ISPRS Journal of Photogrammetry and Remote Sensing 65 (2010) 134–142

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Augmented reality and photogrammetry: A synergy to visualize physical and virtual city environments

Cristina Portalés *, José Luis Lerma, Santiago Navarro

Photogrammetry and Laser Scanning Research Group (GIFLE), Department of Cartographic Engineering, Geodesy and Photogrammetry, Universidad Politécnica de Valencia, Camino de Vera s/n, Building 7i, 46022 Valencia, Spain

ARTICLE INFO

Article history:
Received 14 July 2008
Received in revised form
16 September 2009
Accepted 5 October 2009
Available online 17 October 2009

Keywords:
Augmented reality
Model
Multisensor
Tracking
Navigation
Real time

ABSTRACT

Close-range photogrammetry is based on the acquisition of imagery to make accurate measurements and, eventually, three-dimensional (3D) photo-realistic models. These models are a photogrammetric product per se. They are usually integrated into virtual reality scenarios where additional data such as sound, text or video can be introduced, leading to multimedia virtual environments. These environments allow users both to navigate and interact on different platforms such as desktop PCs, laptops and small hand-held devices (mobile phones or PDAs). In very recent years, a new technology derived from virtual reality has emerged: Augmented Reality (AR), which is based on mixing real and virtual environments to boost human interactions and real-life navigations. The synergy of AR and photogrammetry opens up new possibilities in the field of 3D data visualization, navigation and interaction far beyond the traditional static navigation and interaction in front of a computer screen.

In this paper we introduce a low-cost outdoor mobile AR application to integrate buildings of different urban spaces. High-accuracy 3D photo-models derived from close-range photogrammetry are integrated in real (physical) urban worlds. The augmented environment that is presented herein requires for visualization a see-through video head mounted display (HMD), whereas user's movement navigation is achieved in the real world with the help of an inertial navigation sensor. After introducing the basics of AR technology, the paper will deal with real-time orientation and tracking in combined physical and virtual city environments, merging close-range photogrammetry and AR. There are, however, some software and complex issues, which are discussed in the paper.

1. Introduction

Close-range photogrammetry is currently an efficient tool to derive geometrical information from digital imagery in a fast and economic way. Nowadays, most often the derived photogrammetric product is a 3D photo-realistic model, and two different techniques for modelling objects coexist with success: 'virtual' reality and 'visual' reality (Grussenmeyer et al., 2002). While in the former the texture and object shapes do not need to correspond to real objects, the 3D models in the latter are a replica of physical objects, and they help one to understand their corresponding real objects. Derived photogrammetric models can provide real measurements, as well as virtual and visual realities, depending on the images used to finally texture both objects and environments.

In the last few years, with the increasing capabilities of standard personal computers, new tools have appeared that allow the addition of multimedia content to 3D models such as sound, text, videos, etc. Moreover, navigation in real time is possible via the Internet, and some kind of user interaction is usually possible. The high potential value of interactive 3D photo-models in many different areas is undeniable. It is possible to find many good examples concerning photo-models in different fields, such as the study and conservation of cultural heritage sites (Behan and Moss, 2006; Obleby, 1999), visualization and dissemination over the Internet (Dorffner and Forkert, 1998; Lerma and García, 2004) or scene reconstructions (Fraser et al., 2005).

Nowadays, there are new approaches that can bring further potential to acquired models by close-range photogrammetry, as is the case of augmented reality. This is an emerging technology that is leading to new kinds of visualizations, navigations and user interactions, opening new possibilities for the understanding of virtual photo or non-photorealistic models. Due to its novelty, the research interest has so far been focused on tracking and registration, display technology, rendering, interaction devices and techniques, presentation and/or authoring (Bimber and Raskar, 2005). Nevertheless, nowadays research is also focussing on the applications that can be derived from this technology to issues

* Corresponding author.
E-mail addresses: criporri@upvnet.upv.es, criporri@gmail.com (C. Portalés), jllerma@cgf.upv.es (J.L. Lerma), sannata@topo.upv.es (S. Navarro).

0924-2716/$ – see front matter © 2009 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2009.10.001

related to the final user, such as physical human-to-human or human-to-computer interactions (Cheok et al., 2006).

This article sets up a synergy between conventional close-range photogrammetry and innovative AR approaches in order to explore enhanced visualization, spatial data management, user navigation and/or interaction. We introduce a low-cost outdoor mobile AR system to allow the visualization of virtual building models acquired by multi-convergent close-range photogrammetry which are integrated into physical environments. 3D models can be achieved via terrestrial laser scanners or computed via image-based modelling. Although the former approach is nowadays becoming a standard source for input data in many applications, the latter remains the most widely used in general 3D applications; the requirements of hardware are less demanding and low-cost implementations are possible with existing technology.

Image-based modelling is applied herein. The images were taken with a low-cost conventional digital camera, following the well-known CIPA 3×3 rules (Waldhäusl and Ogleby, 1994). Afterwards, a self-calibration bundle block adjustment was performed to build up the photo-model. This paper depicts a scenario where virtual photo-realistic 3D models are combined into different real (physical) scenarios to figure out how different architectonic constructions might change citizens' identity and behaviours against governmental decisions. The case study presented below is fictional, but it points out the power of combination of two different purpose technologies, AR and photogrammetry. The former is more concerned with real-time processing, high speed and continuous visualization over rough models, while the latter is more concerned with high accuracy, perfect geometry and matching between 3D models and imagery. The photogrammetric technique contributes the visual reality environment to the augmented 3D world.

Section 2 presents a review of AR technology and points out the general benefits it can bring in this field, giving some hints on the composition of an AR environment. Section 3 presents a case study in which a historical building of the city of Valencia is placed on our campus, and topics such as AR system configuration, reference frame and problems due to occlusions are tackled. Section 4 reviews the programming environment, carried out within the Max/MSP Jitter (Cycling '74, 2008) software. Finally, Section 5 presents a discussion and conclusions.

2. AR technology

2.1. Description

AR is a relatively new technology that is based on mixing computer generated stimuli (visual, sound or haptic) and real ones, keeping a spatial relationship between synthetic and physical data and allowing user interaction in real time, as described in Azuma (1997). In recent years, it has been introduced in many different areas, mainly to visualize both virtual data and real environments all together. Such emerging areas include education (Kaufmann and Schmalstieg, 2003; Portalés Ricart et al., 2007), entertainment (Cheok et al., 2003; Wagner et al., 2004), GIS (King et al., 2005), media arts (Benford et al., 2006; Levin, 2006), psychology (Juan et al., 2005), robotics (Stilman et al., 2005), surgery (Glossop and Wang, 2003; Wacker et al., 2005) and urban planning (Ben-Joseph et al., 2001; Portalés Ricart et al., 2005), amongst others. The dizzily increasing number of AR applications in the last few years is due to the new potential that this technology brings. According to Billinghurst and Kato (2002), augmented reality provides:

- *Seamless interaction between real and virtual environments*. An interface seam was introduced by Ishii et al. (1994) and it can be described as a spatial, temporal or functional constraint that forces the user to change between a variety of spaces or ways of working. In AR the communication between users is produced in a natural way as users can still work with traditional tools and are able to see each other at the same time as virtual data, thus allowing face-to-face collaboration. For example, a study is described in Kiyokawa et al. (2000) based on an AR system designed to minimize interaction seams.
- *The ability to enhance reality*. With AR systems, multimodal computer generated stimuli can be added to the physical environment. Moreover, some parts of the reality can be modified or even deleted (e.g. with virtual phantoms that occlude real objects). This is the case presented in Fischer et al. (2005), where different filters are applied to the resulting augmented scene of a video composition.
- *The presence of spatial cues for face-to-face and remote collaboration*. Computer generated objects can be spatially distributed in real time according to physical environments. For instance, in Kato et al. (2001) an AR videoconference system is introduced, where users can distribute the video images of other people around a physical table.
- *Support of a tangible interface metaphor for object manipulation*. In AR systems there exists a close relationship between virtual and real objects, as physical objects can be augmented through computer generated data, allowing a dynamic superimposition of those elements. Therefore, physical objects can be used to directly manipulate virtual data in such an intuitive manner that people with no computer background can still have a rich interactive experience. For example, in Cheok et al. (2006) a review of several applications in the field of entertainment is made, highlighting the natural human-to-human and human-to-computer interactions that can arise with this technology.
- *The ability to transition smoothly between reality and virtuality*. According to Milgram's continuum (Milgram and Kishino, 1994), also known as virtuality continuum (Fig. 1), a classification of interfaces can be made depending on the amount of synthetic data in proportion to the real environment, thus discerning between augmented reality (closer to a real environment) and augmented virtuality (closer to a virtual environment). Therefore, interactive 3D photo-models would be close to the right side of the continuum, inside the area of augmented virtuality, because of the photorealistic texture applied to the virtual models of real objects.

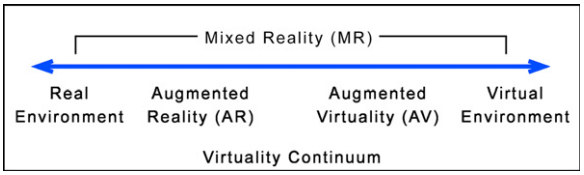

Fig. 1. Simplified schema of a virtuality continuum. From (Milgram and Kishino, 1994).

These issues together with the increasing computational capacities of standard personal computers encourage scientists, researchers and other members of the public to improve, develop and use AR systems.

2.2. Composition of augmented reality environments

Compatibility between both real and virtual data is an important issue in AR (Wang and Dunston, 2006). To properly combine virtual and real objects in such a way that the augmented scene appears to be plausible to the user, the real camera should be mapped to the virtual one in such a way that the perspectives of both environments match. Other issues that may be considered to increase the realism of the rendered augmented scenes are occlusion between real and virtual objects, lighting, shadowing and reflections.

To correctly match real and virtual worlds, computer generated objects need to be accurately registered to the real world, so that

they appear to the user as fixed in the environment. According to Bimber and Raskar (2005), accurate tracking is one of the most significant challenges in AR research today, as accuracy is frequently crucial and depends essentially on the type and resolution of the sensors (e.g. GPS, INS, vision based). If absolute tracking with a global coordinate system is required, it can be distinguished between outside-in and inside-out tracking. The first case refers to those systems where sensors are fixed on the environment and register a set of emitters on mobile objects. The second type makes use of sensors fixed to mobile objects. Nevertheless, the acquisition of the exterior camera orientation in real time for a wide area is not always possible with the use of a single tracking technology due to limitations in sensors. For example, magnetic and radiofrequency sensors are influenced by metallic interferences; GPS receivers suffer from the multipath problem and cannot be used inside buildings; vision-based tracking depends strongly on lighting conditions and visibility. Furthermore, it is not always possible with a single technology to either register the camera's six degrees of freedom (DOFs) or add additional user interaction. This means that some authors integrate different tracking technologies. For example, in Kiyokawa et al. (2000) a collaborative design AR system is presented where camera orientation and user interaction are achieved via a combination of several magnetic sensors and push buttons. In Piekarski (2006) a mobile outdoor AR system is described where camera tracking is achieved via a combination of a GPS receiver and an inertial sensor, whereas user interaction is fulfilled within a data glove and an optical system.

One has to point out the importance that mobile technology (mobile phones and PDAs) is achieving in the field of outdoor AR-based applications. The increasing power of these devices with broadband network, integration of high-resolution cameras and low-cost GPS receivers, leads to small equipment and high performance. Many authors have used this technology in collaborative AR environments, where user communication with other users and/or remote resources is crucial (Benford et al., 2006; Díez-Díaz et al., 2007).

On the other hand, occlusion is a well-known problem within AR research. When doing a video composition of real and virtual scenes, virtual objects are always mapped on top of the images of the physical environment. Thus, non-desirable occlusions can occur. Several authors have tried to solve this problem in different ways. For example, in Lepetit and Berger (2000) a method is developed based on a semi-automatic occluding boundary reconstruction from different camera points of view, whereas in Fischer et al. (2003) a method is presented based on detecting occlusions in front of static backgrounds. To further increase realism, the lighting conditions of the virtual scene should coincide within the ones obtaining in the real environment. Furthermore, reflections and shadowing of virtual objects onto real ones can also be considered. Some of these issues have been implemented by various authors, such as Gibson et al. (2003), Jacobs and Loscos (2006), Stauder (1999), Tatham (1999) and Wang and Samaras (2003). Nevertheless, some of these techniques are complex and require high processing speed, which makes them too cumbersome to be applied in real time. Therefore, in our implementation, only occlusions and lighting on the Serrano Towers are considered (see Sections 3 and 4).

3. Case study

We have developed an AR application for urban visualization based on building models acquired by close-range photogrammetry. In this section we show a test carried out at the campus of the Universidad Politécnica de Valencia (UPV), where a 3D model of a landmark gate of the city centre, the Serrano Towers (*'Torres de Serranos'*), dating back to the XIVth century, is spatially integrated with the physical modern buildings (late XXth century and early XXIst century) on the university campus. The ancient Serrano Towers were part of the old city walls and are located on the river banks delimiting the city centre. The distance from the Northeast suburbs where the UPV is situated to the Serrano Towers is approximately 3 km. User orientation and positioning in real time is acquired with a combination of an inertial sensor and a vision-based tracking procedure, applying basic photogrammetric rules. Other issues such as lighting of virtual models and occlusions are also considered.

Fig. 2. User carrying the devices needed for our AR mobile application.

3.1. AR configuration

Our application can be classified as an outdoor mobile AR system where the user wears the devices needed for the tracking, generation and rendering of the augmented environment. These devices are (Fig. 2):

(a) A standard laptop inside a backpack or carry bag.
(b) A display system. Specifically, we used an I-glasses SVGA video-based HMD, with a resolution of 800 × 600, 26° diagonal field of view and modifiable brightness; it is 2D based, i.e., the same image of the scene is rendered for both eyes.
(c) A standard web cam, with 640 × 480 resolution and USB 2.0 connection.
(d) An inertial sensor. We used the MT9 Xsens miniature inertial sensor, which operates at frequencies up to 512 Hz, provides angular resolutions of 0.05° and accuracies better than 1° RMS in the three axes.
(e) Batteries for the HMD and INS.

The user carries the backpack where the laptop and batteries are integrated. The laptop processes all the received data in real time and performs target oriented processing to simulate an augmented urban environment through an application written in Max/MSP Jitter. The displays used in outdoor mobile AR applications are usually seen through HMD or small hand-held devices (mobile phones or PDAs). In the second case, there exists the disadvantage that these devices are not yet powerful enough to manage complex augmented environments. Some authors also use screen-based displays (tablet-PC or standard monitors), although they have to

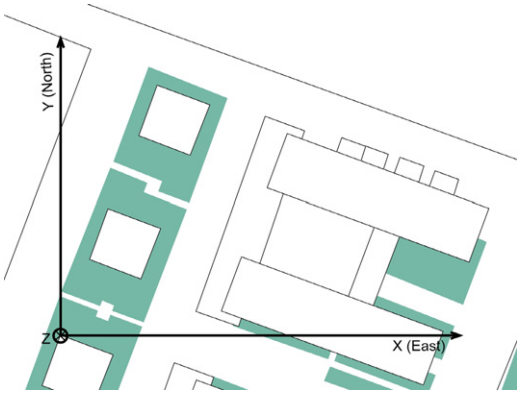

Fig. 3. Local terrestrial reference frame.

be shaded from the incoming sunlight due to the usually limited luminosity of the screen (Wilde et al., 2003). In our application, we can adjust the brightness and contrast of the see-through video-based HMD according to the user needs. Attached to the HMD there is a Webcam whose projection centre is approximated to the user's point of view. We found that the optimum camera resolution is 640 × 480 pixels, thus allowing procedures in real time within 30 fps. The inertial sensor is also attached to the HMD in such a way that its axes are aligned to those of the camera.

3.2. Campus reference frame and tracking

To geometrically combine both the virtual and the physical scenes, the exterior orientation of the virtual camera has to be equal to that of the real one, so a common reference frame has to be established. To that purpose, we considered a Cartesian local reference frame, with the *Y*-axis pointing to the North and *X*-axis to the East. The *Z*-axis is perpendicular to the *XY*-plane and pointing to the zenith. The origin is placed at the corner of a building (Fig. 3).

The user's positioning and orientation in real time is achieved via a combination of an inertial sensor and vision-based tracking. Both sensors – inertial and optical – are calibrated and aligned. Observations are integrated into a system of central projection equations. In our application rotations are directly measured by the inertial sensor.

Certain topographical measurements were needed in order to compute the 3D coordinates of the control points. 3D coordinates can be determined making use of GNSS, urban maps or architectonic façade plans. Herein, a reflectorless total station was used for this purpose. These points were introduced in the physical environment as circular flat elements of different colour (Fig. 4). The minimum number of control points needed to calculate the exterior orientation of the moving camera is two, due to the rotation outputs of the inertial sensor. Nevertheless, extra points were measured to keep them in reserve for further work. It must be pointed out that the control points should be distant in order to increase the accuracy, although it must also be ensured that they remain inside the field of view of the camera. If one of the two control points is out of the field of view, then the superimposition of the virtual photo-model will not be successfully overlaid and tracking will fail.

3.3. Photo-model generation

In this application two kinds of virtual model are needed: the virtual model of the Serrano Towers, and the virtual model of the buildings belonging to the area where the application takes place, the UPV campus. The latter is necessary in order to solve occlusions (see Section 3.4).

Fig. 4. Circular coloured elements acting as control points for the tracking in real time.

To acquire the set of imagery for the purpose of modelling the Serrano Towers, the so-called CIPA 3 × 3 rules have been applied. These rules have been described for simple photogrammetric documentation of architecture in those cases where non-metric cameras are used. They are structured in three triplets (Waldhäusl and Ogleby, 1994):

- Three *geometrical rules*, where the preparation of control information, the photographic coverage and stereo-partners are considered.
- Three *photographic rules* regarding the inner camera geometry, illumination and camera format.
- Three *organizational rules* consisting of making proper sketches, protocols and a final check.

The control points for the whole building were measured according to a pre-defined local coordinate system. Additional tie-points were measured in each image. A bundle block adjustment was applied in order to both calibrate the cameras and generate the points that would be used to build up the 3D model. A total of 76 images were used to generate the 3D model (Fig. 5). All the computations were carried out with FotograUPV, a home-made photogrammetric software based on multi-convergent imagery. However, a uniform texture was applied to the whole model based on its real texture, keeping in mind that the textured model had to be rendered responding to the user movements in real time, and complex photorealistic models are difficult to manage with standard laptops. Afterwards, the model was exported into *obj* format, in order to be readable by the software (see Section 4).

Regarding the model of the university campus, a detailed VRML of the Higher Technical School of Geodesy, Cartography and Surveying can be found at Muñoz Santamaría (2006). This 3D model was achieved by photo-tacheometry due to the high number of planar features. Nevertheless, as this model was used as a 3D mask (i.e. it was not visualized), its geometry was simplified and its textures were not considered in order to avoid extra computational processing (Fig. 6).

3.4. Solving occlusion

Real and virtual objects are all integrated together inside the campus reference frame. In Fig. 7, the Serrano Towers are virtually placed and seen together on the university campus. The Towers are placed in the middle of a square garden in front of the main entrance of the School.

As mentioned in Section 2.2, occlusion is an issue to be solved in AR scenarios because virtual objects are always placed in the foreground of the incoming video image. However, a user is able

Fig. 5. The Serrano Towers: (a) images; (b) 3D model.

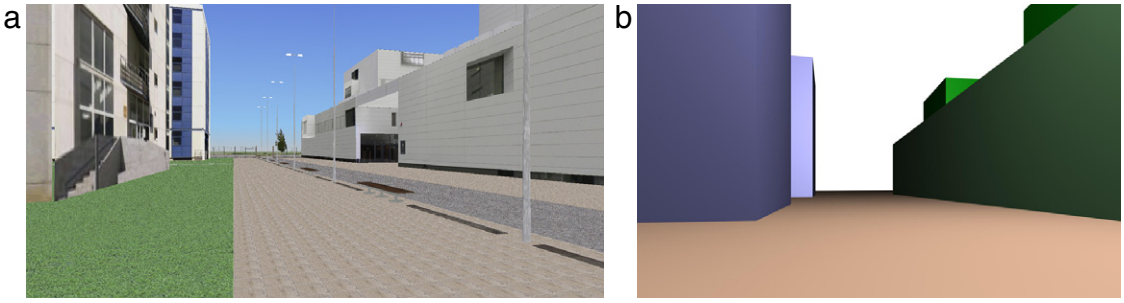

Fig. 6. Simplified VRML model of the Higher Technical School of Geodesy, Cartography and Surveying: (a) with photorealistic textures; (b) simplified model without textures.

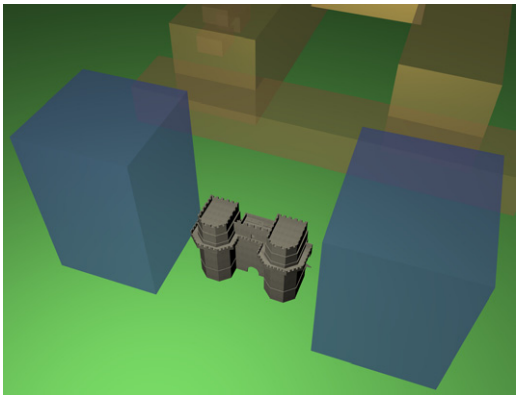

Fig. 7. The Serrano Towers integrated in the University campus.

Fig. 8. False augmented environment due to wrong occlusion analysis.

to walk around the considered study area without constraints. Therefore, it might happen that depending on his real-time spatial position, the visual model would appear in the background of the real buildings, and vice versa. If occlusions were not analytically solved, the resulting augmented image would not be plausible, Fig. 8. Indeed, the Serrano Towers are placed in the foreground of the image, when they should be behind the physical buildings according to the user's point of view. This situation is called a 'false' augmented environment.

In the application under study the occlusion issue was solved with the consideration of 3D masks. These masks are in fact simplified 3D models of the buildings containing texture and an alpha channel, which is blended in the visual model. If the visual model is spatially located in front of the 3D masks (according to the user's point of view), these masks will not affect its visualization; in contrast, if the 3D masks are placed in front the visual model, the former will visually delete parts of the latter, thus showing through the video image that is backwards. This process is described in Fig. 9, whereas in Fig. 10 other scenes of the same augmented environment can be seen (note that in Fig. 10b no occlusion is produced).

4. Software environment

The presented AR system is self-implemented into Max/MSP/Jitter (Cycling '74, 2008), which is a multipurpose software. Jitter is basically a set of video, matrix, and 3D graphics objects for the Max graphical programming environment, especially suited for real-time video processing, custom effects, 2D/3D graphics, audio/visual interaction, data visualization, and analysis. The architecture of the application is shown in Fig. 11. The data acquired by the camera in real time is managed through the *jit.qt.grab* object. Afterwards, it is analysed by a colour tracking algorithm that extracts the image coordinates of control points on site. This information together with the camera interior parameters, the known spatial position of control points and the inertial sensor data are introduced into the central projection equations to determine the spatial camera position. Both position and attitude are assigned to the *jit.gl.render* object to build the virtual scene according to the real one. On the other hand, the 3D virtual models are managed by the *jit.gl.model* object, where transparency is assigned to the masks models of the surrounding buildings. This processing is managed from a main patch and a set of sub-patches.

Fig. 9. Steps to solve occlusion: (a) 3D masks; (b) addition of the visual model; (c) addition of the video image; (d) resulting augmented environment.

Fig. 10. Augmented environment within different points of view.

The sub-patches are *p camori, p leastsquares, p trackpoint, p model* and *p light*. Within the *p camori* and *p leastsquares* sub-patches the mathematical processes to achieve camera exterior orientation are carried out. *P camori* contains the rotations of the inertial sensor (*roll, pitch, yaw*) calculated from the *mt9* object, which shows acquired data in the form of the rotation matrix. *p leastsquares* has the minimum squares procedure to solve the system of four equations given by the two control points. The solution of these equations gives the 3D coordinates of the moving camera (X_O, Y_O, Z_O), which are sent to the *pak camera* object of the main patch. Furthermore, considering that the Z_O coordinate of the camera has a small variation (as it coincides with the user's height), an additional constraint can be added to the system.

In the *p trackpoint* sub-patch the colour tracking of control points is achieved with the *tap.jit.colortrack* object of Tap Tools 1.5 (Electrotap, 2008). This object registers simultaneously up to four different colours based on a hue, saturation and brightness analysis of image pixels. The resulting values are the image coordinates of the window enclosing the coloured area. The centre of the window corresponds to the image coordinates of control points that are sent to the *p camori* sub-patch.

Sub-patch *p model* contains the properties of each virtual model that should be kept in *obj* format. Several attributes can be modified, such as lighting, shading, texturing, mapping or blending. The textures applied to the 3D models that act as masks are of black colour. These models are blended with the Serrano Towers according to the user's point of view in the following way: for those areas of the 3D models containing black colour information in the final scene, pixel information is replaced by the video image of the real environment. Therefore, those areas of the Serrano Towers being at the back of the 3D masks are not rendered in the final scene. Finally, in sub-patch *p light*, general light conditions of the virtual models can be controlled: ambient, diffuse and specular light, as well as spatial position of the light source, can be approximated to the incoming sunlight direction depending on the time of day to correctly shadow virtual objects. These properties are sent to *jit.gl.render* in the main patch.