Interact: A Mixed Reality Virtual Survivor for Holocaust Testimonies

Minhua Ma
University of Huddersfield
Queensgage, Huddersfield, UK
m.ma@hud.ac.uk

Sarah Coward
National Holocaust Centre & Museum
Acre Edge Road, Newark, UK
sarah.coward@nationalholocaustcentre.net

Chris Walker
Bright White Ltd
Swinegate, York, UK
chris@brightwhiteltd.co.uk

ABSTRACT
In this paper we present *Interact*—a mixed reality virtual survivor for Holocaust education. It was created to preserve the powerful and engaging experience of listening to, and interacting with, Holocaust survivors, allowing future generations of audience access to their unique stories. *Interact* demonstrates how advanced filming techniques, 3D graphics and natural language processing can be integrated and applied to specially-recorded testimonies to enable users to ask questions and receive answers from that virtualised individuals. This provides a new and rich interactive narratives of remembrance to engage with primary testimony. We discuss the design and development of *Interact*, and argue that this new form of mixed reality is promising media to overcome the uncanny valley.

Author Keywords
Mixed reality, virtual human, natural language processing, question-answering, holocaust survivor.

ACM Classification Keywords
H.5.1. Information interfaces and presentation (e.g., HCI): Multimedia information systems; H.3.3. Information storage and retrieval: Information search and retrieval.

INTRODUCTION

A key part of educational experience on the Holocaust topic is listening to, and interacting with, a Holocaust survivor. In some memorial centres, Holocaust survivors speak to audience, sharing their story and answering the questions that individuals care about. Listening to and meeting a Holocaust survivor in person provides an opportunity for people to attend to a person's full story, from which they can gain deeper insights, rather than *snippets*, and builds empathy and understanding between the audience and the survivor, from which they can develop their views as to the Holocaust and genocide.

There are approximately 800 Holocaust survivors remaining in the UK, with even fewer actively sharing their story. Each year survivors pass away, or become too frail to deliver their testimony in person. There is an urgent need to capture their experiences.

CONVERSATIONAL NATURAL LANGUAGE INTERFACES

Conversational agents and natural language interfaces have been used to improve the communication between human and computers such as information retrieval systems. They can be text-based or in the form of embodied agents.

An Embodied Conversational Agent (ECA) is a computer-generated virtual avatar that has a 2D or 3D representation and human-like behaviour while interacting with the user. Besides the back end of an ECA, i.e. a text-based conversational program, an ECA may involve visual/audio input and output components such as speech synthesis (output), voice recognition (input), animation for conversational behaviours such as gestures and facial expressions (output), and face/expression recognition (input). To date, ECAs have been widely used for various purposes: museum and tour guides (Swartout et al., 2010), enhancing consumer experience in e-commerce, and computer assisted learning etc.; across many platforms: web-based, smart phones, and online virtual environments.

QUESTION ANSWERING ABOUT THE HOLOCAUST

Holocaust is a rare application domain for closed-domain question answering in NLP: apart from Filatova (2008) & Psutka et al. (2010), there are very few NLP applications dealing with questions about the Holocaust. Most of these QA systems are text-based. Only one ongoing project (Artstein et al. 2014) allowing multimodal conversation based on video testimony and spoken question answering, at a high production cost.

Previous Holocaust archives consist of written records and spontaneous speech from oral history interviews, e.g. the Malach corpus (Byrne et al., 2004) is a large archive of about 8,000 segments from interviews of Holocaust survivors, liberators, rescuers and witnesses. Question-answering system based on these archives are limited in term of narrative immersion and user interaction.

DESIGN AND DEVELOPING *INTERACTION*

At the outset we established solid design principles, which informed the process and approaches throughout the project. These were (1) to recreate, preserve and replicate today's experience in the National Holocaust Centre (NHC). (2) authenticity: to recreate the survivor's presence using non-interventionialist documentary

techniques, and this is desirable in order to make the entire project more meaningful as a historical document.

Mapping Current Interaction

The Holocaust is a pre-defined domain with words, phrases, people, places, ideas and testimonies that has been widely referenced. Each survivor overlays new areas of domain specific to their life experience, often in finer resolution than the general topic domain. For example hometowns, siblings, birthday gifts, family events. In our case, a survivor talks about a decade of his life in enough detail to carry their message within usually one hour.

The current proceedings between museum visitors and survivors at the NHC happen as described in Figure 1. Three parties: the facilitator, the survivor and the audience, are involved. The dark blue elements denote active engagement (talking); the light elements denote passive engagement (listening). The passive survivor engagements (light blue elements) are of indeterminate length, and require special measures to replicate. We use photorealistic 3D virtual human to replicate these stages.

Figure 1. Interaction of Holocaust testimony and QA.

The *Interact* System

Figure 2 shows how a question is processed and answered by the virtual survivor. The audience question is scanned in realtime for recognised exchangeable terms; the same dictionary used to standardise pre-recorded questions is used to standardise the live audience queries. The information retrieval component uses a statistical relevance model to match the question to one of the Q-A pairs recorded with the survivor. If a selected answer (identified by a unique asset ID) passes the customer defined threshold, the audio-visual assets associated with the ID is played back to the audience.

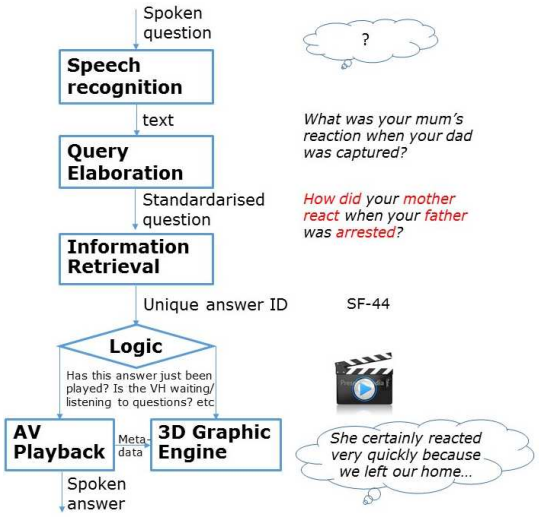

Figure 2. The flow chart of *Interact*.

Regarding the technological development of the system elements, some components were developed based on third party software, e.g. Nuance technology for speech recognition and NPCEditor (Leuski and Traum, 2011) for information retrieval.

Question Generation Methodology

We define questions and answers as a pair. Semantic variants of the question are ignored during pre-production. Variants will be introduced later in the process but, when generating questions, we are looking for unique question-answer pairs, rather than different phrasings of the same question. For example: *Have you ever experienced survivor guilt?* and *Have you ever felt guilty for surviving when so many others perished?* are the same question, count as one question, and was therefore asked once. However, *Have you forgiven the perpetrators?* and *Have you forgiven those involved?* are different questions, since the survivor may treat the perpetrators and those who did nothing or stood by as events unfolded differently.

We established two categories of question that can be posed: (1) questions that are specific to the survivor and his/her testimony, e.g. places, times, people, objects and events laid forth during the testimony. It would not be possible to ask this type of questions without having experienced the talk; (2) subjective questions. The audiences wishes to know what view, opinion, interpretation or emotion the survivor attaches to any aspect of the domain, whether that be the domain defined during testimony, or common-knowledge domains.

We use a lifeline chart (Figure 3) to develop testimony specific questions. This allows a group of people to navigate and visually view a life story. Its principle aim is to facilitate and enable question generation through group working. The Holocaust lifeline works on two common and basic principles, that survivors got older, and were displaced (they were moved around by the Nazis). These two variables, age and displacement represent to two axes of the lifeline graph. Starting at the bottom left, the survivor was born in their hometown. As they grow older, they are displaced through various camps. Some survivors have extremely complex lifelines, others are relatively straightforward.

We believe that our lifeline graph projected on time and displacement coordinate system is applicable not only to the Holocaust domain but also in wider narrative to define the *Hero's journey* for documentary practices in art and exhibitions.

Testimony-specific questions were generated at all-day meetings with that sole purpose. The best question sets arise when many different perspectives are brought to the table, always remembering that the profile of the question generation group should always be matched to the profile of the audience. Our sessions typically involved 8 to 10 people for each question generation session. At the time of writing, 6 survivors' testimony have been processed in this way, and the team generates approximately 550 subjective questions and 500 testimony-specific questions per survivor.

The question processing stage removes duplicate questions and stop words while not breaking up a grammatical sentence, and standardises each question making it as succinct as possible and following a high standard of grammar.

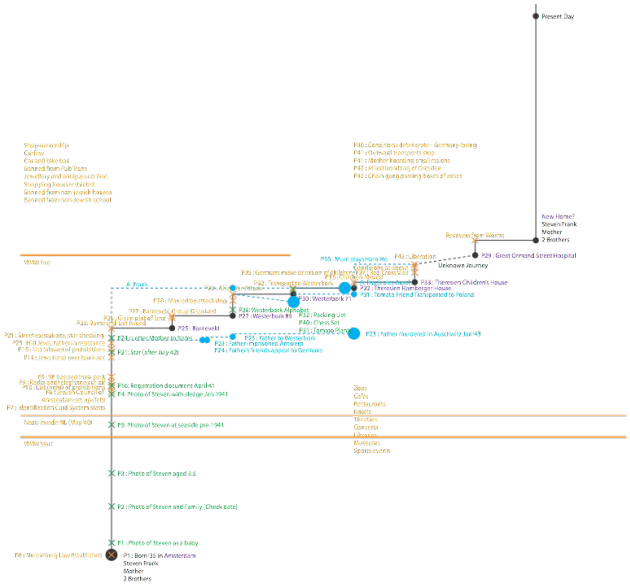

Figure 3. The lifeline chart of a Holocaust survivor.

Filming and 3D Data Capture

Survivors were filmed over a five-day period each at the studio. We trained the survivor to start and end each answer by looking straight into the camera, but to address the whole audience (our standby staff carefully placed around the studio) whilst they were giving the testimony and answers.

Figure 4. Holocaust survivor giving his testimony.

We use stereo pair cameras and a facial close-up camera for video recording of testimony and answers, and also photographic and facial scanning of the survivor for generating a 3D model of virtual human.

Creating Virtual Survivors

In the interaction chart (Figure 1), the active engagements (dark blue) of the survivor are linear pre-recorded sequences; the passive engagements (light blue) are of indeterminate length and require CGI to replicate conversational behaviours like nodding, head tilting, gaze and other idle motion. To maintain the flow of the session, *Interact* virtualises the survivor during the passive engagements, i.e. we switch to a virtual 3D model of the survivor whilst he is not speaking.

The survivor's bodily pose at the beginning and end of each answer was recorded in meta-data associated with the answer. Once an answer has been selected for immediate display, the runtime application reads these poses and in realtime configures the virtual survivor into those poses, cross-fading into the virtual survivor in-between answers. The virtual survivor continues to move naturally, based on a series of collected body language signatures. This means that neither the real nor virtual survivor has to return to a control position, they are free to move naturally.

The appearance of the virtual survivor is photorealistic, but the main front studio light is switched off so the survivor is slightly silhouetted. It acts as if the focus light has moved away from him/her.

A key output of the virtualisation is that a fully-detailed posable 3D model of the survivor is created. This will be of use to teams in the future looking to upgrade the experience for unforeseeable future display technologies.

The virtual survivor was created using a 3D laser scan as the basis, then a 3D modeller develops the model, using a large number of photographic reference images taken whilst the survivor is in the studio. It was important that time was booked in to create this reference, and that the survivor did not change their clothes during the week-long filming sessions.

The Uncanny Valley and a New Form of Mixed Reality

A number of factors play important roles for user satisfaction when interacting with embodied conversational agents. These include personality, believability of non-verbal behaviours (e.g. facial expressions, lip synchronisation, gestures, body postures, gaze) and emotions, visual fidelity in terms of the appearance of virtual human and the naturalness of their motion, and audio fidelity of synthesized voice (e.g. prosodic features of the utterance such as intonation, pauses, accent, and stress).

Computer Generated (CG) virtual humans face another challenge, the uncanny valley (Brenton et al., 2005), on appearance and movement of the animated agent. Since *Interact* is a mixed reality virtual human based on pre-recorded video testimony and 3D character generated from 3D scanning of real human, most of the above challenges can be avoided, if the transition between video recordings and photo-realistic virtual human is seamless. The focus lighting approach is effective as it not only *hides* noticeable flaws of the CG character but also appears natural, i.e. when the survivor is not talking the lights are dimmed.

Mixed reality, a.k.a. augmented reality, is defined as a live view of a physical, real-world environment whose elements are augmented by CG input. It usually overlays

virtual components on real world environment, creating an *augmented* reality scene (Milgram and Kishino, 1994). As a result, the technology functions by enhancing one's current perception of reality.

We differentiate three forms of *mixed reality*, as illustrated in Figure 5. The first is the most common form of augmented reality, where CG elements are overlaid on the real world environment. The second form, which we call 'time-based augmented reality', has multiple points in time overlaid onto the physical world environment. It often provides information about multiple points in time for a single object and has become popular in the construction industry for construction site monitoring and documentation.

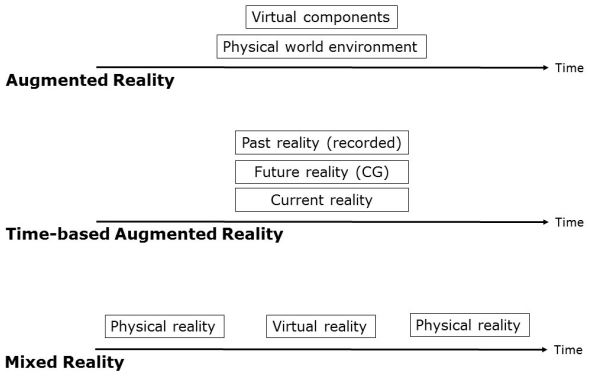

Figure 5. Three forms of augmented and mixed reality.

The third form is what we defined as 'mixed reality', where instead of augmenting physical reality with virtual elements or past reality, it mixed physical reality and virtual reality in different point of time and seamlessly transition between them. The components in the virtual reality replicate those in the physical reality using photorealistic rendering of automatically generated 3D models from laser scanning and photogrammetry data. The *Interact* project belongs to this category. We believe that combining blending techniques and focus lighting the mixed reality could achieve the highest visual fidelity and it is the most promising media to overcome the uncanny valley.

Query Elaboration and Expansion

User questions are processed at the lexical, syntactic, and semantic levels. Discourse level analysis has not been considered due to the one-to-many conversation.

The question expansion process developed accepts variant forms of each question and replaces them with its primary form based on a set of rules, such as stop words removal, re-organise reverse questions. The semantic model of question understanding and processing would recognize equivalent questions, regardless of how they are presented. A semantic ontology for the Holocaust domain were created in the query expansion process. The ontology was built offline using pre-established rules to extract specialised semantic knowledge. Each entry consists a primary term and a number of secondary terms (exchangeable terms).

When generating the ontology, we considered: 1) English word frequency list based on the British National Corpus for conversational and task-oriented speech; 2) semantic relations for different parts of speech (examples in Table 1 are taken from transcripts of a survivor's testimony and answers) based on WordNet synsets (Fellbaum, 1998); and 3) Holocaust domain specific terms such as interchangeable place names or names in other languages, e.g. *Theresienstadt/ Theresien/ Terezin*.

POS	Relations	Examples
Noun	Hypernyms – hyponyms	flower-daffodil; clothes-shoes, coat; food-bread, porridge, potato; building-barrack, house
	Meronym – holonym	foot-toe, sole; building-roof, attic
	Instance	Auschwitz-concentration camp
Verb	Troponym	run-scarper, flee, escape
	Entailment	beat-hit
	Derivationally related form	remember-memory, recall, remembrance, recollection; hate-hatred, hostile, dislike; murder-kill, slay, execute, death
	Hypernym	emotion-hate, love
Adj	Hyponym	fear-scare, panic, dread, afraid
	Synonym	downtrodden-oppressed, crushed, persecuted

Table 1. Semantic relations of Holocaust related words.

In the ontology, the primary word is a selected keyword or phrase in British English language. They make for very rigid forms of speech and carry the meaning of all the secondary forms, which is rich in slang, common speech, dialects, and regional uses for words and phrases.

If a different territory showed an interest in hosting our virtual survivor: any regional features of popular speech, spellings, words and phrases can be represented in the ontology as secondary terms. Similarly, over decades, English language evolves, the ontology could be updated, to reflect shifts in the language.

EVALUATION AND CONCLUSION

Experiments have been carried out to evaluate relevance of answers and user satisfaction. Initial results showed a subjective rating of 4.2 for average user satisfaction and 4.08 for average quality of answers on a 5-level Likert scale. Details of the evaluation and objective measures of precision and recall will be reported at a later stage.

We discovered that the system was capable of dealing unexpected questions. Due to the asymmetry of the Q-A data set, the answer data includes more information than required by the questions. When subjects asking about the professions of parents after the war and the favourite food of the survivor; although we didn't ask these questions in our filming sessions, the answers were present inside the answer to another question, and were successfully retrieved.

Apart from applications within museum settings, *Interact* provides substantial opportunities for the wider arts sector to create conversations between a pre-recorded photorealistic virtual human and audience.

ACKNOWLEDGMENTS

The *Interact* project was supported by the Digital R&D Fund for the Arts, which is jointly funded by NESTA, Arts and Humanities Research Council and public funding by the National Lottery through Arts Council England.

REFERENCES

Artstein, R., Traum, D., Alexander, O., Leuski, A., Jones, A., Georgila, K., Debevec, P., Swartout, W., Maio, H. and Smith, S. Time-offset interaction with a holocaust survivor. In Proceedings of the 19th International Conference on Intelligent User Interfaces, ACM, New York, USA (2014), 163–168.

Brenton, H., Cillies, M., Ballin, D. and Chatting, D. The Uncanny Valley: does it exist? In the 11th International Conference on Human-Computer Interaction. Lawrence Erlbaum Associates, Las Vegas (2005).

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., et al. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Speech and Audio Processing (2004), 12(4): 420–435.

Fellbaum, C. WordNet. Blackwell Publishing Ltd. (1998).

Leuski, A. and Traum, D. NPCEditor: Creating virtual human dialogue using information retrieval techniques. AI Magazine (2011), 32(2):42–56.

Milgram, P. and Kishino, A. F. Taxonomy of Mixed Reality Visual Displays. IEICE Transactions on Information and Systems (1994), 1321–1329.

Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., Williams, J., Leuski, A., Narayanan, S., Piepol, D. Ada and Grace: Toward realistic and engaging virtual museum guides. 10th International Conference on Intelligent Virtual Agents (2010), 286–300.