TimerManager is the class responsible by all the timers in the game and SoundManager for all the sounds in the game.

