Semantically Annotated 3D Material Supporting the Design of Natural User Interfaces for Architectural Heritage

Valeria Cera
Department of Architecture, University of Naples "Federico II"
Naples, Italy
valeria.cera@unina.it

Antonio Origlia
URBAN/ECO Research Center, University of Naples "Federico II"
Naples, Italy
antonio.origlia@unina.it

Francesco Cutugno
Department of Electrical Engineering and Information Technology, University of Naples "Federico II"
Naples, Italy
cutugno@unina.it

Massimiliano Campi
Department of Architecture, University of Naples "Federico II"
Naples, Italy
campi@unina.it

ABSTRACT

With the advent of artificial intelligence and natural user interfaces, the need for multimedia material that can be semantically interpreted in real time becomes critical. In the field of 3D architectural survey, a significant amount of research has been conducted to allow domain experts represent semantic data while keeping spatial references. Such data becomes valuable for natural user interfaces designed to let non-expert users obtain information about architectural heritage. In this paper, we present the architectural data collection and annotation procedure adopted in the Cultural Heritage Orienting Multimodal Experiences (CHROME) project. This procedure aims at providing conversational agents with fast access to fine-detailed semantic data linked to the available 3D models. We will discuss how this will make it possible to support multimodal user interaction and generate cultural heritage presentations.

CCS CONCEPTS

• **Human-centered computing** → **User centered design**; **Information visualization**;

KEYWORDS

Semantic annotation, architectural survey, interaction design

ACM Reference Format:
Valeria Cera, Antonio Origlia, Francesco Cutugno, and Massimiliano Campi. 2018. Semantically Annotated 3D Material Supporting the Design of Natural User Interfaces for Architectural Heritage. In *Proceedings of 2nd Workshop on Advanced Visual Interfaces for Cultural Heritage (AVI-CH 2018)*. Vol. 2091. CEUR-WS.org, Article 7. http://ceur-ws.org/Vol-2091/paper7.pdf, 4 pages.

1 INTRODUCTION AND RELATED WORK

Recent advances in graphics hardware, together with the availability of professional video-game engines, have opened a number of possibilities to develop innovative approaches for cultural heritage presentation. The use of game engines has been shown to produce beneficial effects on interaction quality with systems based on advanced knowledge representation and dialogue-based interaction (e.g. [6]). In particular, the use of conversational agents, represented in the form of 3D avatars moving in virtual reconstructions, provides a natural way to access information. Establishing a dialogue with an artificial character is becoming a more and more frequent way to interact with technological devices.

The annotation of digital models lets scholars associate spatial shapes with the heterogeneous data describing them through the use of semantic descriptors. The most relevant approach to this kind of semantic annotation is presented in [1] and it is based on the geometrical segmentation of architectural digital artefacts. These become collections of separate elements, organised using part-whole relationships. Each entity is identified by a precise concept in a specialised domain thesaurus: the architectural dictionary. Different geometrical representations (point clouds, nurbs, textured meshes, etc…) are linked to the objects represented by the terms, included in the dictionary, depending on the specific descriptive objectives. Each geometrical element can be linked to a single semantic descriptor, while a semantic descriptor may be associated to multiple geometrical elements. More recently, the original methodology has been updated [5] and implemented as a cloud-based service called *Aioli*[1]. Using the projective relationship between bidimensional and tridimensional representations, the semantic annotation of digital models, obtained through a set of reference images, is produced by segmenting the same reference images, thus removing the need of a geometrical segmentation. Images sharing the same semantic label may be linked to one or more specific terms in a controlled vocabulary, or they may be characterised with customised attributes. Semantically annotated 3D models contain a significant amount of data that, to promote cultural heritage, may be used to let non-expert users navigate cultural contents by developing interactive technologies. These technologies should be designed to assist the exploration of the large amount of information available for cultural heritage (texts, images, 3D models, etc…) in an engaging way. To tackle this problem, we pursue the use of conversational agents, in the form of 3D avatars, immersed in the digital representations of cultural artefacts. Using semantic

[1] www.aioli.cloud/

AVI-CH 2018, May 29, 2018, Castiglione della Pescaia, Italy

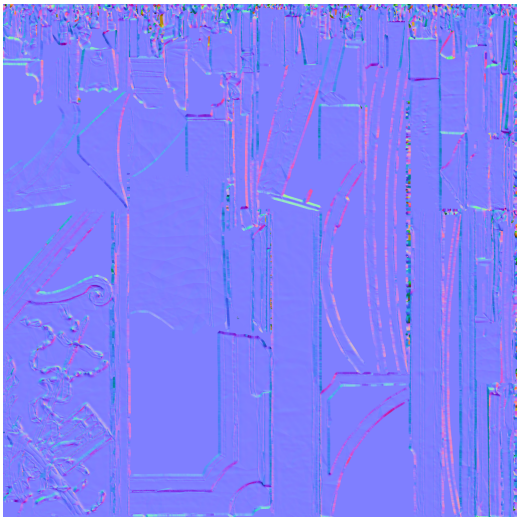

Figure 1: Normal map of a sample segment. RGB values represent the normal vector coefficients driving the lighting simulation of details.

Figure 2: Color map of a sample segment. RGB information is computed by comparing high and low poly meshes.

processing techniques coming from different domains (e.g. Natural Language Processing, Computer Vision, etc…), it is possible, using semantic labels, to link separate sources of information and generate a consistent presentation.

In this paper, we present the architectural data collection pipeline we adopted to obtain the 3D meshes representing relevant parts of the San Lorenzo Charterhouse in Padula (Italy) and how we annotated them with semantic information. The obtained data represent a multi-faceted documentation of architectural heritage describing both geometrical detail and visual experience. We also present the work in progress on a software architecture designed to link the semantically enriched 3D data to textual resources describing the represented artefact. This architecture will be used to support natural user interaction [10] through the use of Social Signal Processing [8] techniques and game engines.

2 DATA COLLECTION

The Charterhouse of San Lorenzo, in Padula, and its monumental staircase represents the selected case study, which is used to test the developed pipeline, spanning from the 3D data acquisition to the semantic annotation process. The staircase, made of local white stone, was built towards the end of the eighteenth century, has an elliptical plan and a double ramp. Closed outside by an octagonal tower, it leads to the first floor of the great cloister, used by the Carthusians for their weekly walk. Several surveying techniques were employed to produce a 3D reality-based model, suitable for dissemination purposes in virtual and interactive environments. In order to obtain a geometrically accurate 3D model, the survey was performed using a terrestrial laser scanner (TLS). Given the morphology of the staircase, its materials and colors, the geometrical data has been integrated with data collected during a photogrammetric campaign. This results in a physically accurate model that also delivers a photorealistic view of the surveyed cultural site, based on state of the art techniques.

Starting from the entrance, the positions of the different acquisitions have been organised to cover the entire volume of the monument, taking into account the tangency of the surfaces and shadows. A Continuous Wave Faro Focus 3D S120 laser scanner was used to perform a total number of 40 scans, positioning the scanner uniformly along an ascending path - for the eastern side - and a descendant one - for the western side -, with a spatial resolution of 6 mm at 10 m. A terrestrial photogrammetric survey was carried out mainly for texturing purposes. Using a Reflex Canon EOS 1300D and a zoom 18-55 lens set at 24 mm view, about 380 images were acquired to obtain a better color information for the final texturing of the 3D digital model.

3 DATA PROCESSING

The complete range-based 3D point cloud was obtained employing a classical processing procedure: the adjacent TLS stations were aligned using a solid-rigid transformation based on planar printed checkboards targets and spheres. A final point cloud of about 500 millions points was obtained.

After a manual cleaning of vegetation and artefacts caused by noise, a polygonal mesh model was generated using a Delaunay triangulation algorithm. A final mesh of about 392.4 millions of triangles was obtained this way. Once the triangulated model editing was completed, a texture mapping was carried out, using the images from the photogrammetric survey. To optimise the computational management of the models during online rendering, instead of generating a whole mesh, the process was set to divide the result into subparts. Each part is defined by an automatic subdivision of the model using a constraint of keeping a maximum of 5 million vertices per subpart. Considering the aims of dissemination and communication, the textured model was simplified using a successive geometric optimisation. The quadratic edge collapse algorithm

Figure 3: The rendered 3D model of the great staircase in the San Lorenzo Charterhouse.

was applied to obtain a polygonal mesh that allowed fluid real-time rendering while preserving an adequate level of perceived detail. The mesh produced with this procedure was collapsed with a target 1% vertices use from the initial mesh.

The deviation between the original and the decimated meshes was measured by calculating the Hausdorff distance. The approximation error was below 1 cm. To retain geometrical fidelity in the visualisation task, we compensate this error by computing normal maps that result from the comparison of the high poly and the low poly meshes, shown in Figure 1. Following the same approach used to bake normals on the low poly mesh, color information is baked using the high poly mesh. The resulting color texture is shown in Figure 2. This way, although geometrical data is lost during decimation, the simulated behaviour of light in the rendering engine takes into account the effect of the removed details. A rendered example of the final result is shown in Figure 3.

From a cultural heritage documentation point of view, it is desirable to preserve both geometrical fidelity and the visual experience. Considering that the error of the measures acquired with the laser scanner is approximately 2 mm, for geometrical documentation purposes an error of 1 cm is considered significant so the high poly mesh must be stored. On the other hand, to document the visual experience, it is only necessary to retain the effect geometry has on lighting. Normal maps allow to retain these effects, although the original geometry is not present in the low poly mesh, and let a rendering engine keep high frame rates. To improve the final quality of the visualised 3D models, Ambient Occlusion maps were also computed.

4 ANNOTATION

Semantic annotation consists of linking abstract concepts to the relevant parts of a 3D mesh. Since these concepts must be represented in standard formats, in order to be recognisable, a reference source must be selected. In our case, we refer to the Art and Architecture Thesaurus (AAT) [7]: a controlled domain vocabulary containing generic terms and other data concerning the represented concepts. These are connected using hierarchical, equivalence and associative relationships. The annotation consists in the use of maps describing semantic concepts applied to the 3D model like a texture, thus avoiding the need to geometrically segment the architectural artefact. For each concept in the AAT found in the digital architectural model, a

Figure 4: A semantic map for the *pediment* concept.

semantic map is created and assigned the same unique ID the concept it represents is recorded with in the AAT. As an improvement with respect to previous approaches, the semantic information is represented as a grayscale map: each map records which polygons, in the digital model, are relevant for the concept it represents by using the model's UV map. In our approach, white indicates high relevance, while black indicates no relevance. An example of semantic map is shown in Figure 4. Using semantic maps and reference IDs for the annotated concepts allows the integration of multiple sources of information (texts, images, audio recordings, etc…) sharing the same annotation scheme. Cross-referencing these sources opens the possibility to produce advanced interfaces to link the descriptions a specific artefact has in separate domains.

The possibility of using gradients in the map lets annotators refine the quality of the semantic data. This way, it is possible to express, more than a binary relevance of each vertex for a given concept, a relevance *level* for that concept. This is important in the field of architectural heritage, as it is not always possible to classify an element in a unique and precise way, and becomes useful when an architectural element cannot be assigned to a specific category. The same applies to situations in which it is not possible to indicate where, exactly, an architectural element becomes another one. This also makes it possible to consider semantic maps produced by multiple annotators to obtain a final map by computing the mean values for each UV coordinate, similarly to what has been done in other fields where annotation uncertainty is important, like for emotions [4].

5 INTERACTION MANAGEMENT

In the scenario of automated information providers for architectural heritage, semantically annotated material can be used to generate cross-domain presentations. Moreover, the possibility of adding virtual characters to the scene allows the elicitation of social signals and the use of natural, multimodal commands. To take advantage

of these possibilities, we designed a software architecture combining specialised modules for interaction management and knowledge representation. This architecture includes: a) a graph database (Neo4J) to represent knowledge, b) a dialog manager (Opendial) to handle interaction, c) a game engine (Unreal Engine 4) to control the virtual character, d) a voice synthesizer (Mivoq) and e) a Kinect sensor to collect users' activity data.

Neo4J [9] is an open source graph database manager that has been applied to a high number of tasks related to data representation (e.g. [2]). Opendial [3] is a dialogue management framework based on probabilistic rules aiming at merging the best of rule-based and probabilistic dialogue management. Probabilistic rules, in Opendial, are used to setup and update a Bayesian network consisting of variables that represent the current dialogue state, including uncertainty. A utility based approach is used to compute the next system action, if any. The virtual avatar and the real time rendering of the obtained artefacts is controlled using the Unreal Engine 4[2]. The voice of the avatar is dynamically generated using the Mivoq Voice Synthesis Engine[3]. User gestures and speech are detected with the Kinect sensor and are continuously forwarded to the game engine.

With this approach, the task of handling raw user data is assigned to a module designed to manage complex, dynamic interfaces that include video, audio and user control systems, while high-level decision processes are delegated to the dialogue manager. This component accesses the encyclopedic knowledge represented in the graph database to select the most appropriate response to a user input. The response consists of an abstract *plan* that may include text extracted from the knowledge base, clarification requests or generic action instructions (e.g. *enter another environment*). Deciding how to implement the action is assigned to the game engine. This choice is motivated by the dynamic nature of the interaction between the users and the avatar, but also between the avatar and the 3D surroundings: a reactive behavioural logic is needed to manage interrupts caused by both implicit or explicit user activity. Also, social signals generation and monitoring must be performed in real time to ensure consistency with the users' behaviour. Lastly, the relative position of the avatar with respect to the concepts that are relevant for the generated utterances must also be evaluated in real time in order to generate pointing gestures.

To support multimodal commands from the users and allow a richer interaction, the user skeletons provided by the Kinect sensor will be used: by exploiting the raycasting system included in the engine, it is possible to emit a single, invisible, ray of light from the tip of the arm bone to capture collision events between the ray and the objects in the scene. From the data included in the collision event, it is, then, possible to extrapolate the UV coordinates of the vertex that is closest to the collision point. These UV coordinates can then be used to query the semantic maps of the object the ray collided with to extract relevance information for the annotated concepts. These can then be passed to the dialog manager when a speech command is detected and multimodal fusion has been performed. The details of the interaction management strategy will be formalised on the basis of audiovisual recordings of expert art historians presenting the Campanian Charterhouses to small groups of visitors, which are currently being collected in the framework of the CHROME project.

[2]www.unrealengine.com
[3]www.mivoq.it

6 CONCLUSIONS AND FUTURE WORK

We have presented the work in progress in the framework of the CHROME project. We have described the work flow leading from 3D architectural data collected with laser scanners and photogrammetry to an interactive system designed to present such data in a rich and entertaining way. Using high and low poly meshes with normal maps to retain the necessary details in real time rendering, we document architectural heritage from a geometrical and from a visual experience points of view. Furthermore, an original method to semantically annotate the low poly meshes has been developed to allow a direct link between concepts in the AAT thesaurus and geometric parts, introducing the possibility to represent uncertainty. The semantic data produced with this work flow will allow the development of 3D conversational agents able to refer to the reconstructed environment.

7 ACKNOWLEDGMENTS

Antonio Origlia's work is funded by the Italian PRIN project *Cultural Heritage Resources Orienting Multimodal Experience* (CHROME) #B52F15000450001.

REFERENCES

[1] Livio De Luca. 2006. *Relevé et multi-représentations du patrimoine architectural Définition d'une approche hybride pour la reconstruction 3D d'édifices.* Ph.D. Dissertation. Sciences de l'Homme et Société. Arts et Métiers ParisTech.
[2] Felix Dietze, Johannes Karoff, André Calero Valdez, Martina Ziefle, Christoph Greven, and Ulrik Schroeder. 2016. An Open-Source Object-Graph-Mapping Framework for Neo4j and Scala: Renesca. In *International Conference on Availability, Reliability, and Security*. Springer, 204–218.
[3] Pierre Lison and Casey Kennington. 2016. OpenDial: A toolkit for developing spoken dialogue systems with probabilistic rules. *ACL 2016* (2016), 67.
[4] G. McKeown, M. F. Valstar, R. Cowie, and Maja Pantic. 2010. The SEMAINE Corpus of Emotionally Coloured Character Interactions. In *Proc. of ICME*. 1079–1084.
[5] Tommy Messaoudi, Philippe Véron, Gilles Halin, and Livio De Luca. 2018. An ontological model for the reality-based 3D annotation of heritage building conservation state. *Journal of Cultural Heritage* 29 (2018), 100–112.
[6] Antonio Origlia, Piero Cosi, Antonio Rodà, and Claudio Zmarich. 2017. A dialogue-based software architecture for gamified discrimination tests. In *Proc. of GHItaly*. http://ceur-ws.org/Vol-1956/
[7] Toni Petersen. 1990. Developing a New Thesaurus for Art and Architecture. *Library Trends* 38, 4 (1990), 644–658.
[8] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and vision computing* 27, 12 (2009), 1743–1759.
[9] Jim Webber. 2012. A programmatic introduction to Neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*. ACM, 217–218.
[10] Daniel Wigdor and Dennis Wixon. 2011. *Brave NUI world: designing natural user interfaces for touch and gesture.* Elsevier.