Web-Powered Virtual Site Exploration Based on Augmented 360 Degree Video via Gesture-Based Interaction

Maarten Wijnants Gustavo Rovelo Ruiz Donald Degraen
Peter Quax Kris Luyten Wim Lamotte
Hasselt University – tUL – iMinds, Expertise Centre for Digital Media
Wetenschapspark 2, 3590 Diepenbeek, Belgium
firstname.lastname@uhasselt.be

ABSTRACT

Physically attending an event or visiting a venue might not always be practically feasible (e.g., due to travel overhead). This article presents a system that enables users to remotely navigate in and interact with a real-world site using 360° video as primary content format. To showcase the system, a demonstrator has been built that affords virtual exploration of a Belgian museum. The system blends contributions from multiple research disciplines into a holistic solution. Constituting technological building blocks include 360° video, the Augmented Video Viewing (AVV) methodology that allows for Web-driven annotation of video content, a walk-up-and-use mid-air gesture tracking system to enable natural user interaction with the system, Non-Linear Video (NLV) constructs to unlock semi-free visitation of the physical site, and the MPEG-DASH (Dynamic Adaptive Streaming over HTTP) standard for adaptive media delivery purposes. The system's feature list will be enumerated and a high-level discussion of its technological foundations will be provided. The resulting solution is completely HTML5-compliant and therefore portable to a gamut of devices.

Author Keywords

Virtual exploration; HTML5; 360° video; Non-Linear Video; gestural interaction; MPEG-DASH; Augmented Video Viewing.

ACM Classification Keywords

C.2.5 Computer-Communication Networks: Local and Wide-Area Networks—*Internet*; H.5.1 Information Interfaces and Presentation: Multimedia Information Systems—*Artificial, augmented, and virtual realities*; H.5.2 Information Interfaces and Presentation: User Interfaces—*Input devices and strategies, Interaction styles*; H.5.4 Information Interfaces and Presentation: Hypertext/Hypermedia

INTRODUCTION AND MOTIVATION

There is an increasing tendency to disclose real-world events and spaces in the virtual realm, to enable hindered people to participate from a remote location. Systems that allow for cyber presence at physically distant sites hold value for heterogeneous application domains, including tourism, entertainment and education.

3rd International Workshop on Interactive Content Consumption at TVX'15, June 3rd, 2015, Brussels, Belgium.

The last few years, technological advances in divergent research disciplines have emerged that hold great promise to increase the persuasiveness of interactive video-driven virtual explorations. A first important example is situated in the video capturing field, in the form of 360° video authoring. 360° video cameras produce video footage that has an omni-directional (i.e., cylindrical or spherical) Field of View. Compared to classical video, 360° video content unlocks options for increased user immersion and engagement. Secondly, the traditional keyboard/mouse interaction technique is recently witnessing increasing competition from more natural alternatives like, for example, touch- or gesture-based schemes. In this context, an exploratory study performed by Bleumers et al. has found mid-air gesture-based interaction to be the preferred input method for the consumption of the 360° video content format [1]. A final notable evolution is the increasing maturity of the Web as a platform for media dissemination and consumption. The HTML5 specification for instance covers all necessary tools to develop a 360° video player that affords typical Pan-Tilt-Zoom (PTZ) adaptation of the viewing angle inside the omni-directionally captured video scene.

SYSTEM OVERVIEW AND USE CASE DESCRIPTION

The work described in this article can best be summarized as offering an interactive multimedia content consumption experience akin to Google Street View, yet hereby relying on 360° video instead of static imagery, and at the same time offering advanced interaction opportunities that go beyond basic "on rails" virtual navigation. The specific use case that will be focused on in this manuscript, is the virtual visitation of a particular Belgian museum. Users can move along predefined paths that have been video captured in 360 degrees. When reaching the end of such a path, users are offered the choice between a number of alternative contiguous traveling directions, very much like choosing a direction at a crossroad. As such, a NLV scenario arises in which the user is granted a considerable amount of freedom to tour the museum at his personal discretion. While navigating along the predetermined paths, users can play/pause the video sequence, dynamically change their viewing direction, and perform zoom operations.

The use case furthermore includes a gamification component in the form of simple "treasure hunting" gameplay. In particular, users are encouraged to look for salient museum items that have been transparently annotated to increase user engagement with the content. Finally, both mouse-based and gestural interaction is supported by the use case demonstrator. The two control interfaces are identical expressive-wise, in the sense that they grant access to the exact same set of functionality (i.e., direct manipulation

of the user's viewport into the 360° video content, making navigational decisions and performing gamification interaction through pointing and selection, video playback control).

IMPLEMENTATION

Except for its gesture-related functionality, the demonstrator has exclusively been realized using platform independent Web standards. The typical execution environment of the player component of the demonstrator is therefore a Web browser.

The involved media content was recorded using an omni-directional sensor setup consisting of 7 GoPro Hero3+ Black cameras mounted in a 360Heros rig (`http://www.360heros.com/`). The resulting video material was temporally segmented according to the physical layout of the museum in order to yield individual clips for each of the traversable paths. The collection of paths (and their mutual relationships) is encoded as a directed graph. This graph dictates the branching options in the NLV playback.

Media streaming is implemented by means of MPEG-DASH. The separate video clips from the content authoring phase were each transcoded into multiple qualities, temporally split into consecutive media segments of identical duration (e.g., 2 seconds), and described by means of a MPD. The resulting content was published by hosting it on an off-the-shelf HTTP server. The W3C Media Source Extensions specification is exploited to allow for the HTML5-powered decoding and rendering of media segments that are downloaded in an adaptive fashion using JavaScript code.

While the playback of a path is active, the initial media segments that pertain to each of the potential follow-up routes (as derived from the NLV graph representation) are pre-fetched from the HTTP server. The total number of media segments to pre-fetch is dictated by the corresponding path's `minBufferTime` MPD attribute. By making initial media data locally available ahead of time, the startup delay of the selected follow-up path is minimized.

The gesture set that was defined for the demonstrator consists of composite gestures in the sense that they involve performing a sequence of discrete, gradually refining postures. As such, it becomes feasible to organize the available gestures in a tree-like topology, where intermediate layers represent necessary steps towards reaching a leaf node, at which point the gesture (and its corresponding action) is actually actuated. It also allows gesture clustering and organization on the basis of their respective sequence of encompassed postures. Two gestures whose posture series are identical up to some intermediate level, share a branch in the tree up to that level and only then diverge topologically.

The gestural interface is implemented by means of a mid-air gesture recognizer (which currently relies on a Kinect 2.0 for skeleton tracking purposes). It adheres to a walk-up-and-use design, which implies that it provides supportive measures that empower users to leverage the system without requiring training. The supportive measures take the form of a hierarchical gesture guidance system that exploits the tree-like organization of the gesture set to visually walk the user through the subsequent steps needed to perform a particular gesture (see Figure 1).

To encode and present the navigation options at NLV decision points and to add interactivity to the treasure hunt objects that appear in the video footage, the demonstrator resorts to the AVV

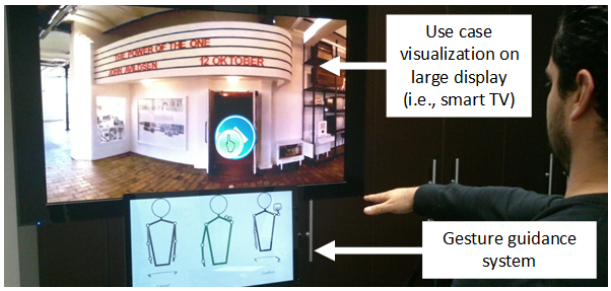

Figure 1. Gesture-based interaction with the demonstrator.

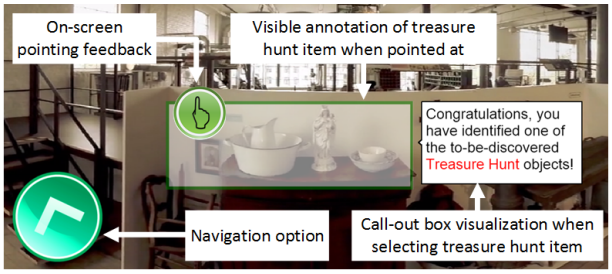

Figure 2. Three applications of the AVV methodology in the demonstrator.

methodology [2]. This methodology (and its Web-compliant implementation) is intended to transform video consumption from a passive into a more lean-forward type of experience by providing the ability to superimpose (potentially interactive) overlays on top of the media content. The navigation options are represented as arrows indicating the direction of potential follow-up paths; their visualization is toggled when the playback of the current path is about to end. The treasure hunt objects on the other hand are (invisibly) annotated by means of a transparent overlay. When the user points to such an object, the visual style of the associated AVV annotation is on-the-fly transformed (through CSS operations) into a semi-transparent one. If the item is subsequently selected, an informative AVV-managed call-out widget is visualized. Finally, the AVV methodology is also exploited to present visual feedback of the user's current pointing location. This is realized by dynamically instantiating an overlay (carrying a green hand icon) as soon as the user enters pointing mode and by continuously updating its coordinates as the pointing operation is being performed. When pointing offscreen (only possible with the gestural interface), the hand icon is clamped to the nearest on-screen position and turns red. Figure 2 illustrates the three applications of the AVV approach in the demonstrator.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610370, ICoSOLE ("Immersive Coverage of Spatially Outspread Live Events", `http://www.icosole.eu`).

REFERENCES

1. Bleumers, L., Van den Broeck, W., Lievens, B., and Pierson, J. Seeing the Bigger Picture: A User Perspective on 360° TV. In *Proc. EuroITV 2012*, ACM (2012), 115–124.
2. Wijnants, M., Leën, J., Quax, P., and Lamotte, W. Augmented Video Viewing: Transforming Video Consumption into an Active Experience. In *Proc. MMSys 2014*, ACM (2014), 164–167.