Building interactive spatial and temporal models using multimodal data

Rui Nóbrega
Nova School of Science and Technology

Building interactive spatial and temporal models using multimodal data

Doctoral Program in Computer Science Thesis Plan

CITI, Departamento de Informática
Faculdade de Ciências e Tecnologia, FCT
Universidade Nova de Lisboa 2829-516 Caparica, Portugal

Rui Pedro da Silva Nóbrega
rui.nobrega@gmail.com

Supervisor: Prof. Nuno Manuel Robalo Correia

October 30, 2009

Abstract

The construction of virtual world models is an important step in several applications from simulations to virtual reality. These three-dimensional worlds are usually constructed offline from scratch or using large datasets of data extracted from satellites. Once they are constructed they are usually static or hard to alter automatically. What is proposed in this thesis is a method to create methods that would take advantage of several types of low-cost sensors and computer vision to quickly create and update virtual models. These virtual models will have live feeds from the sensors and from images and videos. The resulting model can be consulted in terms of space and time, turning on and off several layers of information. These models can then be used in several scenarios such as retrieving information from a physical space, creating simulations or creating augmented reality scenes. One of the main ideas proposed is the fact that changes in the virtual model can be reflected back to the images and videos that helped in its construction. This reconstruction of reality can have several applications to visualize the results of environmental simulations such as pollution or disaster simulations, showing the affected areas in real life images. Additionally it can be useful to superimpose virtual objects to be used in augmented reality, turning the room where the user is into a completely different scenario. The construction of the models involves the study of several image processing algorithms and techniques. Additionally, to support the fast creation, visualization and interaction with the models several tools will have to be developed. The interaction should explore new paradigms different from the mouse and keyboard. Essentially it should take advantage of the computer vision knowledge learnt in the construction of the models. This document addresses the problems involved in this area, presents related work, preliminary solutions and a work plan for the thesis.

1 Introduction

The current information systems are usually unaware of the physical space where they are in. The construction of virtual worlds and models is usually done manually with 3D editors, with 3D scanning methods or using data from satellites. These methods are not usually simple for a common user. One low cost alternative is to use computer vision, image processing and sensors to build these models. Using these methods the computer systems will be more aware of a certain physical space and can interact better with it. With this context awareness it is possible to bridge the gap between the real world and the virtual one. Augmented reality, embedded interactions where part of the actions happen in simulations while controlled with real objects, these are results that we can expect when the information systems are aware of their context.

Computers have for a long time been a black box deprived of sensorial capabilities. The common human interface present in most computers is the keyboard and mouse input. It is time to add new senses such as vision and touch that can give more context to the information system. These are essentially raster sensors where a grid of information has to be interpreted to retrieve some knowledge.

Most computers today are equipped with web cameras and allow the connection of several other devices using USB ports. Using these devices it is possible to explore the world from another perspective. Instead of doing actions in the real world, why not transpose the real world to the virtual world and do endless inexpensive changes there? This might be an ambitious goal in a generic situation but transposing this objective to specific situations may turn this into a feasible approach. As an example, of things that are possible we can track people inside a space, monitor the level of a dam, repaint a room or create 3D representations of objects.

One of the main tasks in this thesis is to evaluate in what situations is it possible to create, maintain or update a virtual model. It must be noticed that the models must already exist in some form but should be prepared to receive input from several sources of video, image or other media (i.e. analogic or digital sensors). Having the models built it would be interesting to explore the information from several dimensions such as time, space, color, concepts, affinity and others.

The topics explored in this work are computer vision, electronic sensor integration, computer graphics visualization, simulations and interaction. These are research areas that have been the center of attention lately in several projects (see section 2.3) and have some integration potential. This can be achieved through several multimedia appliances that present the information in innovative forms and use sensors and cameras to receive input and control.

1.1 Problem Statement

The main problem addressed in this work will be how to create spatial and temporal models enriched with large sets of indexed information from images, videos and sensors while providing an intelligible and interactive form of data manipulation.

The spatial and temporal model should be prepared to be useful in visualization and simulation systems. The acquisition of the multimedia data should be done in an almost automatic fashion using image and video databases, computer vision, sensor data and geographic systems. The model should be prepared for retrieving information from the multimedia data but also some of the changes to the model due to user interaction or a simulation should be reflected back to the image and video data. The investigation on the interaction level should explore the affordances of the spatial model in order to reach an innovative approach based on gestures, haptics or multi-touch tables.

Taking into consideration the above premises it is possible to consider three smaller problems: (a) Acquisition, (b) Model Creation, (c) Visualization and Interaction.

In the acquisition (a) problem several data is collected from a chosen physical location. Possible locations can be a room, a building or a large terrain area. The study will essentially focus on the feasibility of several sources of information. Possible candidates are pictures or videos taken from large internet databases or from cameras. Also information from ultra-sound or infra-red sensors will be considerated.

To create the spatial model (b) several known computer vision algorithms will be integrated and adapted. Part of the goal is to understand what kinds of algorithms are useful for model parameterization, including possible simulation parameterization. Additionally there will be research on innovative ways to interconnect data with the information sources. In other words, how to change the original images and videos with the results of the evolution of a simulation model?

Finally there will be the need to build tools to (c) build, visualize and interact with the data model. Although this is not a primary objective it is an essential part of the workplan.

These are challenging goals which touch several research areas such as computer vision, information retrieval, simulation modeling and human-computer interaction. The main novelty here is the two-way connection between the computer vision data and the model. Changes in the model can change real data and vice-versa.

Additional details will be given in section 3. The research context will be presented in section 2.

1.2 Main Goals

One of the main outcomes of this thesis will be a series of software and physical prototypes. These initial proposals are described in section 3.4 and include an automatic creation of a virtualized room, automatic parameterization of the model of an emergency simulation, exploration of a building through a collection of images and time and space exploration of images and videos in a geographical model.

Furthermore it is the author's objective to publish his work, user studies and analysis in international conferences such as the ones organized by ACM, IEEE or Springer including CHI, Interact, Siggraph, Eurographics, ACM Multimedia, UIST or ACE.

1.3 Document Organization

This document is organized as follows: the next section provides a research context describing the topics involved in this thesis and giving some background and related work. In the third section, the research statement is explained in more detail with some attention to the concrete problems addressed and possible solution paths. Section four presents a road map for the work to be done. In the end, conclusions and future directions are presented.

2 Research Context

In this section, the authors, notions and concepts that are more relevant in the context of this thesis will be presented.

In general, the topics addressed in this thesis will be related with multimedia from a computer science point of view. Multimedia consists of, as described by Li et. al. [24], "applications that use multiple modalities to their advantage". Although this is not a closed definition, it is an idea of convergence of all media in one single experiment. Multimedia usually also implies some level of interaction which can be achieved using several input devices or sensors.

Visualization and interaction are two topics that are closely related in multimedia. Authors such as Ben Shneiderman [38] explored new forms of presenting information in a visual form. At the same time computer graphics started to be used to synthesize reality, in an attempt to recreate the world inside the computer. This was an effort that received a large contribution from the entertainment industry. One of the main references in this well established area is the work done by Foley et al. [11]. With the increasing attention given to graphical user interfaces (GUI) new questions were considered. How do users perceive them? Do they use them as expected? And if not who is to blame, the user or the interface? Ultimately this led to the study on how computers and humans interact.

Alan Dix in his book [8] defines several topics on Human-Computer Interaction (HCI) presents methods for building and testing interfaces. One important definition in HCI is the notion of affordances. In this work, the definition of affordance that will be used will be the properties of a certain interface, object or environment that suggest and lead the user to perform a certain action. To imagine a new form of interaction one must be aware of the affordances of the physical and digital space that surround the user.

In this work one of the main goals is to increase the knowledge of the computational systems by using a range of sensors and computer vision techniques. This is actually a way to overcome a well known problem generally known as "Semantic Gap". This problem is usually associated with the conversion of analogical information into computational digital information, and is usually associated with the information retrieval area. The problem can be further subdivided in three sub problems: Sensorial, Semantic and Subjective gaps. When the data is collected from the sensors and is transformed into digital information using, for example, quantization what happens is that there is a Sensorial gap where information may be lost due to the resolution, quality or captured noise of the sensor. In the same way when the raw data obtained by the sensors is converted into semantic concepts, there is a Semantic gap. Finally a Subjective gap may occur when the concepts are related and interpreted according to the logic of a given application to give some kind of result.

Taking into consideration the main goals defined in section 1.3 and in the problem statement we can categorize the construction and tuning of the sensors as related to the Sensorial Gap. The automatic model construction and parameter acquisition are related to the Semantic Gap part. The final simulation, visualization and interaction are applications have implications in the Subjective gap.

2.1 Background Work

In order to achieve the goals proposed there is previous background knowledge and experience that supports the development of this research.

The PhD plan proposed here has some relations and contact points with the master thesis [28] developed in the context of the project Life-Saver [29]. This project was created in the Portuguese context of Dam Break Emergency Management (DBEM) and is basically a flood emergency plan simulation system. My main contribution for this project was the proposal and study of an innovative visualization and interaction system.

The objective of this project is to develop a system that can effectively validate existing emergency plans, through simulation of a flooding and of all the actors intervening in the situation. This information is handled by an agent-based simulation engine, which automatically feeds relevant spatial information to the visual interface component as needed. The simulator is able to define emergency scenarios which will include available DBEM (Dam Break Emergency Management) resources, actors and roles. The system dynamics are visualized and manipulated with a graphical interface representing the emergency scenario, while parameters characterizing this dynamics are registered during the simulation period, for later analysis.

A case study scenario was created in order to test the interaction and visualization system. The simulator runs an evacuation plan from the Portuguese authorities where the Alqueva dam collapses creating a flood. The data was collected by LNEC (Portuguese National Civil Engineering Laboratory) and includes a full topographic characterization of the valley as well as a listing of all important buildings and infrastructures of the area. This data collecting process involved several teams from different entities in several projects. The data collection period took several years and the latest data is now 10 years old, when the Alqueva Dam was not even built. This is one of the main problems and inspirations for this work; any simulation with this type of data is outdated.

A model that could collect and update data from images and cameras from the valley would improve this simulation. This will be one of the prototype case studies explained later in section 3.1.

2.1.1 Interaction Studies

The final prototype of the Life-Saver project featured a three-dimensional application with support for multi-touch interaction. To support this type of interaction two multi-touch tables where constructed. The first one is very simple and is based on detecting the shadows of the hand. The last is still in construction but is based on the Frustrated Total Internal Reflection (FTIR) approach developed by Jeff Han [16] and on a Reactable [20] like setup. The FTIR technique consists on using a display made of acrylic with some diffuser material in one of the sides and rear project the images on it. On the edge of the acrylic are placed several infra-red LEDs, and when the user touches the acrylic screen the light is reflected. An infrared camera on the back of the display captures the light. The captured image will have several points or blobs that can be computer tracked and used as input for applications.

Several new techniques appeared in multi-touch such as finger detection using planar laser reflection or using silicone solutions to intensify the infrared reflection. The NUI group [15] is the leading source in multi-touch related discussion, technologies and solutions.

Most multi-touch or multi-point technologies presented in these early examples rely on infra-red cameras to detect fingers or visual symbols. This actually transforms the detection problem into a computer vision problem.

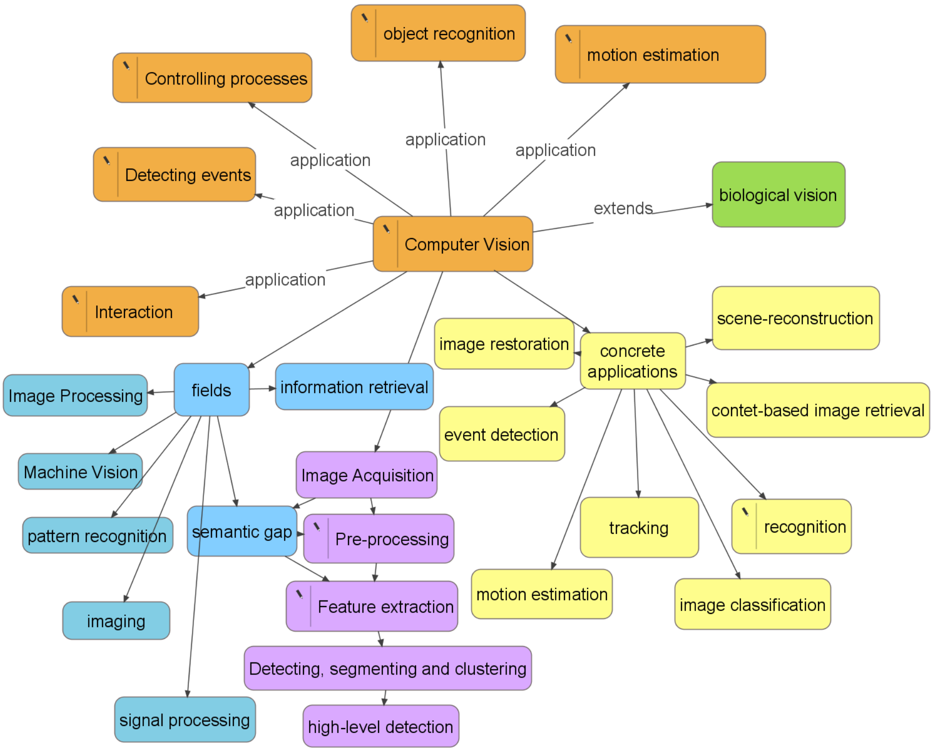

Figure 1: Computer Vision, aplications and research fields.

Instead of using a specialized sensor to detect the position of the fingers, the multi-touch tables optimize the image capture and transfer the rest of the problem to the software. Most of the work of the table is to reduce the visual noise, eliminating unwanted sources of light and highlighting the fingers that are touching or near the surface. The camera captures a raster image that can be processed to retrieve the desired information. Using computer vision it is possible to create rich interactions with applications; the next section will detail some of the possibilities regarding the use of cameras.

2.2 Computer Vision

The ultimate goal of computer vision is to replicate and enhance the human vision and the capacity of visual recognition and perception of the world around us. As presented in figure 1, computer vision has many applications and is known by several names depending of the area of application.

There are two main stages in computer vision. The first is image acquisition with all the problems that come from signal noise, weak illumination, distortion, weak signal, media conversion and compression. The second stage corresponds to the application of the algorithms to accomplish the desired effect. It is then fundamental to work in controlled environments whenever this is possible. This involves preparing the physical scenario filtering most of the noise using hardware filters to spare the computational power of software filters. To process images one of the most common libraries used in computer vision is OpenCV [31]. This is a library with the most common algorithms for face and body recognition, pattern matching, blob tracking, motion detection and many others.

For a deep understanding of how most algorithms work there is extensive literature [9] on the subject. In their book, Forsyth and Ponce [12] illustrate most of the common techniques in Computer Vision. From the understanding of how do cameras and the human eye work to color models, filters, edge and texture detection, tracking, classifiers or recognition. Klette et al. [21] specialize in translation from 2D images to three-dimensional data using stereo analysis, geometric object models and studying shapes. Another contribution comes from Bhanu and Pavlidis [3] where the authors explore computer vision using light spectrums different from the visible one. Among the several experiments they use laser radar recognition, thermal infrared recognition, magnetic resonance image analysis or ultrasound images. It is important to understand that although these are different techniques of obtaining raster images the problem is the same.

2.3 Related Projects

In this section are presented some related projects which contributed to the development of the ideas proposed in this thesis.

2.3.1 Computer Vision and Sensors in Applications

Most computer vision systems and sensor applications are limited to small experiences in a lab controlled environment. In its survey Abidi et al. [1] present several methods for integrating sensors and cameras in wide area surveillance. It uses as case study a harbor security scenario, but the same conclusions are extrapolated to airports, parking lots and traffic control. Using these examples they try to track several objects, do 3D reconstruction and long-range face recognition. The most important contribution of this paper is that it summarizes several results from different authors and compares the long-range behavior of several input sensors such as radar range sensors, thermal infrared sensors, visual optical sensors and laser range sensors.

Another type of application is presented in [18] where Hsieh et al. present a system that assesses coastal cliff erosion by creating a virtual reality model of the cliff using a LIDAR. LIDAR stands for Light Detection And Ranging and is an optical technique used for 3D scanning. It works by scanning a scene with a laser in several close lines. The idea is, at each point, to measure the distance to the scene by counting the time that the light takes to travel and reflect itself. Using a LIDAR it is possible to construct a very accurate 3D scene. There are some occlusion problems, so probably the scene may have to be scanned from different points, but it is a very effective method. In the example, the several scans of the cliff are made regularly and then compared to see if it has eroded.

A well known project that takes advantage of stereo vision is the Mars Rover [25]. Maestro is a java application that lets the user explore the Martian soil using data taken from the two camera system present on the rover. This is an interesting application where computer vision is used to navigate a robot in a hostile environment.

2.3.2 Geographic Simulations

There are many examples of simulations that take advantage of geographical spatial data or explain their results using maps. The MIT Senseable [23] city lab has done extensive research on obtaining information using a wide range of sensors and produce geographical statistics and simulations. In [4] the Real Time Rome project is detailed. The project used a system from Telecom Italia for the real time evaluation of urban dynamics based on the anonymous monitoring of cell phone networks. In addition, data was supplemented based on the instantaneous location of buses and taxis using GPS. All data was then processed and updated to provide information about urban mobility in real time, from the condition of traffic to the movements of pedestrians and foreigners in the city. This way the city of Rome was monitored in real time using a variety of sensing systems.

Many geographical simulations use geo-applications to provide map support. One of the most successful is Google Earth [14], an application which has a photorealistic representation of the earth with additional spatial information data. There are many similar projects to this, such as NASA World Wind [43] open source project. This is a 3D virtual Earth which allows researchers to use NASA satellite data freely. Terravision [40] was the predecessor of these systems, and the first to provide a seamless navigation and visualization in a massively large spatial data environment. Terravision is a virtual representation of the earth based on satellite images, aerial shots, altitude data and architectural data. It serves as an environment to organize and access information spatially. Users can navigate seamlessly from overviews of the earth to extremely detailed objects and buildings. Historical aerial shots and architectural data are offered in the system; this allows users to navigate not only spatially but also through time. All data is distributed and networked and is streamed into the system according to the user needs. The system navigation is done with a large sphere referencing the globe to pilot the planet; a 3D mouse to fly around; and a touch screen to interact with objects on the virtual earth.

An interesting geographic application is presented by Girardin et al. [13] where the user related spatiotemporal data is explored to give significant information. This work tries to find where the tourists in Barcelona are spending most of the time by analyzing their digital footprint on the web. Whenever a user introduces a geo-referenced photo in Flickr related with Barcelona, the system detects it. Finally a visualization system was created that has a visual representation of the city with all the information collected from images to density of data. The idea is then to extrapolate where the tourists are heading.

2.3.3 From 2D to 3D

In an early work [39], Paul Debevec and his colleagues developed a novel approach for recognizing the polyhedral that compose an architectural photographic scene. Using their approach, it is possible to automatically build from a photo a block model of a set of buildings. In a more recent work [19] some experiments were made on how to create real-time 3D holograms. The process involves capturing a 3D representation of a person in one place and displaying it in another place using a holographic approach.

PhotoSynth [34] is a project done in cooperation by Microsoft and the University of Washington. The main idea of this application is to take several unrelated images of the same scene and present them in a 3D space. Images taken from the same position are presented overlapped together. When the user rotates the scene, the camera is shifted to the nearest image taken from that position. The main advance of this technology is the ability to stitch each photo in the right place thus creating the appearance that all images are aligned over the photographed object. Using techniques such as the ones described in the next section, PhotoSynth examines images for similarities to each other and uses that information to estimate the shape of the subject and the focal point from where each photo was taken.

Another interesting work on bringing 2D information into a 3D world is the Video Flashlights project [37]. This is a system that projects a live video texture from static and moving cameras in a static 3D world. The 3D world can be then navigated, but the textures from the geometry will be updated with the feed from the 2D cameras. This is an augmented reality solution for security and monitoring system. Typical security systems use video monitoring with a grid of 2D display, limiting the global awareness of the space. This system provides an integrated view of all cameras in the 3D environment. The user selects a spot on the map and sees all the cameras in that area mapped as textures. This requires the distortion of the 2D image to fit the 3D world exactly in the same area which is being filmed.

2.3.4 Recognition and Classification

One of the key points of this work is the detection of the content of the images. In [17], Schmid et al. present a detailed comparison of several methods for interest region description. They introduce several descriptors such as the SIFT and LBP that are used to classify different images in different classes such as Urban or Trees. These classes correspond to a known [6] dataset used worldwide to test the quality of these algorithms. This is an important work because it summarizes the current state of the art in scene and object recognition.

To test and run recognition algorithms it is necessary to retrieve many test photos. A common source is Flickr [10], an open library of millions of photos that are freely submited by people around the world. In [7] the authors investigate how to organize a large collection of geo tagged photos, working with a dataset of about 35 million images collected from Flickr. This is one of the requisites for the project described in the solution section. In their approach, they combine content analysis based on text tags and image data with structural analysis based on geospatial data. This analysis is then used to predict where are the pictures taken from or what are the most popular places and landmarks on a given area.

3 Research Statement

Taking into consideration the problem statement already defined in section 1.1, the research here proposed aims to transform physical spaces and locations into virtual information that can be accessed seamlessly in visualization and simulation systems.

In this thesis, the main challenges that will be addressed are related with: (a) Acquisition, (b) Model Creation, (c) Visualization and Interaction.

3.1 Information Acquisition

The acquisition of information will eventually come from three sources: cameras (image and video), internet media databases or specific sensors.

Using cameras is a very simple solution because there are several software libraries which support them. There are image processing and computer vision solutions such as OpenCV [31], and the literature on the subject is extensive [3, 12, 21]. Using simple USB web cameras connected to the computer or mobile device to obtain visual information avoids the use of complicated hardware bridging the gap between the lab setup and the home user. With cameras it is possible to capture still images or videos, with the possibility of real-time capture.

Other source of images and videos are the open image database available on the internet. There are many types of databases, but it is possible to divide them in two groups. The first group corresponds to databases specially prepared to train, test and validate computer vision algorithms. As an example, the PASCAL Visual Object Classes [6] is a database composed of classified images that are then used in object recognition. Another group of databases belong to websites that use user-generated content such as Flickr [10]. These websites, sometimes associated with social networks, have very large quantities of images and videos generated and classified by their users. Most of them have APIs that allow the retrieval of information for other purposes [13]. It is possible, for example, just by doing a simple query, to retrieve hundreds of photos of a given known location or building, from different years, time of the day or angle.

The use of sensors is another possibility to obtain data. There are many types of sensors that can be used. The main focus will be given to two known hardware kits that are simple, available and low-cost: Arduino [2] and Lego Mindstorms [26]. Although some exploration on this topic is needed, these sensors can be used to detect the orientation of a camera, measure the distance of objects or detect the presence of motion. The use of sensors combined with cameras is also a possibility to enhance the data obtained. One example of this is the ZCam [44] recently incorporated in the Microsoft project Natal [36] for the Xbox.

For outdoor scenarios another source of information are the Geographic Information Systems. These can be open and accessible through APIs such as the Google Earth and Google Maps [14] or can be specific data from a certain place with 3D terrain information. In the project Life-Saver [29]the geographic data was supplied by LNEC and contained a detailed 3D representation of a river valley. An important aspect that could be explored is the intersection of information from images, cameras, GPS, maps and 3D terrain representations.

3.2 Model Construction

The construction of the models will resort to several techniques from computer vision. Depending on the availability of raster data, the algorithms that will be explored will try to retrieve 3D information from images. This 3D information can be obtained solely with images [21] or using complementary data from other sensors as done by the ZCam [44].

The techniques to transform 2D in 3D include using several images to find relations between them. This can be done with images from a scene as it is done by PhotoSynth [34]. This matching can be helped by using geographically annotated images that already have GPS coordinates and orientation. Another method is to use several close images has done by Peters et al. [33]. These are refined methods of using stereo vision [27] in order to triangulate the 3D position of the objects. Once a certain amount of images is positioned in the 3D space, other images that are close can be related using panoramic [22] techniques.

Object recognition and tracking [42, 45] may also be important to insert or monitor data into the virtual models.

The study of these methods will enable to understand which techniques are more adequate for data input. Additional information can be inferred from the meta-data of the images, parameters such as date, time, size, resolution or dominant color can provide information that may be important.

A detailed overview of the virtual models that will be experimented will be given next, in subsection 3.4.

3.3 Visualization and Interaction

To support the construction and interface with the virtual model and tools, several multimedia applications will have to be constructed. Development will rely on open and free frameworks that are available online. Two of these frameworks are Processing [35] and OpenFrameworks [32]. These libraries integrate several modalities such as video, sound, display, cameras, computer vision support and many others. Following the previous experience in [28] the graphic 3D library OGRE3D [30] may also be used.

On the interaction level, several experiments will take place using cameras which may also be in mobile devices. Although it is not a primary objective, some level of innovative interaction may arise from the computer vision studies. Additionally there is the intention of exploring an interaction with multi-touch tables due to some previous work in this area and its applicability to the problem.

3.4 Prototypes and Experiences

Until now, in this document, only the broader objectives have been discussed. These are in general, study computer vision, information retrieval from several sources and coherent integration of this information in models. In this subsection, several specific prototypes and experiences that are meant to be carried out in this PhD are explained. Figure 2 is an overview of the main directions that this thesis will follow.

3.4.1 Virtualized Room

The idea of this prototype is to automatically recreate an indoor space. This should be a faster procedure that any user with no special skills could do. To do this a set of cameras and sensors should be used to obtain the three-dimensional picture of the space. The 3D model can then be presented in interactive applications where the elements such as colors or textures of certain objects can be changed. Additionally the user should be able to select a certain object to be tracked in future data updates. This can be useful to track changing objects in a room. Using the virtual model of the room,

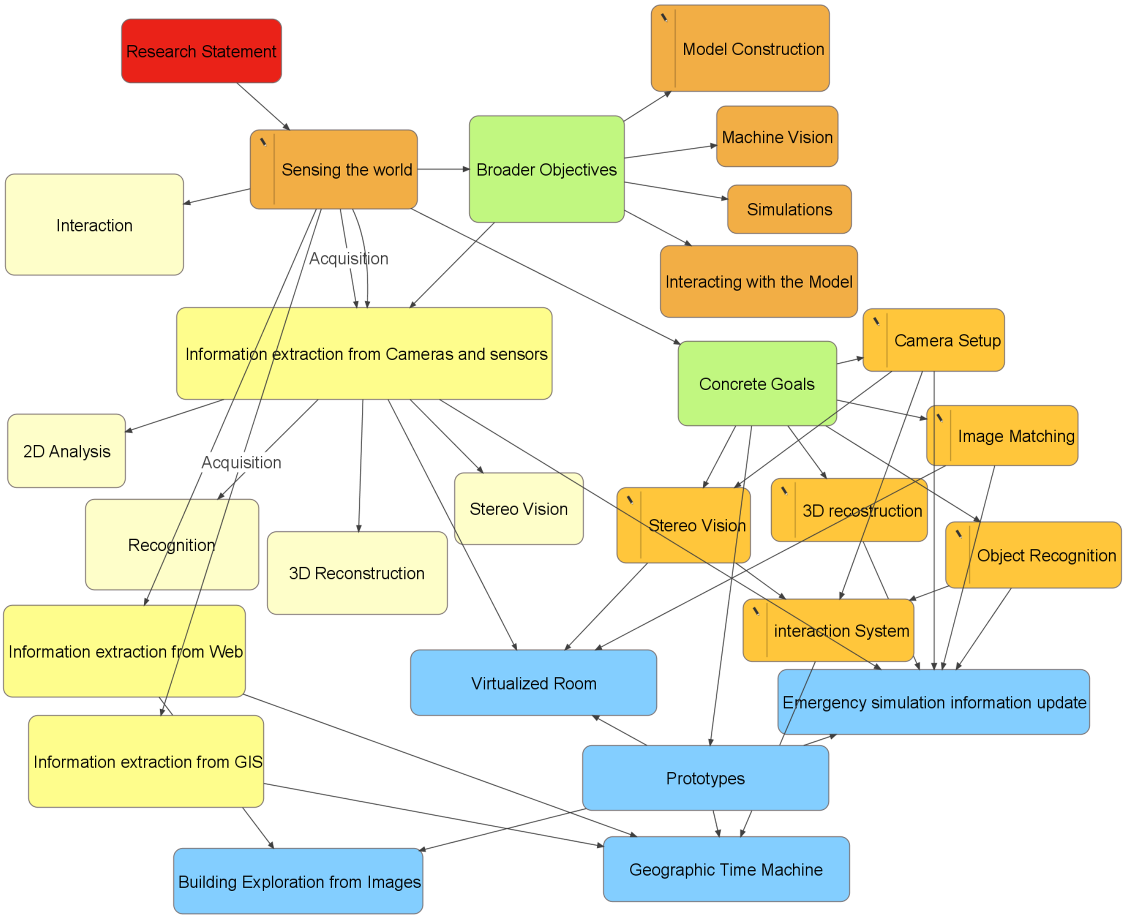

Figure 2: Map representing the different areas of the proposed work.

synthetic objects can be added and superimposed into the reality as augmented reality. Applications of this technique could be testing redecoration of rooms, navigation inside buildings or virtual stores that imitate the real ones but have different content in virtual reality.

3.4.2 Enhancing Emergency Simulations

Simulations are important in several aspects of our life, and they allow us to predict, plan and reduce costs, but sometimes constructing a simulation model is not a simple process. The amount of data necessary may not encourage the use of simulations. The idea proposed here is the use of a computer vision methodology to enhance simulation models. The method works by extracting information from images and videos, collected by the user or surveillance cameras. These images are integrated in a three-dimensional representation of the scenario and then processed with recognition algorithms that identify the content of the raster images. The classified content is then added as features to the simulation model.

As an example, in a typical fire emergency simulation [5, 41], it is important to know what kind of buildings or vegetation are present in the scenario in order to extrapolate the speed of the fire propagation. Currently most open web geography applications simply represent an area as being an urban or rural place. A national park is usually a green area in the map with no information about the type of vegetation or the infrastructures that compose it, such as, sand roads, huts, small rivers or fences. The idea here is to identify these elements by combining computer vision with 3D representations of the terrain and GIS. Additionally it would be interesting to have live feeds of information (i.e. live images of a fire or flood coming in).

Other interesting applications would use the simulation results, reverse the process and synthesize the original images with additional visual information taken from the data of the simulations. Imagining a pollution simulation, where a green water plat infestation spreads trough the river. A photo from the river is used to obtain the water level and the current amount of seaweeds. Using the above described process, data would be inserted in a seaweed growth model simulation. The resulting predictions could be translated back into the original image, where a synthesized area full of water plants would be artificially introduced to represent the future pollution state of the river. This reverse process would use the same ray tracing principle but now using the terrain as a source.

3.4.3 Building Exploration from Images

There are many famous buildings and sites of which thousands of photos were taken. It would be interesting to create an automatic model of one of these sites based solely on collections of pictures and videos taken by several people. The prototype would recreate a virtual model of the building and allow for the observation of the place from several perspectives. Although this idea is somewhat inspired by PhotoSynth [34], it is possible to explore the cross-image model using different paradigms such as time, number of people in the vicinity, color, weather or historical relations.

3.4.4 Geographic Time Machine

This prototype would take advantage of geographically tagged photos and videos in large areas such as cities or nature parks. The images and videos used as source have several properties such as size, resolution, color depth; they represent images taken at a given point in time, at a certain period of the day with specific weather and illumination conditions. This poses several problems because each one of these variables is not constant. After a large set of data is collected and classified it becomes important to retrieve only selected portions of information. One idea would be the creation of a time machine application, which would select the images and videos from a given period of time, and present its projections on the 3D terrain. The information could also be selected using as criteria the illumination, the season or any identifiable characteristic from the images. The different criteria could be used to create different layers of information that can be totally or partially enabled in the geographic terrain, with certain areas showing one kind of information and others showing other at the same time.

4 Work Plan

The research work is planned to last three years addressing several tasks such as: (a) Research, (b) Experimentation, (c) Prototype Implementation, (d) Discussion, (e) Peer-Reviewing and (f) Final Thesis Writing.

The (a) Research stage was initiated before the writing of this document and will be continuously updated considering new proposals and the directions of work that are followed. Albeit this, the main research effort will take place in the beginning, with special focus to the methods that will help in the definition and implementation of the prototypes. This will happen in the Experimentation (b) phase where several small prototypes will test most of the concepts described in section 3. This experimentation will occur in the beginning of the dissertation, most of it should be done by the end of the first year.

With the experience obtained in the research and experimentation the second year will be devoted to the implementation of final working prototypes such as the ones described in subsection 3.4.. Some of this work may still be in progress in the beginning of the third year. The work should be discussed (d) internally and submitted regularly to international conferences for peer-reviewing (e). This documentation of the work in paper should occur regularly across the dissertation, reflecting the results, even if preliminary, that are being obtained.

In the final year the writing of the thesis will progressively occupy most of the time.

4.1 Evaluation and Final Remarks

To evaluate the feasibility of the prototypes there will be several user studies and computational benchmarks. The user studies will be conducted in the experiments that are meant to be used by an end-user, such as the four prototypes presented in section 3.4. There are different types of tests that will be considered. For example, in some prototypes such as the virtualization room (section 3.4.1) and the simulation model (section 3.4.2) the subject of the test will be the creation of the model and the subsequent interaction with it. In the other prototypes only the interaction will be studied. The studies will be conducted using techniques such as the ones described in [8]. Low and high fidelity prototypes will be developed to study the interface of the applications. Additionally, the applications will track the user's behavior and measure their response time, the completion time and other possible meaningful data.

References

[1] Besma R. Abidi, Nash R. Aragam, Yi Yao, and Mongi A. Abidi. Survey and analysis of multimodal sensor planning and integration for wide area surveillance. *ACM Comput. Surv.*, 41(1):1–36, 2008.

[2] Arduino. `http://www.arduino.cc/`, 2009.

[3] Bir Bhanu and Ioannis Pavlidis. *Computer Vision Beyond the Visible Spectrum.* SpringerVerlag, 2004.

[4] F. Calabrese and C. Ratti. Real time rome. *Official Journal of the IGU's Geography of Information Society Commission*, 20:247–258, 2006.

[5] Modesto Castrillón, Pedro A. Jorge, Adrián Macías, Antonio J. Sánchez, Javier Sánchez, José P. Suárez, Agustín Trujillo, Izzat Sabbagh, Ignacio J. López, and Rafael J. Nebot. Wildfire forecasting using an open source 3d multilayer geographical framework. In *SIGGRAPH '09: SIGGRAPH 2009: Talks*, pages 1–1, New York, NY, USA, 2009. ACM.

[6] PASCAL Visual Object Classes Challenge. `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2006`, 2006.

[7] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 761–770, New York, NY, USA, 2009. ACM.

[8] Alan Dix, Janet E. Finlay, Gregory D. Abowd, and Russell Beale. *Human-Computer Interaction (3rd Edition).* Prentice Hall, 3 edition, December 2003.

[9] B. Fischer, S. Perkins, A. Walker, and E. Wohlfart. Hypermedia image processing reference 2. `http://homepages.inf.ed.ac.uk/rbf/HIPR2/`, 2009.

[10] Flickr. Photo-sharing. `http://www.flickr.com`, 2009.

[11] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics: Principles and Practice in C.* Pearson, 2nd edition, 1990.

[12] David A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach.* Prentice Hall, us ed edition, August 2002.

[13] Fabien Girardin, Francesco Calabrese, Filippo Dal Fiore, Carlo Ratti, and Josep Blat. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, 7(4):36–43, 2008.

[14] Google. Google earth and maps. `http://earth.google.com`, `http://code.google.com/apis/earth/`, `http://maps.google.com`, `http://code.google.com/apis/maps/`, 2009.

[15] NUI group. Natural user interface group, 2009.

[16] Jefferson Y. Han. Low-cost multi-touch sensing through frustrated total internal reflection. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, pages 115–118, New York, NY, USA, 2005. ACM.

[17] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with local binary patterns. *Pattern Recogn.*, 42(3):425–436, 2009.

[18] Tung-Ju Hsieh, Michael J. Olsen, Elizabeth Johnstone, Adam P. Young, Neal Driscoll, Scott A. Ashford, and Falko Kuester. Vr-based visual analytics of lidar data for cliff erosion assessment. In *VRST '07: Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pages 249–250, New York, NY, USA, 2007. ACM.

[19] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. *ACM Trans. Graph.*, 28(3):1–8, 2009.

[20] S. Jordà, G. Geiger, M. Alonso, and M. Kaltenbrunner. The reactable: Exploring the synergy between live music performance and tabletop tangible interfaces. In *Proc. Intl. Conf. Tangible and Embedded Interaction (TEI07)*, 2007.

[21] Reinhard Klette, Andreas Koschan, and Karsten Schluns. *Computer Vision: Three-Dimensional Data from Images.* Springer-Verlag Singapore Pte. Limited, 1998.

[22] Barbara Krausz, Andreas Brièll, Christian Eckes, and Jobst Löffler. Capturing and processing of 360 panoramic images for emergency site exploration. pages 76–90, 2009.

[23] Senseable City Lab. Mit. `http://senseable.mit.edu`, 2009.

[24] Ze N. Li and Mark S. Drew. *Fundamentals of Multimedia.* Pearson Prentice Hall, first edition, 2004.

[25] Larry Matthies, Byron Chen, and Jon Petrescu. Stereo vision, residual image processing and mars rover localization. In *ICIP '97: Proceedings of the 1997 International Conference on Image Processing (ICIP '97) 3-Volume Set-Volume 3*, page 248, Washington, DC, USA, 1997. IEEE Computer Society.

[26] Lego Mindstorms. http://mindstorms.lego.com, 2009.

[27] Sandino Morales and Reinhard Klette. A third eye for performance evaluation in stereo sequence analysis. In *CAIP '09: Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*, pages 1078–1086, Berlin, Heidelberg, 2009. Springer-Verlag.

[28] Rui Nóbrega and Nuno Correia. *Visualization and Interaction in a Simulation System for Flood Emergencies, Master Thesis.* FCT-UNL, 2008.

[29] Rui Nóbrega, André Sabino, Armanda Rodrigues, and Nuno Correia. Flood emergency interaction and visualization system. In *VISUAL '08: Proceedings of the 10th international conference on Visual Information Systems*, pages 68–79, Berlin, Heidelberg, 2008. Springer-Verlag.

[30] Ogre3D. ver. 1.6.4. http://www.ogre3d.org/, 2009.

[31] OpenCV. http://sourceforge.net/projects/opencv/, 2009.

[32] OpenFrameworks. http://www.openframeworks.cc/, 2009.

[33] Gabriele Peters and Klaus Häming. Take three snapshots - a tool for fast freehand acquisition of 3d objects. In *INTERACT '09: Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction*, pages 842–843, Berlin, Heidelberg, 2009. Springer-Verlag.

[34] PhotoSynth. http://photosynth.net, 2009.

[35] Processing. http://processing.org/, 2009.

[36] Microsoft Research. Project natal. http://www.xbox.com/en-US/live/projectnatal/, 2009.

[37] H. S. Sawhney, A. Arpa, R. Kumar, S. Samarasekera, M. Aggarwal, S. Hsu, D. Nister, and K. Hanna. Video flashlights: real time rendering of multiple videos for immersive model visualization. In *EGRW '02: Proceedings of the 13th Eurographics workshop on Rendering*, pages 157–168, Aire-la-Ville, Switzerland, Switzerland, 2002. Eurographics Association.

[38] Ben Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer-Interaction, third edition.* Addison-Wesley, 1998.

[39] Camillo J. Taylor, Paul E. Debevec, and Jitendra Malik. Reconstructing polyhedral models of architectural scenes from photographs. In *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume II*, pages 659–668, London, UK, 1996. Springer-Verlag.

[40] Terravision. http://www.artcom.de/, 2009.

[41] Sébastien Thon, Eric Remy, Romain Raffin, and Gilles Gesquière. Combining gis and forest fire simulation in a virtual reality environment for environmental management. *ACE*, 2(4):1886–4805, 2007.

[42] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[43] NASA World Wind. http://worldwind.arc.nasa.gov, 2009.

[44] G. Yahav, G. J. Iddan, and D. Mandelboum. 3d imaging camera for gaming application. In *Consumer Electronics, 2007. ICCE 2007. Digest of Technical Papers. International Conference on*, pages 1–2, 2007.

[45] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.