Question-Answering Virtual Humans based on Pre-recorded Testimonies for Holocaust Education

Minhua Ma

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

Sarah Coward

National Holocaust Centre & Museum, UK.
sarah.coward@nationalholocaustcentre.net

Chris Walker

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

Abstract In this chapter we present Interact—a project which builds question-answering virtual humans based on pre-recorded video testimonies 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 virtualised individuals. This provides a new and rich interactive narrative of remembrance to engage with primary testimony. We briefly reviewed the literature of conversational natural language interfaces, discussed the design and development of *Interact*, including how we mapped the current proceedings of testimony and question answering session to human computer interaction, how we generated/predicted questions for each survivor using a lifeline chart, the 3D data capture process, generating 3D human, natural language processing, and argue that this new form of mixed reality is promising media to overcome the uncanny valley. Subjective and objective evaluation is also reported. The chapter is a longer version of a short paper presented at the ACM OzCHI conference (Ma, Coward and Walker, 2015).

Keywords Mixed reality, virtual human, natural language processing, question-answering, holocaust survivor, pre-recorded video testimony

1. Introduction

A key part of Holocaust education is listening to and interacting with a Holocaust survivor. In some museums and education centres such as the National Holocaust Centre and Museum in the UK (NHC), Holocaust survivors speak to audience, sharing their story and answering questions about their experience. 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 listening to *snippets*. This builds empathy between the audience and the survivor, as the audience develop their knowledge and understanding of the Holocaust and genocide.

Listening to a Holocaust survivor's personal experience and interacting with them is a key part of Holocaust education. However, soon this experience will be lost. There are few Holocaust survivors remaining in the UK who are able and willing to share their story publically in person. Each year survivors pass away, or become too frail to deliver their testimony in person. There is an urgent need to capture their experiences.

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.

We aim to create a rich experience which would replicate, as far as possible, the existing experience for visitors, by developing a virtual Holocaust survivor, who could effectively respond to questions in the closed domain of the Holocaust. The basis for the work was informed by research into conversational natural language interfaces.

2. Conversational Natural Language Interfaces

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

Text-based conversational agents are the earliest form of chat bots. In a closed domain conversation, they sometimes fool users into believing that it is a real human through written conversation and applications of conversational programs vary from online help (interactive question answering), accessing an information system, to personalised services. The main areas involved to build conversational programs include Natural Language Processing (NLP), dialogue management, knowledge representation (specialised and common sense knowledge), information retrieval, and reasoning. Conversations with chat bots are virtually unlimited unless topics or

tenors are restricted (closed-domain). In the early years of Turing tests, it has been decided to add rules to limit the topic to a closed domain in order to give computers a chance. The most common method in closed-domain conversations is searching algorithms based on question-pattern and answer pairs in a repository of questions and answers. More recently, closed domain conversational systems started to integrate image processing techniques to utilise multimedia data. For example, the COMPANIONS project (Wilks et al., 2011) resulted in a senior virtual companion who can engage the user in reminiscing conversations about their photographs using face recognition and information extraction techniques.

Speech-based chat bots are based on the Automatic Speech Recognition (ASR), speech synthesis technology and test-based question answering systems. Typical applications are in searching and personal assistant services such as Apple's Siri, Google Now, Microsoft Cortana and Amazon's Echo, which are embedded in smartphones, computers and game consoles. However, most of them only conglomerate data available on the internet and lack sophisticated AI.

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: clinical psychology training (Talbot et al., 2012), museum and tour guides (Swartout et al., 2010), job interview skills training and coaching (Hoque et al., 2013), enhancing consumer experience in e-commerce (Delecroix, Morge and Routier, 2012), and computer assisted learning etc.; across many platforms: web-based, smart phones, and online virtual environments such as Second Life.

2.1. Question Answering (QA) about the Holocaust

Holocaust is a rare application domain for closed-domain question answering in Natural Language Processing (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 information retrieval system though the corpus may be text, speech or videos in single language or cross-lingual. There is only one ongoing project (Artstein et al. 2014) allowing multimodal conversation based on video testimonies and spoken question answering, for which the production costs are high.

Previous Holocaust question answering applications have been based on spontaneous speech from oral history interviews. For example, the Malach corpus contains 8,000 segments from 300 interviews of Holocaust survivors, liberators, rescuers and witnesses. Each segment contains ASR outputs from IBM ASR

systems with a 40% word error rate and automatically generated thesaurus terms and a set of human generated data, including person names mentioned in the segment, thesaurus labels and 3-sentence summaries.

3. Design and development of *Interact*

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.

3.1 Mapping Current Interaction

The Holocaust is a pre-defined closed domain with words, phrases, people, places, ideas and testimonies that has been widely referenced, and the audience also brings a degree of knowledge of the domain with them. This domain is not static merely because the events happened in the past; new interpretations and discoveries happen all the time.

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 duration of testimony and answers are roughly equal. That is to say that a fuller testimony (better-defined and organised) will result in fewer questions due to fewer loose ends being left, and that a scant testimony will leave many questions. When considering the application of this research and development to other programmes, the talk length needs to be carefully considered.

We believe that there is a penalty to the overall sum duration (it will increase) with short testimonies since there is a higher chance of exploratory questions, and those eventualities need to be provided-for. In other words, the framework of the story is unclear and will be discovered by questions.

We worked on the principle that regardless of the fullness of the testimony, the audience will have questions for the survivor that either related to facts in his narrative or about his opinions, interpretations and emotions. Mathematically,

```
Duration of captured media = TM + NA +SA
```

where `TM` is the length of testimony; `NA` is answers relating to the narrative; and `SA` is subjective answers.

```
NA = (1/TM) * Discovery Factor
```

When the narrative is badly defined due to a short testimony (TM), some of the questioning (NA) by the audience is spent on discovering the general facts of the story rather than probing more deeply into the survivor's experience and emotions.

Our decision on length of testimony was made to allow the survivors to talk for their natural duration, which is usually one hour, and in some cases 40 minutes. Organisations seeking to consider applicability of this technology to other domains should be aware that these observations are only true for the Holocaust domain, which is large, representing huge swathes of human experience, and therefore our talk length and number of questions is correspondingly large. Smaller domains, such as an artist talking about a specific work, or an architect talking about a specific building, are likely to imply shorter talk lengths and fewer questions.

The current proceedings between museum visitors and survivors at the NHC happen as described in Fig 1. The format is a typical talk-plus-QA session. Three parties: the facilitator, the survivor and the audience, are involved. Without the facilitator, the interaction does not work well. As host, the facilitator introduces the survivor, defines the periods when the audience should be listening, and encourages the audience to interact. They also help to ensure fairness in giving as many of the audience as possible the chance to ask questions. We were only concerned with re-creating the active and passive engagement by the survivor, as the facilitator and the audience are real and present people.

The dark blue elements denote active engagement (talking); the light elements denote passive engagement (listening). In the cases of the facilitator and the audience, the passive and active engagement are *live* processes (they are living people) unlike in the case of the survivor, the active engagements can be replaced with linear pre-recorded sequences. The passive survivor engagements (light blue elements) are of indeterminate length, and require special measures to replicate. We use a photorealistic 3D virtual human to replicate these stages.

Fig.1. Interaction of Holocaust testimony and QA

3.2 The Interact System

Fig. 2 shows how a question is processed and answered by the virtual survivor. The audience question is scanned in real-time 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.

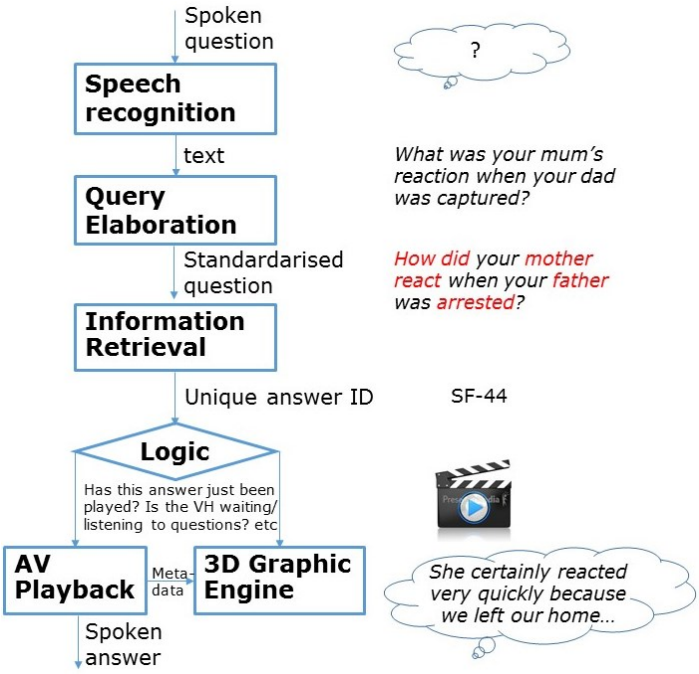

Fig. 2. The flow chart of *Interact*

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.

3.3. Question Generation Methodology

The NPCEditor requires us to 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. They are regarded as two distinct questions and both were asked.

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.

Below is the procedures created by the team to develop testimony specific questions.

1. Survivor testimony was recorded as a guide. The video was trimmed, compressed, and uploaded to the collaborative secure document system.
2. The testimony was transcribed and marked up to identify people, objects, events, times, places and digressions.
3. A lifeline chart was drawn-up to include all mark-ups that could lead to a valid question.
4. A team was appointed, given the materials, and asked to methodically go through each entry on the lifeline, generating questions as they go. On the whole, the nature of these questions is an attempt to increase the definition or resolution of the domain.
5. The survivor was asked selected questions to complete the domain, and progress stories.
6. Further testimony-specific questions (2nd round) were generated based on 5.
7. Subjective questions were generated--questions that were related to feelings, opinions, and views. These questions in general could be asked to any survivor.
8. The questions were collated, processed, and approved by the team.
9. The questions were prioritised.

We use a lifeline chart (Fig. 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 geographically displaced (e.g. by being moved from camp to camp; by changing location to seek to avoid persecution; by going into hiding). 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.

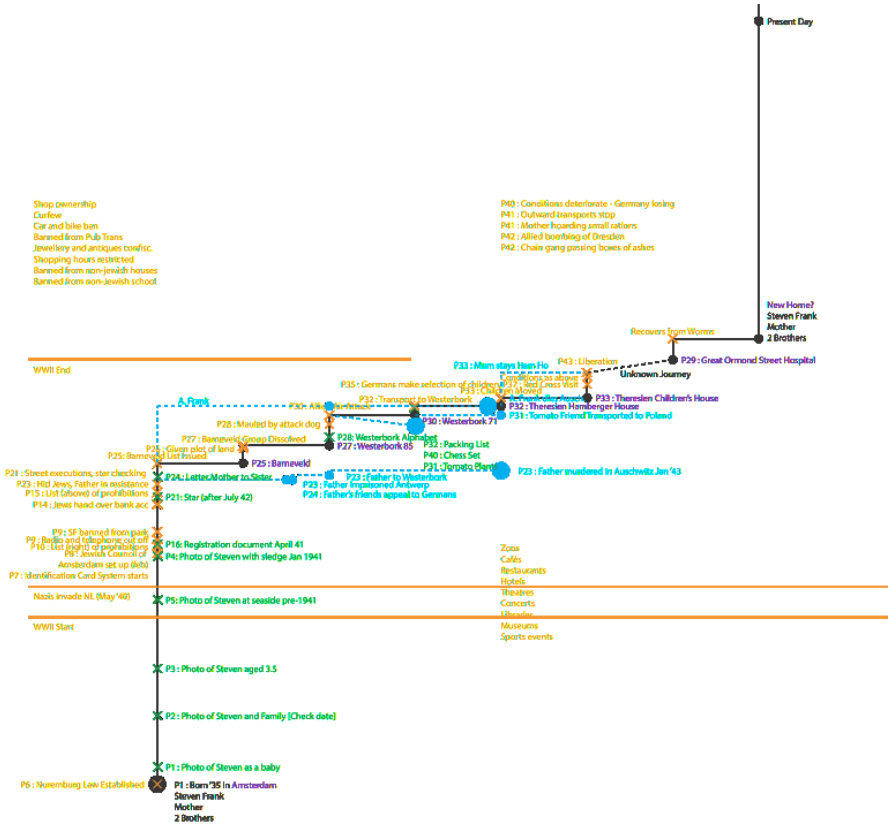

Fig. 3. The lifeline chart of a Holocaust survivor

Storytelling and narration have played a significant role in Holocaust education, contemporary art, materializing as a trend that has developed alongside the increasing popularity of documentary practices in art. Storytelling seems to be capturing everyone's attention as an ever-increasing number of exhibitions feature strongly narrative work. 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.

Subjective question generation took place in the same forum, but with a different approach. An analysis of the questions asked by schoolchildren to the survivor, supplied by the Holocaust Centre, has led to the identification of a map of topic prompts. Distinctly different from the logical discovery of the domain through

questions relating to the domain, the subjective questions seek to discover how the survivor in question felt at given times, how their faith was effected, how their interpretation or opinions may differ from the norm or the history books. Many of the questions are generic, but not all. Each mind set is adopted by the question generation group, as they aim to methodically predict as many subjective questions as possible.

At the time of writing, 10 survivors' testimonies 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.

3.4. Video Recording 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.

We use a stereo pair camera 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. Fig. 4 and 5 show survivors giving their testimonies and answers in a filming session. In terms of audio I/O, we use stereo overhead, Lapel mic, and a microphone for the questioner.

A key principle of *Interact* is authenticity; the collected data will not be processed in any way. No grading, colouration or editing will take place other than to normalise the image. The captured data remains a primary source historical document. We adjusted the lighting and colour of the renders of 3D virtual human in the post-production and testing phase to match those in pre-recorded videos.

Fig. 4. Holocaust survivor giving his testimony

Fig. 5. Video recording of a Holocaust survivor in the studio

3.5. Creating Virtual Survivors

In the interaction chart (Fig. 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 real-time 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 (Fig. 6), 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.

Fig. 6. Photorealistic 3D representation of a Holocaust survivor

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.

3.6. 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, Fig. 7), on appearance and movement of the animated agent. The uncanny is a feeling of uneasiness triggered by unreal or unnatural artefacts of an animated character. The theory was originally developed for evaluating the realism of humanoid robots, but has been extended to animated characters. Unnaturalness in appearance is easily to be spotted when an embodied agent is in motion. For example, Pandorabots' conversational agent Captain Kirk (Pandorabots.com), the user will soon discover the flaw on the texturing of his eyes and teeth when he is moving or talking. The problem is not obvious when Captain Kirk is still, but it immediately throws the users out of the flow of natural conversation once they noticed the nuance.

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.

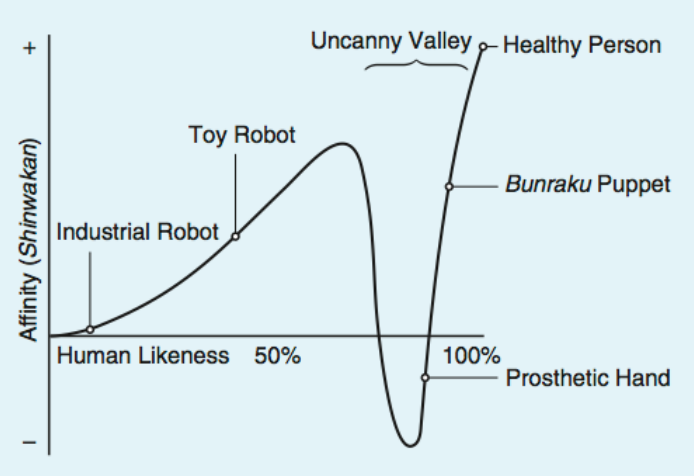

Fig. 7. Uncanny valley (Brenton et al., 2005)

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 Fig 8. 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.

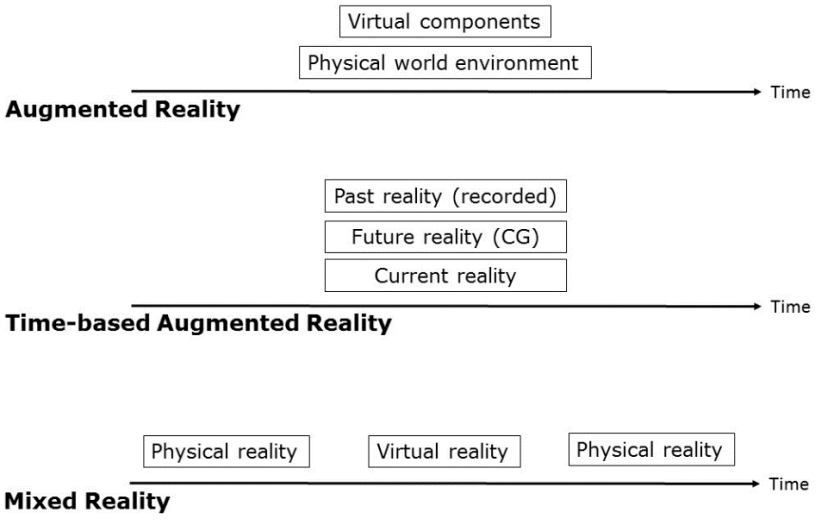

Fig. 8. Three forms of augmented and mixed reality

The third form is what we defined as *mixed reality*, where instead of augmenting physical reality with virtual elements or past reality, it mixes physical reality and virtual reality at different points in time and transitions 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.

3.7. 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. It was evident that different types of query required the use of different strategies to find the answer. Typical question classes we considered were:

- Polar questions that seek to find one of two answers, typically yes or no),
- Wh-questions (what, why etc., which implies a range of possible answers),
- Questions that request description (usually with an imperative form of command, e.g. *how would you describe your school?*)

Topics in each question class were investigated, for example, wh-questions include names, age, opinion, fate (factual and hypothetical), reaction (factual and hypothetical), awareness, comparison, slow realisation, and revelation, etc. The same information request can be expressed in various ways, some interrogative (*What are the names of x?*) and some assertive (*Tell me the name of x*).

The question expansion process developed accepts variant forms of each question and replaces them with its primary form based on a set of rules, including:

- Removing stop words in the question ("Did you *ever* go back to any of the camps?"), i.e. high frequency common words that have a low weight and contribute little to the relevance score, such as *any, ever, always, specific*.
- Using bare infinitive form of verbs in polar questions as the primary form ("Did your brother marry?" / "Did your brother get married?");
- Accepting assertive forms of the question as secondary forms ("What was your daily routine in [x]?" / "Tell me your daily routine in [x]", "Can you tell me your daily routine in [x]?")
- Accepting reverse questions, which does not have the Wh-word at its beginning and is equivalent to a question that does, as secondary forms. ("What year were you deported to concentration camp?" / "You were deported to concentration camp in what year?")

The semantic model of question understanding and processing would recognise equivalent questions, regardless of how they are presented. Due to the presence of a facilitator/compère (Fig. 1), a more complicated semantic model that would enable the translation of a complex question into a series of simpler questions, identify ambiguities and treat them in context or by interactive clarification, is not required in this context. It is important to therefore recognise the importance of the facilitator in supporting the interaction.

A lexicon for the Holocaust domain were created in the query expansion process. The lexicon 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 lexicon, 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 survivors: assuming that the principle language is English, any regional features of popular speech, spellings, words and phrases can be represented in the lexicon as secondary terms. Similarly, over decades, English language evolves, the lexicon could be updated to reflect shifts in the language.

3.8. Interact Hardware

Development and roll-out takes place on any desktop or laptop computer built within the last 3 years. The installation hardware is a custom-integrated system designed to project 4k stereoscopic images onto a stage, complete with audio and parallel projection channels to support pre-recorded PowerPoint presentations and facial close-ups.

The system requires an integration service, but all elements are standard other than the facilitator's microphone which is a bespoke construction that integrates a momentary pushbutton into the microphone, allowing the facilitator to indicate when a question is being asked.

It is important to remember that the hardware requirements for our application are very high; more affordable systems can be enabled, for example, 4k resolution could be replaced by 1080p or 720p; stereoscopic 3D videos could be replaced by traditional 2D videos; projection could be replaced by screen-based display. As an illustration of the scalability of the technology, a 2D screen-based 720p implementation would run on an ordinary desktop computer.

4. Evaluation

Interact has been successful in demonstrating that integrated technologies can be applied to help audiences engage with key individuals who have unique knowledge or experience: providing the opportunity for people to engage with a pre-recorded filmed individual and virtual human to explore their experience. Experiments have been carried out to evaluate relevance of answers and user satisfaction.

Initial testing on the dataset was performed using a body of questions authored by the United States Holocaust Memorial Museum (USHMM). The list contains sample questions for interviewing Holocaust survivors. The questions provided a framework for the kinds of question one may ask in an interview with a Holocaust survivor. The body of questions was useful to us because they are high quality and, more importantly, not written by any of the development team, therefore used a different style of language, important to our evaluation.

Our Q-A dataset is asymmetric: the questions are short in comparison to the answers. The set had an average word count of 8.56 for questions, and an average answer word count of 114.48. The goal of *Interact* is to retrieve best-matching passages rather than short answers to questions, which is the goal of most information retrieval or question answering systems currently do, e.g. the TREC Question Answering Track that has motivated most recent research in the field, focuses on fact-based, short-answer questions such as "*Who killed Abraham Lincoln?*"

This led us to the idea that the statistical analysis of questions and the statistical analysis of answers should be different. We tested whole-word level scrutiny of the answers, and sub-word (N-grams) scrutiny of the questions. The latter achieved strong results compared to symmetric or inverse-asymmetric scrutiny.

Sub-word scrutiny of questions exposed inconsistencies in our question data. For example, of the 913 questions, approximately 84 questions were of the same class, which was '*Will you describe…*'. In a small number of cases, when compiling the questions, we had slipped into using the form '*Can you describe…*'. During in-house testing, we found that using the latter form would improve NPCEditor's precision on retrieval. We have never edited words from the survivor's answer due to the requirement for authenticity, but we alter the stylistic form of the question as long the meaning is maintained. We then replace the question with its primary alternate form based on the rules, e.g. *Can you tell us about X?* / *What was X like?*

Testing also exposed that the speech recogniser employed to semi-automatically transcribe the answers altered certain words against its internal lookup table. For example the words 'identity card' was abbreviated to 'ID Card'. The abbreviation ID is not easily transposed to the word 'identity' and its derivatives and therefore important connections were lost. We ensure that any such changes coming from the speech recogniser were identified and either rectified by either modifying the speech recogniser's lookup table, or our own data set. A similar observation was made about the notation of dates and years (e.g. nineteen forty-eight vs 1948).

The Q-A matching was capable of pulling deep answers out. Due to the asymmetry of the Q-A data set, the answer data includes more answers than the number of questions we asked. For example, asking about the professions of parents after the war and the favourite food of the survivor. Although we didn't actually ask these questions in our video recording sessions, the answers were present inside the answer to another question, and were successfully retrieved.

In the subjective evaluation, test subjects watched the filmed survivor giving his testimony, and then ask any question they liked. They gave a subjective rating to each response of the virtual survivor's on user satisfaction and quality of answers. The 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.

Objective performance of precision, recall and quality of answers were measured on a relatively small testing dataset. The definitions of these evaluation measures are below.

1. **Precision**: the fraction of retrieved instances that are relevant, i.e. proportion of relevant answers among all returned answers.
2. **Recall**: the fraction of relevant instances that are retrieved. Non-response is considered here. It is calculated by the number of relevant answers returned, divided by the total relevant answers in the dataset. Since Interact only returns the best answer or no answer, the recall is calculated differently from conventional information retrieval system evaluation. For example, for 10 user questions, Interact returned 7 relevant answers, 1 irrelevant answer and 2 non-response. In the dataset, we are able to find relevant answers for the 2 questions which returned an irrelevant answer or no answer. The recall will be 7/9=77.8%
3. **Quality of answer** is measured by comparing *Interact* response with a real person's response. Of course, it might not be possible to compare virtual survivor's response with the real person whose answers were recorded, a team member who is very familiar with the survivor's story and scripts acted as the human evaluator. We compared the answers returned by *Interact* and the best answers given by the human evaluator using the existing answers in the dataset, and compare how close they are. In 10 questions, if 6 are same from *Interact* and from the human evaluator, then the human likeness or quality of answer is 60%. Those responses that are relevant to the question but not the best answer in the dataset were not counted.

Our testing dataset has 42 Q-A pairs. The system returned 31 relevant answers, 25 of them are the best answer in the dataset; 7 irrelevant answers, 5 of which has a better answer in the dataset; and 4 non-answers, 2 of which has a relevant answer

in the dataset. Therefore, the precision of Interact is 81.6%; the recall is 81.6%; and the quality of answer (human likeness) is 64.3%.

We are in the processing of collecting more data from the Q-A sessions at the National Holocaust Centre when the audience interact with the virtual survivors and plan to analyse the data on a much bigger test collection.

5. Conclusions and Future Work

We have presented a viable approach to creating a question-answering virtual human for educational use within Holocaust education. *Interact* provides a significant opportunity for Holocaust museums and centres to preserve this vital educational experience and continue using testimony to support museum-based learning, to ensure that museum visitors of the future are able to access an experience that would be lost to them without the project, and to expand its audiences, by providing multiple opportunities to listen and interact with a survivor and providing access to the experience off-site in the future. Apart from applications within museum settings, *Interact* provides substantial opportunities for the wider arts sector to employ the model to create conversations between a pre-recorded photorealistic virtual human and audience. Future work should conduct experiments comparing a real Holocaust survivor with the virtual survivor over a video conference interface like Skype, i.e. a new variation of the Turing test, in order to investigate and evaluate its impact on human computer interaction.

Acknowledgments

The *Interact* project research and technical development 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. The project as a whole was also supported by a range of funders including the Pears Foundation and the Association of Jewish Refugees.

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.

Delecroix, F., Morge, M. and Routier, J. (2012) A Virtual Selling Agent Which Is Persuasive and Adaptive. In Agreement Technologies, S. Ossowski (Ed.), 625-645. Springer Netherlands.

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

Hoque, M., Courgeon, M., Martin, J. Mutlu, B. and Picard, R. (2013) MACH: my automated conversation coach. In the Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, 697-706. ACM New York, USA.

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

Ma, M., Coward, S. and Walker, C. (2015) Interact: A Mixed Reality Virtual Survivor for Holocaust Testimonies. In the Proceedings of 27th Annual Meeting of the Australian Special Interest Group for Computer Human Interaction (OzCHI '15), 250-254, ACM New York, USA.

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

Pandorabots.com (n.d.) Chatbot Captain Kirk. [Online] Available at: http://sheepridge.pandorabots.com/pandora/talk?botid=fef38cb4de345ab1&skin=iframe-voice [Accessed 1 Apr 2016].

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.

Talbot, T., Sagae, K., John, B. and Rizzo, A. (2012) Sorting Out the Virtual Patient: How to Exploit Artificial Intelligence, Game Technology and Sound Educational Practices to Create Engaging Role-Playing Simulations. In International Journal of Gaming and Computer-Mediated Simulations, 4(3): 1-19.

Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R., Field, D. and Cheng, W. (2011) A prototype for a conversational companion for reminiscing about images. In Computer Speech and Language 25 (2011): 140–157, Elsevier.