REALISTIC NATURAL INTERACTION WITH VIRTUAL STATUES IN X-REALITY ENVIRONMENTS

G. Margetis[1], G. Papagiannakis[1,2], Constantine Stephanidis[1,2]

[1] Foundation for Research and Technology – Hellas (FORTH), Institute of Computer Science (ICS), Heraklion, Crete, GR-70013 - (gmarget, papagian, cs)@ics.forth.gr
[2] University of Crete, Department of Computer Science, Greece

KEY WORDS: X-Reality, Diminished Reality, Spoken Dialogues, Natural Interaction, 3D model realistic deformation, Virtual Statues, Cultural Heritage

ABSTRACT:

Augmented, Virtual, and Mixed Reality have successfully transitioned from the laboratory environment to becoming mainstream technologies available to almost everyone, mainly due to important advancements in terms of hardware. These technologies, along with recent advancements in Artificial Intelligence, have the potential to deliver compelling experiences. In the Cultural Heritage domain, they can be used to achieve natural interaction with virtual statues, making visitors feel as if physical statues "come to life" and can engage in dialogue with them. This paper presents the fundamental technological components towards this vision and how they can be orchestrated through the description of a conceptual architecture and a use case example. Open research issues are also discussed, providing a roadmap for future research in this area.

1. INTRODUCTION

The recent proliferation of affordable Virtual Reality (VR) and Augmented Reality (AR) devices paved the way for leveraging humans' interplay with digital media and assets, towards the unification of the various extended reality facets that are available today. For example, the popularity and success of Pokemon GO can be largely attributed, among others, to the availability of low-cost components embedded in smartphones, providing to users the ability to interact with AR, nearly anywhere in their physical world.

On the other hand, technological advancements – as they have resulted from multidisciplinary research conducted in the fields of Computer Vision (e.g. body tracking, hands articulation, face detection) and Artificial Intelligence (AI) (e.g. natural language processing - NLP - and generation, conversational systems) – can act as key enablers for enhancing users' experience with AR, VR and Mixed Reality (MR) applications. AR, VR and MR applications are collectively referred to as X-Reality (Extended Reality) or XR applications (Fast-Berglund et al., 2018).

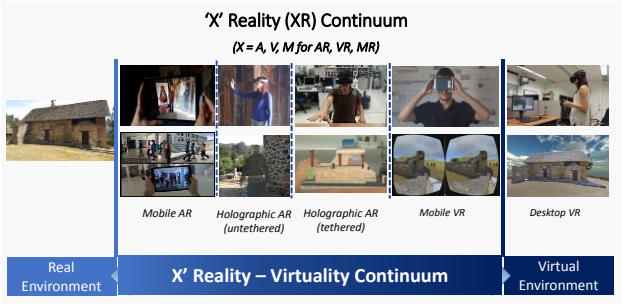

Figure 1. Extended Reality (X-Reality or XR) taxonomy diagram

Figure 1 illustrates the X-Reality taxonomy diagram, including the technological approaches that lie within the Real and the Virtual Environment. These technological approaches require a spectrum of diverse interdisciplinary research topics in order to be substantiated, such as computer vision (e.g., diminished reality), high-end computer graphics (e.g., real-time rendering and animation of 3D realistic models) as well as natural interaction (e.g., gestures and spoken dialogues with virtual characters).

Technologies, such as 3D scanning and computer vision, enable real-time rendering and facilitate the development of better AR and MR devices. Modern headsets include built-in tracking systems commonly based on GPS, Bluetooth/Wi-Fi sensors, or optical sensors, offering high levels of accuracy and precision. Additionally, newly introduced and rapidly advancing technologies, such as speech recognition and gesture recognition, provide natural means of interaction for AR and MR headsets. MR and AR platforms featuring ready-to-use software development kits (such as Apple's ARKit[1], Google's ARCore[2], and Snapchat Lens Studio[3]) have emerged to facilitate and improve MR and AR application development, thus enabling companies to reach different customer bases and produce customizable AR and MR solutions for different industry verticals. Moreover, the technological advancements in this field are entailing new ways for better and more affordable MR and AR devices, thus increasing its outreach to a significantly wider consumer group.

XR technologies, in combination with AI, open ground-breaking opportunities across multiple domains, with cultural heritage being one of the most prominent ones. In particular, the adoption of technologies in the fields of realistic 3D modelling, rendering and animation, diminished and mediated reality, as well as natural interaction can lead to a paradigm shift in terms of interpreting, visualizing and interacting with elements of cultural heritage artefacts. The use of such technologies has the

[1] https://developer.apple.com/arkit/
[2] https://developers.google.com/ar/
[3] https://lensstudio.snapchat.com/

potential to enrich the information of cultural heritage artefacts, such as statues or other museum exhibits. The enriched information (e.g. their shape, color, materials, constructive elements, original representation) can be presented in a seamless fashion across the physical and virtual space.

Additionally, XR technology can be used to introduce virtual guides to present the history of an exhibit or even have the exhibit come to life and share its own story. In the field of Cultural Heritage, the idea of creating virtual agents who will tell the story of the exhibits of a museum or an archaeological site is quite old. However, the focus of previous efforts has been on fictional characters that reside at the same place with the users and that engage in interplay with the users, through AR approaches. There are story-telling agents (Ibanez et al., 2003) and virtual augmented characters who re-enact dramatic events (Ioannides et al., 2017); but, to the best of our knowledge, a holistic approach that brings the archaeological artefact into life in a realistic manner enabling their natural interaction with the visitors does not exist.

One important challenge that cultural heritage environments enriched with XR technologies should address is to captivate users' presence. Presence can be defined as a psychological perception of being immersed in the XR environment (Sanchez-Vives et al., 2005) and is essential for engagement and cognitive connection to the content. This involves high quality and authentic/certified content, which is relevant and coherent in terms of the social and cultural context, including aspects such as cultural values, recognition and significance, representation of emotional intelligence, semantic time, space, provenance and uncertainty (Ioannides et al., 2018).

This paper discusses the fundamental technological prerequisites for leveraging end-users' presence in cultural heritage spaces through XR technologies, focusing on realistic natural interaction with statues. Through the proposed approach, passive visitors are turned into active participants engaged in an interactive and immersive blend of physical and virtual as if it was a single unified "world". Moreover, this paper discusses the fundamental challenges toward the realization of such a conception, as well as potential directions for addressing them, and proposes a conceptual architecture that harnesses the necessary technologies within one concrete system. The paper is structured as follows: Section 2 discusses related work in the fields of diminished reality, virtual character reconstruction, and natural interaction with virtual agents; Section 3 presents the conceptual architecture; Section 4 discusses conclusions and provides direction for future research.

2. RELATED WORK

This section discusses related work and fundamental challenges toward the realization of the proposed conception.

2.1 Diminished Reality

The idea of blending user experiences between the real and virtual (digital) world entails the capability of fading real parts of the environment and substituting them with a plausible background. The notion of diminished reality was coined by the wearable computing pioneer Steve Mann and describes a reality that can remove, at will, certain undesired aspects of regular reality (Mann et al., 2001). During the last decades with the proliferation of AR applications, several diminished reality approaches have emerged that can be clustered in two main categories: those that require prepared structure information and registered photos and those that achieve real-time processing and work without pre-processing the target scene.

Zokai et al. introduce a technique for removing an object or collection of objects and replacing it with an appropriate background image (Zokai et al., 2003). It uses multiple calibrated views of a scene to remove an object or region of interest and to replace it with the correct background, by using a paraperspective projection model and has flexibility to recover a crude or fine detailed background. Although the aforementioned algorithm provides adequately good results, it requires the use of multiple fixed cameras, capturing the target from different angles.

Enomoto and Saito (2007) propose a system for diminished reality, which is based on multiple handheld cameras, thus providing bigger degrees of freedom. Although the proposed system provides the appropriate resilience, even in cases of moving cameras, moving obstacles or changing objective scene, it is solely based on ARTags (Fiala, 2005) for combining the images acquired by different handheld cameras, which makes it an unrealistic approach for contemporary AR applications.

Siltanen, (2006) proposes method of generating a photorealistic texture, in real-time, for hiding distracting planar markers used for image registration by AR applications. Another real-time capable Diminished Reality approach for high-quality image manipulation is PixMix (Herling et al., 2010), achieving a significantly good performance and image quality for almost planar but non-trivial image backgrounds. Meerits and Hideo (2015) propose an interesting diminished reality multi-camera system utilizing an RGB-D camera to hide arbitrary trackable objects from a scene. The scene background does not have to be planar for the system to work and scene changes can be handled in real time.

However, such techniques conclude to be inappropriate approaches for realistic interaction in real environments, since they either require fixed calibrated cameras, or are based on unrelated to the environment planar markers. Furthermore, the objects to be replaced by their 3D alternatives are usually non-planar, as well as the surface behind the object—which can also be far away from the objects to remove.

In this respect, a new diminished reality method for 3D scenes considering background structures has been proposed (Kawai, et al., 2013), which is not constrained by the assumption of planar features. Instead, it approximates the background structure by the combination of local planes, correcting the perspective distortion of texture and limiting the searching area for improving the quality of image inpainting.

An alternative approach of diminished reality is based on 3D information provided by large photo collections on the Internet (Li et al., 2013). Specifically, as a first step the algorithm is making use of internet photos, registering them in 3D space and obtaining the 3D scene structure in an offline process. Then, reference images of the scene are selected and unobstructed regions are detected automatically; the patches synthesized by one transformation warping from the reference images are more realistic than those by pixel-wise reconstruction methods.

Lepetit et al. (2000) present a different approach for outlining 3D objects in real environments, in order to be able to be

substituted with virtual 3D models. Given a video sequence, starting from an outline of the target in all the images of the sequence, the method replaces it by a 3D reconstruction of the scene, constructed from the same video sequence. One of the main strengths of the algorithm concerns its ability to handle uncertainties on the computed motion between two frames. However, the method was designed to proceed offline, which makes the viewpoint computation and the object detection much easier than in the case of real-time conditions.

Although different approaches have been studied for realizing realistic inpainting and providing diminished reality for AR applications, there are shortcomings for realistic and real time elimination of obstacles and substitution by plausible background. In light of AI techniques able to address difficult computer vision challenges efficiently and quickly, the concept of diminished reality can be tackled by methods based on 3D geometry and Deep Learning to remove objects from the user's point of view in real-time. These approaches should avoid visible artefacts, by developing techniques to match rendering of the background with the viewing conditions.

2.2 Realistic reconstruction of 3D models from the real world

Virtual characters play a fundamental role for attaining a high level of believability in XR environments and they are a key-element for transferring knowledge and presenting scenarios in different Cultural Heritage applications. In this respect, the immediate interaction of users with realistic virtual narrators, representing historical personalities or infamous artwork, plays a vital role in the presentation of Cultural Heritage. Embodied agents providing instructions to the users and transferring knowledge about the history of the Cultural Heritage artefact, through realistic interplay which imitates human-to-human communication, can endorse the users' feeling of presence in the XR environment, which tend to prevail against other approaches for transferring knowledge to the users through immersive and interactive experiences.

Regarding 3D reconstruction, Papaefthymiou et al. (2017) compare three different 3D reconstruction techniques, namely using Agisoft Photoscan Software, Fast Avatar Capture application and the Occipital Structure Sensor. They propose that the third technique, in which the structure sensor is used, is the most efficient, since it is fast and provides a more satisfying and robust result compared to the other two techniques. The interactive reconstructed virtual characters of this technique support a wide range of different behaviors like performing gestures, speech and lip synchronization. This is also applicable to 3D models that are reconstructed out of statues.

Figure 2. 3D reconstruction of the priest of the Asinou church (middle, right) using the occipital structure sensor (left)

Following this technique, the 3D model, either human-like or not, is scanned with this sensor (which is connected on a mobile device – such as an iPad). The reconstructed 3D model appears on the screen of the mobile device, so that the scanning procedure can stop when a satisfying result is obtained. This technique has been used in the case of the priest of the Asinou church in Cyprus, for the purposes of the ITN-DCH Cultural Heritage project (Papaefthymiou et al 2017, Ioannidis et al 2017) (see Figure 2).

For the realization of the proposed concept, state of the art 3D reconstruction techniques should be further enhanced, in order to produce realistic 3D reconstructions that will be appropriate for deployment in real-time XR environments. Practically, the 3D reconstruction techniques that will be used for each object depend on the object itself (dimensions, type of object, etc.). Each time the correct 3D reconstruction technique is suggested to be used, in order to achieve the best possible result in the lowest possible time. The output of the reconstruction process in general should be a high-quality 3D mesh that contains geometry and texture information for the modelled physical object.

In the context of modelling cultural heritage artefacts, a challenge that needs to be overcome is that antiquities, and especially statues, are often incomplete (e.g. broken, or partly recovered). Therefore, a typical 3D reconstruction technique would fail to produce virtual objects as representations of the original physical objects. Such a restoration to the artifacts' original form would require the expertise of archaeologists and curators, as well as sophisticated 3D modelling skills to manually extend a partially reconstructed model, which is a resource-demanding approach. To this end, semi-automated approaches should be pursued. Such an endeavor still remains an open challenge, which can only be addressed through an interdisciplinary approach involving the fields of archaeology, museology and curation, as well as computer science.

2.3 Real-time mixed-reality virtual character rendering and animation

In the field of XR, rendering of deformable objects and characters, attaining a high level of believability and realism of real-time registration between real scenes and virtual augmentations requires two main aspects for consistent matching: geometry and illumination (Ioannides et al., 2017). First, the camera position-orientation and projection should be consistent; otherwise the object may seem too shortened or skewed (geometrical consistency). Secondly, the lighting and shading of the virtual object needs to be consistent with other objects in the real environment (illumination consistency). Such a combination is crucial for the 'suspension of disbelief' for dynamic scenes in mixed reality. In the past, consistency of geometry has been intensively investigated (Egges, 2007; Zhou, 2008).

On the other hand, few methods have been proposed so far for consistency of real-time illumination to superimpose virtual objects onto an image of a real scene as most methods are targeting offline simulations (Magnenat-Thalmann, 2007). Furthermore, very few research methods are available in the bibliography, in superimposing real-time, dynamic-deformable virtual scenes on real-time AR scenes (Papaefthymiou et al., 2015).

Employing virtual characters as personal and believable dialog partners in multimodal dialogs entails several challenges, because this requires not only a reliable and consistent motion and dialog behavior, but also nonverbal communication and affective components. Besides modelling the "mind" and creating intelligent communication behavior on the encoding side, which is an active field of research in AI (Kasap et al.,

2007), the visual representation of a character including its perceivable behavior, from a decoding perspective, such as facial expressions and gestures, belongs to the domain of computer graphics and likewise implicates many open issues concerning natural communication (Papanikolaou et al., 2015).

Papagiannakis et al. (2014) propose two alternative methodologies for implementing real-time animation interpolation for skinned characters using Geometric Algebra (GA) rotors and show that they achieve smaller computation time, lower memory usage and more visual quality results compared to state-of-the-art animation blending techniques such as quaternion linear blending and dual-quaternion slerp. Moreover, Wareham et al. (2004) propose a method for pose and position interpolation using the Conformal Geometric Algebra model (CGA), which can also be extended to higher-dimension spaces and a method for interpolating smoothly between two or more displacements that include rotation, translation, as well as dilation using CGA for virtual character simulation (Papaefthymiou et al., 2016).

'True AR' has recently been defined to be a modification of the user's perception of their surroundings that cannot be detected by the user (Sandor et al., 2015). The most obvious parameter of the test protocol then should be which senses can be used: even the most sophisticated visual display will immediately fail a test in which users can use their hands to touch objects to tell the real from the virtual.

The proposed approach in this work, aims to improve the consistency of the simulated world with actual reality. We refer to True Mediated Reality as the means of positioning 3D True-AR models in the real world in a very veritable manner, leading to people not being able to notice that the model they are looking at is actually a 3D augmented model. True Mediated Reality will be used as a means of presenting 3D models of statues perfectly adapted to the real-world environment. The 3D models of the statues will also support nonverbal and verbal communication, affective components, and behavioral aspects, such as gaze and facial expressions, lip movements, body postures and gestures. Additionally, the proposed approach aims to achieve "suspension of disbelief", which can be obtained through realistic rendering and animation of virtual objects. This requires both geometrical consistency, i.e. having consistent camera position-orientation and projection, as well as illumination consistency, i.e. having the lighting and shading of a virtual object (Kateros et al., 2015) be consistent with other physical world objects (Vacchetti et al., 2004), (Papaefthymiou et al., 2015).

2.4 Interactive virtual characters supporting realistic deformation approaches

A successful substantiation of the concept that historical personalities or anthropomorphic artefacts (e.g., statues) are represented as embodied virtual agents able to share knowledge with users through storytelling, banks on the realistic representation of the model, as well as the convincing imitation of the human behavior and expressions. To that end, not only skinning methods for the real-time animation of deformable virtual agents, are required, but also a systematic approach for providing human like postures, movements, eye gaze as well as modeling of emotions and human behavior are the basic ingredients for persuasive interactive virtual characters.

Regarding real-time model skinning, a lot of research effort has been put the last decades resulting in high level of technology maturity in the domain. Loper et al., (2015) introduce SMPL, a realistic learned model of human body shape, able to create realistic animated human bodies that can represent different body shapes, deform naturally with pose, and exhibit soft-tissue motions like those of real human. The model was trained on thousands of aligned scans of different people in different poses, thus it is able to learn the parameters from large amounts of data while directly minimizing vertex reconstruction error. Another novel method provides automatic estimation of the 3D pose of human bodies as well as their 3D shape from a single unconstrained image, based on the combination of a CNN-based approach and the SMPL model (Bogo et al., 2016). Such approaches pave the way for bringing into the game Deep Learning algorithms that will leverage the pose estimation and 3D model deformation.

No matter how efficient model deformation algorithms for the rendering of humanoid 3D models are, several aspects that imitate the human behavior should also be considered for a convincing performance by virtual agents. Static body posture offers a reliable source of information concerning emotion, and contributes to our understanding of how emotion is expressed through the body (Coulson, 2004). To that end, embodied agents should communicate and interact with the users not only verbally, but also providing emotional cues through their body posture. Furthermore, the gaze direction of agents towards the users systematically influences emotion perception (Adams et al., 2005), thus fostering the agent-human communication.

Virtual agents' body and emotions should be compelled by decision-making mechanisms, capable to lead the agent towards smooth animation transitions and interaction responses. Such mechanisms can be built upon formal representations of human emotions and behaviors. Emotion Markup Language, which was introduced in Schroder et al. (2011) and constitutes a W3C recommendation[4], provides a standardized manner for emotions description and related states. Kopp et al. (2006) propose a three stage model where the stages represent intent planning, behavior planning and behavior realization. They defined the Behavioral Markup Language (BML) and specify the communicative and expressive behaviors traditionally associated with explicit, verbal communication in face-to-face dialog.

Although many achievements have been reached so far, the challenge of interactive virtual characters supporting realistic deformation still remains open. New models should be introduced following a perception-attention-action process for virtual characters in order to improve the naturalness of their behavior. Such models should include: (i) perception capabilities that will allow virtual humans to access knowledge of states of real users and other virtual humans (for example position, gesture and emotion) and information of both real and virtual environments; (ii) attention capabilities that will model the cognitive process of real human to focus on selected information with importance or interest; and (iii) decision-making and motion synthesis for virtual humans.

2.5 Natural interaction with virtual agents

Natural interaction with technology is a much-acclaimed feature that has the potential to ensure optimized user experience, as

[4] https://www.w3.org/TR/emotionml/

people can communicate with technology and explore it like they would with any real world interaction counterpart: through gestures, expressions, movements, and by looking around and manipulating physical stuff (Valli, 2008). Speech, of course, is also a natural interaction modality, that is gaining popularity. Speech-based interaction with virtual agents is addressed by the field of embodied conversational agents that typically combine facial expression, body posture, hand gestures, and speech to provide a more human-like interaction (McTear et al., 2016). Although currently embodied conversational agents remain rather rare, it is indicative of the popularity of dialogue-based systems the fact that more and more applications enrich their classical Graphical User Interfaces (GUI) with personal chat services.

A major concern with regard to natural interaction with virtual agents is the fusion of multiple modalities into such a complex system. In this respect, adaptive multimodality can be employed to support natural input in a dynamically changing context of use, adaptively offering to users the most appropriate and effective input forms at the current interaction context (Stephanidis, 2012). At the same time, multimodal input needs to be semantically interpreted in order to achieve an efficient interaction and appropriate system response. For example, interaction commands (e.g. speech, gestures) addressed to the virtual agent must be distinguished from interactions with co-visitors or friends, which is a challenging endeavor in crowded real-life settings. Finally, the orchestration of multimodal input and system output is also an issue that needs to be addressed. It requires dealing in real-time with the distribution of input and output so as to provide humans with continuous, flexible, and coherent communication, both with the agent and with others, by proportionally using all the available senses and communication channels, while optimizing human and system resources (Emiliani et al., 2005).

2.6 Realistic spoken dialogues between V-statues and end-users

Several of the conversations generated by conversational agents are driven by artificial intelligence, while others have people supporting the conversation. Lately, AI has been refueled with the emergence of Deep Learning and Neural Networks, which boosted the research results in NLP, as well. Several Neural Network approaches pursue realistic spoken dialogue systems, such as Recurrent Neural Networks, Recursive Neural Networks or Deep Reinforced Models and Unsupervised Learning.

In (Kumar, 2016) a Dynamic Neural Network (DMN) is introduced, which processes input sequences and questions, forms episodic memories, and generates relevant answers. The DMN model is a potentially general architecture for a variety of NLP applications, including classification, question answering and sequence modeling.

In (Sutskever et al., 2014), a general end-to-end approach to sequence learning is presented, able to make minimal assumptions on the sequence structure. This approach uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. Interestingly, the proposed method superseded in terms of performance and accuracy standard short term memory based systems.

Additionally, several NLP pipelines, tools and libraries have emerged recently, aiming to help researchers to focus on the high-level summary of their models rather than on the details. The Stanford CoreNLP toolkit (Manning et al., 2014) is widely used by the NLP community, providing annotation-based NLP processing pipeline for the prediction of linguistic structures. AllenNLP (Gardner et al., 2018) is a toolkit for deep learning research in dialog and machine translation.

Although a large number of conversational agent engines have been developed and several are freely available, there are very limited attempts to address diverse groups of Cultural Heritage visitors, e.g. children and the elderly. In this respect, a completely new mind-set must be adopted in designing and developing conversational interfaces, even when a chat might seem so simple. The typical design patterns that are used in GUIs do not work in a conversation-driven interface. In the design of conversational interfaces, natural-language processing (NLP) remains the biggest bottleneck.

3. CONCEPTUAL ARCHITECTURE

A generic approach that will tackle the aspect of realistic interaction of users with virtual statutes, should be built according to a distributed Service oriented Architecture (SoA) that will interweave the different technologies in a flexible and scalable manner and promote reusability, interoperability, and loose coupling among its components. Figure 3 illustrates the fundamental components that such approaches should comprise.

Overall, the conceptual architecture involves two main components, namely “XR rendering” and “Context-sensitive natural interaction with multiple users”. The first is responsible for delivering the XR experience, while the second for perceiving and interpreting user interactions.

In particular, the XR rendering component is responsible for scene registration and localization that is for identifying the user’s location in the physical environment and objects in their field of view, which is dynamic and may be modified anytime during their interaction. Once the localization and scene registration have been accomplished, the Diminished Reality component undertakes the task of removing, in real-time, physical statues that will be replaced from their virtual counterparts from the user’s view, by substituting them with the appropriate background. The Virtual Agent Rendering component is responsible for delivering realistic representations of the statues that will be interactive in the XR environment, according to their matching 3D models. Eventually, the True Mediated Reality component positions the virtual representations of physical statues in the virtual environment, in a veritable manner, a task that requires realistic rendering and animations.

The “context-sensitive natural interaction with multiple users” involves components that are responsible for perceiving and interpreting users’ natural input commands, namely gestures and natural language. The Natural Language Processing component identifies the received commands employing the embedded NLP knowledge base. The Emotion Detection component is responsible for detecting user emotions, so that the system can be further adapted to the user. All identified gestures, speech, and emotions are fed to the “Context-sensitive interaction decision making” component, which is responsible for determining how the virtual statue will respond, taking also into account other parameters, such as the number of users who

actively interact with the virtual statue and of those who passively attend the ongoing interaction. According to the decisions made, the "Spoken Dialogue Generation" and "Multi-faced Embodied Response" components determine the feedback that will be provided by the virtual statue in the XR environment in terms of virtual agent posture, gestures and emotions, as well as information that will be delivered through spoken dialogue output.

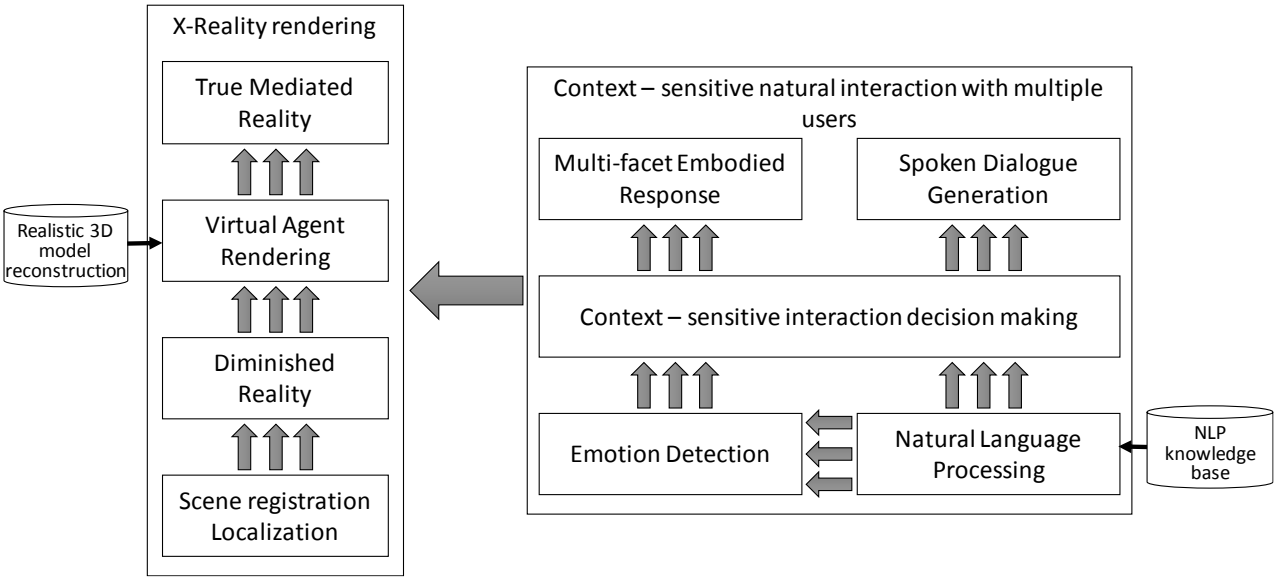

Figure 3. Conceptual architecture for realistic interaction with virtual statues in XR environments

The orchestration of the above components towards delivering a realistic user experience, is exemplified through the imaginary scenario of a museum visitor interacting with a virtual Caryatid[5] statue, that follows.

Physical world: Andrew is visiting the museum with his daughter, Sophia. Standing in front of a Caryatid statue, Andrew puts on his XR display device.

XR environment: The user's location and physical objects in their field of view are identified. The physical statue disappears and is replaced by a virtual Caryatid, placed in the appropriate position in the XR environment as it is rendered for each user. The virtual Caryatis welcomes Andrew and offers guidance: "Welcome visitor (nods), would you like to learn my story?"

Physical world: Andrew agrees "Yes, please" and asks Sophia to join him "Sophia put on your mask to hear the statue speak its story!"

XR environment: The system correctly interprets the first spoken dialogue as a command and ignores the second. As Sophia also wears her XR display device, the XR environment is appropriately rendered and she is asked whether she would like to join Andrew or initiate her own interaction. As soon as she joins, the XR experience is delivered taking into account multi-user interaction aspects. The Caryatis starts narrating its story, beginning from when it was first built to support the Erechtheion porch until the time she was transferred to the British Museum in London. Her narration is accompanied by multimedia to further engage her interlocutors and successfully visualize focal points of her narration (e.g. images, documentary videos).

[5] Caryatids are the famous ancient Greek statues acting as supporting columns of a porch in Erechtheion, an ancient Greek temple on the north side of the Acropolis of Athens, Greece; one of those original six figures, removed by Lord Elgin in the early 19th century, is now in the British Museum in London. The Acropolis Museum in Athens holds the other five figures, which are replaced onsite by replicas.

4. CONCLUSIONS

XR technologies hold the promise of delivering captivating experiences, allowing users to engage in natural interaction with technology as if they would with human counterparts. Advancements in the fields of realistic 3D modelling, rendering and animation, diminished and mediated reality can foster interaction in a seamless manner across the physical and virtual space. Employing such technologies in the Cultural Heritage domain, could produce extraordinary experiences promoting visitor satisfaction and enhancing knowledge acquisition.

This paper has presented a conceptual architecture for blending virtual and physical worlds in a single unified world, where for example statues become alive and narrate their story, or guide visitors to the entire museum. To this end, the necessary technological components have been presented in terms of current state of the art and challenges.

Current work includes the implementation of the presented architecture, while future endeavors will focus on the actual deployment of the proposed concept in cultural heritage sites and the assessment of the user experience entailed.

REFERENCES

Adams Jr, R.B. and Kleck, R.E., 2005. Effects of direct and averted gaze on the perception of facially communicated emotion. Emotion, 5(1), p.3. doi.org/ 10.1037/1528-3542.5.1.3

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J. and Black, M.J., 2016, October. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In European Conference on Computer Vision (pp. 561-578). Springer, Cham. doi.org/10.1007/978-3-319-46454-1_34

Coulson, M., 2004. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. Journal of nonverbal behavior, 28(2), pp.117-139. doi.org/10.1023/B:JONB.000

Egges, A., Papagiannakis, G. and Magnenat-Thalmann, N., 2007. Presence and interaction in mixed reality environments. The Visual Computer, 23(5), pp.317-333. doi.org/10.1007/s00371-007-0113-z

Emiliani, P.L. and Stephanidis, C., 2005. Universal access to ambient intelligence environments: opportunities and challenges for people with disabilities. IBM Systems Journal, 44(3), pp.605-619. doi.org/10.1147/sj.443.0605

Enomoto, A. and Saito, H., 2007, November. Diminished reality using multiple handheld cameras. In Proc. ACCV (Vol. 7, pp. 130-135).

Fast-Berglund, Å., Gong, L. and Li, D., 2018. Testing and validating Extended Reality (xR) technologies in manufacturing. Procedia Manufacturing, 25, pp.31-38. doi.org/10.1016/j.promfg.2018.06.054

Fiala, M., 2005. ARTag, a fiducial marker system using digital techniques. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 2, pp. 590-596). IEEE. doi.org/ 10.1109/CVPR.2005.74

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M. and Zettlemoyer, L., 2018. AllenNLP: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640.

Herling, J. and Broll, W., 2010. Advanced self-contained object removal for realizing real-time diminished reality in unconstrained environments. In 2010 IEEE International Symposium on Mixed and Augmented Reality (pp. 207-212). IEEE. doi.org/10.1109/ISMAR.2010.5643572

Ibanez, J., Aylett, R. and Ruiz-Rodarte, R., 2003. Storytelling in virtual environments from a virtual guide perspective. Virtual Reality, 7(1), pp.30-42. doi.org/10.1007/s10055-003-0112-y

Ioannides, M., Magnenat-Thalmann, N. and Papagiannakis, G. eds., 2017. Mixed Reality and Gamification for Cultural Heritage (pp. 161-199). Switzerland: Springer. doi.org/10.1007/978-3-319-49607-8

Ioannides, M. and Davies, R., 2018. Towards a Holistic Documentation and Wider Use of Digital Cultural Heritage. In Research Conference on Metadata and Semantics Research (pp. 76-88). Springer, Cham. doi.org/10.1007/978-3-030-14401-2_7

Kasap, Z. and Magnenat-Thalmann, N., 2007. Intelligent virtual humans with autonomy and personality: State-of-the-art. Intelligent Decision Technologies, 1(1-2), pp.3-15. doi.org/10.1007/978-3-540-79868-2_2

Kateros, S., Georgiou, S., Papaefthymiou, M., Papagiannakis, G. and Tsioumas, M., 2015. A comparison of gamified, immersive VR curation methods for enhanced presence and human-computer interaction in digital humanities. International Journal of Heritage in the digital Era, 4(2), pp.221-233. doi.org/10.1260/2047-4970.4.2.221

Kawai, N., Sato, T. and Yokoya, N., 2013, October. Diminished reality considering background structures. In 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 259-260). IEEE. doi.org/10.1109/ISMAR.2013.6671794

Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R. and Vilhjálmsson, H., 2006, August. Towards a common framework for multimodal generation: The behavior markup language. In International workshop on intelligent virtual agents (pp. 205-217). Springer, Berlin, Heidelberg. doi.org/10.1007/11821830_17

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R. & Socher, R. (2016, June). Ask me anything: Dynamic memory networks for natural language processing. In International Conference on Machine Learning (pp. 1378-1387).

Lepetit, V. and Berger, M.O., 2000. A semi-automatic method for resolving occlusion in augmented reality. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662) (Vol. 2, pp. 225-230). IEEE. doi.org/10.1109/CVPR.2000.854794

Li, Z., Wang, Y., Guo, J., Cheong, L.F. and Zhou, S.Z., 2013, October. Diminished reality using appearance and 3D geometry of internet photo collections. In 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 11-19). IEEE. doi.org/10.1109/ISMAR.2013.6671759

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. and Black, M.J., 2015. SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG), 34(6), p.248. doi.org/10.1145/2816795.2818013

Magnenat-Thalmann, N., Foni, A.E., Papagiannakis, G. and Cadi-Yazli, N., 2007. Real Time Animation and Illumination in Ancient Roman Sites. IJVR, 6(1), pp.11-24. doi.org/10.1145/1174429.1174432

Mann, S. and Fung, J., 2001. Videoorbits on eye tap devices for deliberately diminished reality or altering the visual perception of rigid planar patches of a real world scene. EYE, 3, p.P3.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60). doi.org/10.3115/v1/P14-5010

McTear, M., Callejas, Z. and Griol, D., 2016. Conversational interfaces: Past and present. In The Conversational Interface (pp. 51-72). Springer, Cham. doi.org/10.1007/978-3-319-32967-3_4

Meerits, S. and Saito, H., 2015, September. Real-time diminished reality for dynamic scenes. In 2015 IEEE International Symposium on Mixed and Augmented Reality

Workshops (pp. 53-59). IEEE. doi.org/10.1109/ISMARW.2015.19

Papaefthymiou, M., Feng, A., Shapiro, A. and Papagiannakis, G., 2015, November. A fast and robust pipeline for populating mobile AR scenes with gamified virtual characters. In SIGGRAPH Asia 2015 Mobile Graphics and Interactive Applications (p. 22). ACM. doi.org/10.1145/2818427.2818463
Papaefthymiou, M., Hildenbrand, D. and Papagiannakis, G., 2016. An inclusive conformal geometric algebra GPU animation interpolation and deformation algorithm. The Visual Computer, 32(6-8), pp.751-759. doi.org/10.1007/s00371-016-1270-8

Papaefthymiou, M., Kanakis, M.E., Geronikolakis, E., Nochos, A., Zikas, P. and Papagiannakis, G., 2018. Rapid Reconstruction and Simulation of Real Characters in Mixed Reality Environments. In Digital Cultural Heritage (pp. 267-276). Springer, Cham. doi.org/10.1007/978-3-319-75826-8_22

Papagiannakis, G., Elissavet, G., Trahanias, P. and Tsioumas, M., 2014. Mixed-reality geometric algebra animation methods for gamified intangible heritage. International Journal of Heritage in the Digital Era, 3(4), pp.683-699. doi.org/10.1260/2047-4970.3.4.683

Papanikolaou, P. and Papagiannakis, G., 2015. Real-time separable subsurface scattering for animated virtual characters. In GPU Computing and Applications (pp. 53-67). Springer, Singapore. doi.org/10.1007/978-981-287-134-3_4

Sanchez-Vives, M.V. and Slater, M., 2005. From presence to consciousness through virtual reality. Nature Reviews Neuroscience, 6(4), p.332. doi.org/10.1038/nrn1651

Sandor, C., Fuchs, M., Cassinelli, A., Li, H., Newcombe, R., Yamamoto, G. and Feiner, S., 2015. Breaking the barriers to true augmented reality. arXiv preprint arXiv:1512.05471.

Schröder, M., Baggia, P., Burkhardt, F., Pelachaud, C., Peter, C. and Zovato, E., 2011, October. EmotionML–an upcoming standard for representing emotions and related states. In International Conference on Affective Computing and Intelligent Interaction (pp. 316-325). Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-24600-5_35

Siltanen, S., 2006. Texture generation over the marker area. In 2006 IEEE/ACM International Symposium on Mixed and Augmented Reality (pp. 253-254). IEEE. doi.org/10.1109/ISMAR.2006.297831

Stephanidis, C., 2012. Human factors in ambient intelligence environments. Handbook of Human Factors and Ergonomics, pp.1354-1373. doi.org/10.1002/9781118131350.ch49

Sutskever, I., Vinyals, O. and Le, Q.V., 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112). arXiv:1409.3215

Vacchetti, L., Lepetit, V., Ponder, M., Papagiannakis, G., Fua, P., Thalmann, D. and Thalmann, N.M., 2004. A Stable Real-time AR Framework for training and planning in Industrial Environments. In Virtual and Augmented Reality Applications in Manufacturing (pp. 129-145). Springer, London. doi.org/10.1007/978-1-4471-3873-0_8

Valli, A., 2008. The design of natural interaction. Multimedia Tools and Applications, 38(3), pp.295-305. doi.org/10.1007/s11042-007-0190-z

Wareham, R., Cameron, J. and Lasenby, J., 2004. Applications of conformal geometric algebra in computer vision and graphics. In Computer algebra and geometric algebra with applications (pp. 329-349). Springer, Berlin, Heidelberg. doi.org/10.1007/11499251_24

Zhou, F., Duh, H.B.L. and Billinghurst, M., 2008, September. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality (pp. 193-202). IEEE Computer Society. http://doi.org/10.1109/ISMAR.2008.4637362

Zokai, S., Esteve, J., Genc, Y. and Navab, N., 2003, October. Multiview paraperspective projection model for diminished reality. In Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality (p. 217). IEEE Computer Society. doi.org/10.1109/ISMAR.2003.1240705