Natural Experiences in Museums through Virtual Reality and Voice Commands

Natural Experiences in Museums through Virtual Reality and Voice Commands

Andrea Ferracani, Marco Faustino, Gabriele Xavier Giannini, Lea Landucci, Alberto Del Bimbo
Media Integration and Communication Center - University of Florence
Firenze, Italy
[name.surname]@unifi.it

ABSTRACT

In this demo we present a system for immersive experiences in museums using Voice Commands (VCs) and Virtual Reality (VR). The system has been specifically designed for use by people with motor disabilities. Natural interaction is provided through Automatic Speech Recognition (ASR) and allows to experience VR environments wearing an Head Mounted Display (HMD), i.e. the Oculus Rift. Insights gathered during the implementation and results from an initial usability evaluation are reported.

CCS CONCEPTS

•Information systems → Multimedia information systems; Users and interactive retrieval; Query languages; **•Human-centered computing → Virtual reality; Interaction techniques;**

KEYWORDS

Voice Commands, Virtual Reality, Cultural Heritage, Museum Experience, Automatic Speech Recognition, Head-Mounted Display

ACM Reference format:
Andrea Ferracani, Marco Faustino, Gabriele Xavier Giannini, Lea Landucci, Alberto Del Bimbo. 2016. Natural Experiences in Museums through Virtual Reality and Voice Commands. In *Proceedings of MM'17, October 23–27, 2017, Mountain View, CA, USA.*, , 2 pages.
DOI: https://doi.org/10.1145/3123266.3127916

1 INTRODUCTION

Nowadays, personalized mobile museum guides, augmented reality systems featuring see-through technology and HMD VR systems are the most popular trends for providing visitors with rich context-aware information in cultural heritage apps. However, such technologies pose limitations to users with motor disabilities as they assume the ability to hold a device or to move and interact with the surrounding real or virtual Immersive Museum Environment (IME) through controllers or natural gestures.In latest years there has been a significant raise of voice interaction in games and 'serious games'. This is due to the proliferation of consumer devices with built-in capabilities for Automatic Speech Recognition (ASR) such as the Microsoft Kinect as well as to improvements of these systems in terms of recognition rates. However, though voice interaction has long been of interest to HCI as perceived like a natural way of communicating with a computer, it has not yet freed itself from being regarded as a supplement to traditional controller-based or gestural input. In fact, there is still little research on how to exploit progress in ASR for developing effective and accessible speech controlled interfaces.Nevertheless, some examples exist of humanoid conversational agents in Museum applications, but dialogue is poorly supported and, also in advanced immersive solutions which exploit HMD, is restricted to few words.In this regard, there are still significant issues related to the use of VCs in 'games' and interactive exhibits that can be summarized as follows: 1) perceived distance between the player and the game character defined as 'identity dissonance' in [2]; 2) the social context where voice interaction takes place (e.g. the quiet environment of museums, privacy concerns); 3) errors in ASR (due to noise, spelling, etc.); 4) restricted freedom of speech in limited domain applications with VCs constituted by simple words or short phrases due to the difficulty of ASR in the wild.

In this demo we propose some ideas on how to alleviate these issues experiencing an IME displayed through an HMD and made walkable using VCs. The system was conceived as a natural interface for users with motor disabilities, so that they can visit a museum not only remotely but also exploiting VCs exclusively. The player can navigate the museum and obtain information through Voice Commands to a Virtual Museum Guide agent (VMG). Commands have a certain degree of freedom since are automatically fed and augmented *via* a semantic storage provided with a reasoner capable of inferring concepts.

2 THE SYSTEM

The system[1] is composed by three main modules, implemented in a library for Unity 3D[2], respectively in charge of: 1) importing and setting up the IME; 2) performing ASR, augmenting and detecting Voice Commands; 3) allowing interaction and navigation in the environment.

Setting up the Environment. The library allows to insert and place artworks (paintings and sculptures) in a 3D Museum model using Unity scripts, that can be attached to 3D objects. Artworks can be described using triples $\{s, p, o\}$ through ontologies imported in or created by the system (e.g. specifying image URIs, authors and artistic movement artists belong to). Possible questions can be defined as instances of the *Question* class through a script attached to the First Person Controller. Multiple ontologies can be used and extended creating new classes, instances and properties which support both literals and resources. For parsing and managing ontologies and triples in Unity the system exploits the dotNetRDF

MM'17, October 23–27, 2017, Mountain View, CA, USA.

DOI: https://doi.org/10.1145/3123266.3127916

[1]Demo video available at https://vimeo.com/miccunifi/museum-voice-commands
[2]https://unity3d.com

Opensource Library[3]. Statements are saved in N-Triples format and then imported in the Apache TDB Jena semantic storage[4].

Speech Recognition. Speech Recognition is performed by the System exploiting the Microsoft Speech API (SAPI) 5.3, the native API for Windows[5], and mapping VCs to a dynamic grammar using rules. This is done in order to allow the user to ask questions and express commands in the virtual space. Rules define patterns and word sequences to be matched against the vocal input. Rules are represented as a graph of states. States (or group of words) are part of a sentence which mark a particular part-of-speech in the context (they identify the relationship of a state with adjacent and related states in the sentence; e.g. subject, predicate and direct objects). Rules and patterns are described in an XML-format grammar that conforms to the Microsoft Speech Recognition Grammar Specification (SRGS) Version 1.0. The grammar contains variants of interrogative, exhortative and desiderative sentences and is dynamically created through SPARQL queries. In this way, questions and requests by the user in the domain are intended as voice commands by the natural interface. A predefined set of instances of a *vc:Question* and *vc:Request* classes has been provided with the library. Requests and questions have three default properties which are *vc:hasSubject*, *vc:hasPredicate* and *vc:hasDirectObject*. The instances of these classes *vc:hasSubject rdfs:range vc:Character*; *vc:hasPredicate rdfs:range vc:Predicate* and *vc:hasDirectObject rdfs:range vc:Artwork, vc:Artist, vc:ArtisticMovement*. Direct objects are resources retrieved dynamically form the semantic storage via SPARQL and added setting up the environment in the Unity 3D Editor (e.g. "I'd like to see 'The Scream' by Edvard Munch"). Inference is also provided by the system. For example, given that:

```
Class(vc:ActionPainter complete intersectionOf(vc:Artist
  restriction(vc:exponentOf someValuesFrom(a:ActionPainting))))
Class(vc:AbstractPainter complete intersectionOf(vc:Artist
  restriction(vc:exponentOf someValuesFrom (vc:AbstractArt))))
Class(vc:ActionPainting partial vc:AbstractArt)
```

The following class inference can be derived:

- an Action Painter is an exponent of the Action Painting;
- Action Painting is a type of Abstract Art;
- an Action Painter is an exponent of the Abstract Art, so must be an Abstract Painter.

When concepts are inferred, the grammar is updated and rules added so that the user may ask additional questions such as "Which types of abstract art are present in the museum?" or "Is Jackson Pollock an abstract painter?". The inference engine solves in part the issue n. 4 expressed in Sec. 1 allowing more flexibility in questions and commands.

Interaction and Visualization. The virtual museum is visualized through the Oculus Rift which provides immersion. The system has been designed to be used by people with motor disabilities in their own rooms or in a dedicated private space in this way excluding the social context of the interaction, and consequently embarrassment and privacy concerns related to VCs, and environmental noise (i.e. issue n. 2 and n. 3 in Sec.1). In order to increase naturalness, the user experiences the environment as a First Person Viewer. He is guided inside the museum by a VMG agent to whom he can ask questions using voice. In this scenario, the player has not to embody himself with a virtual representation. This mitigates the 'identity dissonance' issue (i.e. n. 1 in Sec. 1). Furthermore, verbal immediacy is demonstrated to have a significant impact on learning and sense of presence in IMEs. To ease the user interaction, upon first access the agent lists the possible vocal questions the virtual visitor can ask. There are three default question instances in the semantic storage: 1) "Which artistic movements are displayed in the museum?"; 2) "What artists are there in the museum?" and 3) "What artworks are there in the museum?". Additional questions, that dynamically populate the grammar for ASR, can be added manually through ontologies or inferred by the reasoner. Text-To-Speech (TTS) synthesis is used by the guide to explain possible questions and to give responses. Once the user has asked the question of interest, the ASR takes the audio stream as input and turns it into a text transcription. Acoustic models, lexicons and language models are used to search the best match of the input with the textual instances present in the grammar. Let's say that the user ask question 1). The question is interpreted as a VC and mapped to a SPARQL query. Consequently, the guide will list all the pertinent information retrieved or inferred by the reasoner to the user using TTS. Then she will ask the user which artistic movement he is interested in. So the conversation can go on, and the user can make new requests (desiderative or imperative) to the agent who can satisfy them in two ways: 1) explaining the concept and asking new questions (e.g. listing all the artists of a particular movement) or 2) guiding the user to and describing an artwork of interest if he expresses the desire to know more about it (e.g. "I'd like to see 'The Starry Night' by Vincent Van Gogh"). In the latter case, the VMG guides the visitor to the place where the artwork is located walking through the halls of the museum. The idea is to give the user the natural impression of following behind a guide while she explains what she and the visitor are going to see. To make the guide move naturally through the museum environment avoiding obstacles (e.g. walls, sculptures) the A* algorithm is exploited. A* is an algorithm for path finding which can compute the shortest path between *vertices* in a graph. Given the 2D museum map, all the walkable surface and obstacles are mapped to a fine-grained grid modeled as a graph. The A* algorithm is able to find the least cost path from an initial node to a goal node. How the interaction between the player and the guide works is demonstrated in our demo video.

The usability of the system was preliminarily tested using the popular Standard Usability Scale (SUS) [1]. 10 users were asked to perform the task of navigating the museum using VCs obtaining insights from the VMG on at least an artistic movement and an artwork. Average SUS score was 71.0. Scores are in the range $[0-100]$ and over 68 mean that the interaction design is above average [3].

[3] http://www.dotnetrdf.org/
[4] https://jena.apache.org/
[5] http://bit.ly/2qNEnWF

REFERENCES

[1] John Brooke. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* (1996), 189–194.
[2] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player identity dissonance and voice interaction in games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. ACM, 265–269.
[3] Jeff Sauro and James R. Lewis. 2012. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.