HKUST SPD - INSTITUTIONAL REPOSITORY

Title

Authors

Source

Version

DOI

Publisher

Copyright

This version is available at HKUST SPD - Institutional Repository (<https://repository.ust.hk/ir>)

If it is the author's pre-published version, changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published version.

A Live Storytelling Virtual Reality System with Programmable Cartoon-style Emotion Embodiment

Zhenjie Zhao Feng Han Xiaojuan Ma
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China
{zzhaoao, fhanab, mxj}@cse.ust.hk

Abstract—Virtual reality (VR) is a promising new medium for immersive storytelling. While previous research works on VR narrative have tried to engage audiences through nice scenes and interactivity, the emerging live streaming shows the role of a presenter, especially the conveyance of emotion, for promoting audience involvement and enjoyment. In this paper, to lower the requirement of emotion embodiment, we borrow experience from cartoon animation and comics, and propose a novel cartoon-style hybrid emotion embodiment model to increase a storyteller’s presence during live performance, which contains an avatar with six basic emotions and auxiliary multimodal display to enhance emotion expressing. We further design and implement a system to teleoperate the embodiment model in VR for live storytelling. In particular, 1) we design a novel visual programming tool that allows users to customize emotional effects based on the emotion embodiment model; 2) we design a novel face tracking module to map presenters’ emotional states to the avatar in VR. Our lightweight web-based implementation also makes the application very easy to use. We conduct two preliminary qualitative studies to explore the potential of the hybrid model and the storytelling system, including interviews with three experts and a workshop study with local secondary school students. Results show the potential of the VR storytelling system for education.

Index Terms—Emotion embodiment; social presence; avatar; storytelling; education; visual programming

I. INTRODUCTION

Virtual reality (VR), with its ability to immerse audiences in a scene that captures the real world or is completely imaginary, is a promising new medium for engaging storytelling [1]. Researchers and practitioners have explored applications of VR-enabled storytelling in tourism [2] and history education [3]. For instance, the Google VR Tour [4] is a recent application that allows users to explore tourist attractions via VR technologies. Simulating virtual characters that can interact with users has also shown to be useful for social anxiety treatment [5] and medical ethics training [6]. Moreover, VR has evolved from prototypes in laboratories to customer products that ordinary people can experience [7]. Users can now enjoy live stories through head-mounted devices (HMDs), thanks to advancements in affordable VR technologies.

To improve an audience’s experience, previous research into VR storytelling has focused primarily on enhancing the sensory quality [8], interactivity of a scene [9], and faithfully rendering virtual characters [5], [6]. With the increasing popularity of live streaming, an emerging form of storytelling [10], there is a growing interest in the role of a presenter in engaging

users [11], *i.e.*, telling stories to others. In conventional face-to-face storytelling, the physical presence of a presenter, especially the sharing of emotion, bounds the narrative to experience and renders it unforgettable. Similarly, in the digital field, embodying the presence and emotion of a presenter plays a crucial role for engaging storytelling [12]. For example, VTubers (short for “virtual YouTubers”) are virtual voice-over and motion-controlled avatars of video bloggers who used to speak face-to-face into the camera to their fans [13]. Audiences only see the performance of the avatar whose speech and motion are supplied by a (semi-)professional actor behind the screen. VTubers try to engage audiences by creating a lively digital embodiment of the presenter, which is achieved by a delicately crafted avatar and fine tracking of facial and body movements [13]. However, this approach is technically challenging for an ordinary person. First, to enhance the feeling of presence in a virtual world, a VTuber requires a high-quality, full-bodied motion capture system to track fine movements [13], which is unaffordable for ordinary users who broadcast from a regular personal computer (PC). Second, ordinary people with little or no performing arts backgrounds usually lack the ability to express rich emotions in a vivid yet natural manner.

Existing artists’ practice show that it is possible to create immersive VR experience through composing 2-dimensional (2D) images [14], [15], which is much easier than building 3-dimensional (3D) models. Therefore, to lower the barrier to perform engaging live storytelling in VR, we take a 2D approach to create emotional experience. Specifically, we propose a novel method to embody a presenter’s emotional states by integrating two components: 1) a simple cartoon-style animated avatar to represent emotions directly and 2) an auxiliary display of multimodal effects to enhance emotional intensity. We take inspiration from common emotion expressing methods in comics [16] and cartoon animation [17], and derive three key design requirements (**DR**):

- **DR**-(A). Clearly defined emotional state [17].
- **DR**-(B). Iconic abstraction of emotional meanings [16].
- **DR**-(C). Use multi-modalities to express an emotion [16].

For 1), we use comic-style facial expressions [18] to represent six basic emotions [19] of an avatar, which has a clearly defined emotional meaning for audiences (**DR**-(A)). Thanks to the comic-style [18], it is also abstract enough (**DR**-(B)) so that

no fine motion control is needed like a professional performer, which is important for ordinary users without enough artistic performing background. For 2), we support composition and customization of emotion-embedded special effects in different modalities. In this way, users can unite multi-modalities to give an auxiliary expression (**DR**-(C)) to enhance emotional effects.

To assist ordinary users in live VR storytelling, we adopt a design process [20] and develop a system to create and express the emotion embodiment. Users first edit their customized emotional effects via a block-based visual programming tool. During the live storytelling, a face tracking module is used to map users' facial information directly to the avatar, and they can also actively trigger the pre-edited emotional effects. The self-defined triggering actions and combinations of different emotional effects can alleviate the requirements of a high-quality motion capture system. In addition, with the emotion embodiment, we can also lower the performance requirement of the face tracking module, and make the tool more affordable for ordinary users. Moreover, we implement the system as a lightweight web-based application so that we can run it with minimal hardware and software configuration. We conduct two qualitative studies to explore the potential of our emotion embodiment model and the proposed system, including interviews with three experts and a workshop study with local secondary school students. Through the study, we show that with the emotion embodiment, a presenter will have a stronger feeling of being situated (presence) in the story. The audience will also better perceive the co-existence with the presenter (social presence), which will boost the audience's enjoyment and involvement.

II. RELATED WORK

A. Storytelling in VR

Previous research shows that an immersive experience in VR is more effective than 2D screens on evoking the sense of presence [21] and promoting appreciation of narratives [1]. For example, in [21], a contrast experiment shows that displaying VR stories increases audiences' immersion and presence. In [22], researchers explore the concept of immersive journalism which affords audiences access to the sights, sounds, even feelings and emotions of the news. Previous works have focused primarily on enhancing the sensory quality [8] and faithfully rendering scenes [9]. However, in face-to-face storytelling, the sharing of emotion also plays a non-negligible role on engaging audiences in the story [12].

B. Live Animation

Enabling a presenter's presence in VR through control of a virtual character, also called live animation, can potentially help engage audiences. This models after puppetry, an ancient art form of storytelling and a long time theatrical practice. Previous works of digitalized live animation include Chinese shadow puppetry simulation [23] and controlling marionette performance in a VR environment [24]. The recent development of facial and full body tracking techniques empowers new tools for creating and exercising live animation in the

Fig. 1. Illustration of basic design space of our hybrid emotion embodiment.

digital world [13]. To help users better interact with each other in VR, the Oculus avatar system also provides directly mapping users' facial or even body movements to their virtual counterparts [25]. All these works have focused primarily on personifying the movement of an avatar, while how to embody a presenter's emotion has not been fully explored, especially in live VR storytelling scenarios [25].

C. Emotion Representation

Emotional experience plays an important role in human-human interaction [26]. Ekman's discrete model is shown to be cross-cultural, and thus has been widely used in emotion recognition and representation [19]. In a virtual environment, users are deprived of many of their common channels of emotion communication [27]. Researchers thus seek alternative means to display and perceive emotions through various sensory experiences, *e.g.*, temperature [28], color [29], and vibration [30]. However, for live storytelling in VR, how to embody the emotion of a presenter is still a challenging problem. First, there are many channels of a presenter to convey emotion in offline settings. Mapping them one by one to an online counterpart is non-trivial [25]. Second, it is technically challenging to sense presenters' spontaneous emotion changes, especially through micro-facial expressions [31]. It is thus difficult for a system to automatically capture the emotional cues of users who are not trained to convey the feelings and moods of their characters.

III. HYBRID EMOTION EMBODIMENT IN STORYTELLING

Following the three key design requirements: clarity, iconic, and multi-modalities [16], [17], we propose to use a hybrid model to embody a presenter's emotions in the virtual counterpart, which combines a central expressive avatar and auxiliary displays of special effects. The avatar has clearly defined emotions (**DR**-(A)), and the auxiliary effects are iconic (**DR**-(B)). Users can combine them to express an enhanced emotional

Fig. 2. A: The system interface for the presenter, including A-a) menu, A-b) audience view, A-c) viewing angle, A-d) VR button, A-e) script editor, A-f) main canvas, A-g) camera view, A-h) recognized emotion distribution, A-i) storyline view. B: Block-based visual programming tool, including B1) the editing interface and B2) a multimodal effect example.

state (**DR-(C)**). The basic design space of the proposed model is illustrated in Figure 1. We present the implementation of this model via visual programming in Section IV.

A. Facial Expression of Avatar

To enable unambiguous communication of emotional states, it is critical to extract facial features that facilitate the perceptual process. In [18], Ma *et al.* summarize salient features for conveying Ekman’s six basic emotions in comics by conducting interviews with comic artists and analyzing human labels of cartoon emotions collected through crowdsourcing. We adopt the design guidelines from this work, and invite two professional graphics designers to create two virtual characters, a boy and a girl, with the ability to animate the basic facial expressions (Figure 1-center). Even though it is possible to design more sophisticated avatars that comply with these guidelines, in this paper, we as a proof-of-concept would like to investigate if the simplest graphic design could achieve the desired outcome.

B. Auxiliary Multimodal Display

While the facial indicator can deliver a clear emotional state to audiences, it alone is not adequate for evoking a strong emotional resonance. It is necessary to enhance and enrich the expression of emotion with other iconic modalities through different sensory channels. These modalities are particularly helpful for specifying emotional intensity, which is also important for expressive storytelling.

1) *Texture*: To animate the common background texture with motion lines in comics [16], we propose to configure a texture effect with five parameters. First, users can select an icon [18] or an emoji from our built-in library or some external source as the texture element. They can further set the size of the element as small, middle, or large. They can also define the motion of the texture from three aspects: velocity three levels (low, medium, or high), outgoing motion direction (four

cardinal and four intercardinal directions), and motion anchor. Motion anchor contains two options: a point anchor that emits texture images from a particular point in the VR scene, and a full screen anchor that emits texture images from a line anchor which is useful for mimicking rain, snow, *etc.*

2) *Text*: Similar to comics, text can be used as onomatopoeia to draw the audiences’ attention, or as a caption to assist in the narration of the storylines. Recent research on live streaming [32] finds that textual commentary, or danmaku in particular, is an effective means to reveal feelings. In the current design, users can input onomatopoeia, caption, and short comments as text elements, and specify the font size.

3) *Sound*: Non-speech sound is also an effective way to communicate emotions. We allow users to trigger several emotional human sounds, *e.g.*, laughing and crying, and add customized background sound effects at will.

IV. SYSTEM FOR SUPPORTING EMOTION EMBODIMENT

A. Overview

To enable ordinary users to compose and perform narrative in VR, we design and implement a live storytelling support system. A live VR story coming out of our system consists of three components: 1) a VR scene, 2) a verbal story told by a presenter, and 3) our emotion embodiment module. The emotion embodiment module contains an avatar and auxiliary multimodal effects. We design and integrate a block-based visual programming tool into our system so that users can easily create and customize such effects. During storytelling, a face tracking module analyzes the presenter’s facial information in real-time and directly maps the detected emotion, eye state, and head position to the avatar. The presenter can also actively trigger the pre-edited multimodal effects through keyboard and mouse. Meanwhile, the audience wears a HMD to explore the VR scene and watch the performance.

B. User Interface

1) *Presenter's Live Performing*: the interface used by the presenter during live performing is shown in Figure 2-A. In particular, a) is the menu for importing and exporting edited effects; b) is the real-time VR view mirroring the two screens in the HMD; c) is an overhead view to show the different viewing angles of the presenter and the audience as marked in the figure; d) is the button for starting and stopping displaying in the HMD; e) is a script editor for users to add, modify, and delete performance scripts; f) is the main canvas where users can drag the scene to change the presenter's view angle; g) is the camera view that captures the presenter's face; h) is an emotion recognition visualization view that displays the real-time recognition results across six basic emotions; and i) is the *storyline* view that records all the default and user-edited effects, where we call *storyline* as a group of edited effects. Users can add, delete, and (re-)name a storyline. Pressing the button **EDITOR** takes users to a visual editing tool (Figure 2-B1, introduced in the next section) to edit effects.

A typical user path through our system usually starts with presenters' editing and rehearsing a storyline using the visual editing tool. They can later save the composed effects in the storyline view i) and compile the associated script in the script editor e). Once ready, presenters can initiate live storytelling to an audience browsing in a VR environment. During the performance, the audience will see a virtual avatar presenting the narrative in front of them, whose facial animation is directly driven by the facial movement of the actual human presenter behind the camera. Presenters can monitor the audience's viewing angle through c). When they notice that the audience is no longer engaged with the story, *i.e.*, the audience is out of focus (Figure 2-c), presenters can try to attract their attention through triggering a pre-defined effect, *e.g.*, prompting a big smile on the avatar with flying bubbles and laughter onomatopoeia "haaaaa~" in the background (Figure 2-f) to display excitement by hitting the **a** key in the keyboard.

2) *Visual Programming of Emotion Embodiment*: following visual programming design guidelines [33], we develop a tool for intuitive event-based visual programming of animated effects. Such authoring environment is friendly to even novice users, allowing them to customize and combine different emotion expressing units by simply adding, configuring, and connecting the corresponding functional blocks (Figure 2-B1). By employing intuitive visualizations to represent function units and their relations [33], our tool is easy to master while providing sufficient adaptability and expandability.

In general, there are three kinds of blocks in the tool: triggering blocks, logic blocks, and functional blocks. Triggering blocks are used to specify triggering actions, including keyboard, mouse, counter, and facial expression. Logic blocks are used to connect multiple triggering blocks by logic conjunction (*And*) and disjunction (*Or*), which are helpful for designing flexible effect activation schemes. Functional blocks are used to define and change the state of any element occurring in

Triggering	
Keyboard	send a triggering signal by pressing 0-9, a-z, or A-Z
Mouse	send a triggering signal by pressing left or right mouse button
Counter	send a pre-set number of triggering signals
Facial Expression	send a triggering signal when the selected emotion is detected
Logic	
And	send a triggering signal when all inputs are triggered
Or	send a triggering signal when any input is triggered
Avatar	
Emotion	change the facial expression of the avatar
Position	change the position of the avatar
Role	change the role of the avatar, boy or girl
Reset	reset the position of the avatar to the center of the view
Texture	
Point	create a point texture
Full Screen	create a full screen texture
Attraction	create an attraction point to decorate the effects of a kinetic texture
Text	
Text	create a configurable text
Sound	
Human	configure human sound
Customization	configure emotion sound
Scene	
Background	change background image
Rotation	rotate the scene

TABLE I

SUPPORTED BLOCKS AND THEIR FUNCTIONS IN OUR TOOL.

the virtual environment. Each functional block corresponds to a different effect of our emotion embodiment (including avatar, text, texture, sound) and background scene. The full description of the currently supported blocks is provided in Table I. We cannot exhaust the full design space and the current blocks are implemented as a proof-of-concept. The system can be easily extended to incorporate richer options.

Users can load in new blocks from the drop-down menu on the top and connect different blocks to compose multimodal effects. For example, Figure 2-B demonstrates how to use the visual editing tool to create a hybrid surprised effect. Specifically, users can first choose a triggering block and assign a control, *e.g.*, keyboard **a** to be the effect handler. Users can then connect it to a facial emotion block, a texture block, a human sound block, and a text block. The output of each functional block can be configured by adjusting the corresponding parameters. When the keyboard **a** is pressed, all four effects will be activated simultaneously (Figure 2-B2). Users can preview the result and make adjustment iteratively.

3) *Audience's Interface*: An audience with a HMD can look around in 360 degrees, and the avatar and auxiliary visual effects are always displayed on a 2D plane in front to attract his/her attention. However, the presenter can adjust the position of the avatar and auxiliary effects in the 3-dimensional (3D) space through the visual programming tool.

C. Face Tracking for Live Storytelling

To animate the avatar a real-time manifestation of presenter's facial expressions and movements, we develop a face tracking module to capture users' emotion, eye state, and head position. We develop it as a local module to prevent possible network delay problems brought about by using online services. Video frames from the front camera are first processed by a feature point detection algorithm to get important 2D

facial landmarks. These landmarks and the corresponding face image are then sent to three separate modules, including emotion recognition, blink detection, and particle filtering. Finally, the outputs of the three modules are fused together to control the animation state of the avatar.

1) *Emotion Recognition*: We train a neural network on the FER2013 dataset [34] to recognize six basic emotions automatically. We use a state-of-the-art network [35] and a multi-class focal loss to deal with the unbalanced training sample distribution in FER2013. Apart from the fearful emotion, our model can reliably recognize other emotions in most cases ($\sim 60\%$), which is acceptable compared with human recognition accuracy [18].

2) *Blink Detection*: To classify different eye blinking states, we train another neural network on the CEW blink detection dataset [36]. Since this task is relatively easier than emotion recognition, we simply apply a common VGG16 network architecture and binary cross entropy loss. We use the same model for detecting both left and right eye. The final validation accuracy is about 96%.

3) *Particle Filtering*: To project the presenter's head position onto the VR scene, we bind the 2D position x, y of the avatar on the vertical plane to the detected 2D position of users' head. During live storytelling, the location of the avatar follows that of the users. To alleviate the drift and jitter problem, we adopt a standard particle filtering algorithm [37] for face position tracking.

D. Implementation

We implement the system as a lightweight web-based application by Javascript. For training neural networks, we use a PC with Intel(R) Core(TM) i9-7900X CPU 3.30GHz, 64GB RAM, Nvidia GeForce GTX 1080 Ti ($\times 2$). Our system can run with an ordinary camera and PC supporting HMD display. The current implementation only supports one-to-one storytelling.

V. USER EVALUATION

We explore the use of our system by secondary school students through a qualitative user study. We evaluate our system to see if it can help a presenter enhance his/her sense of presence and joy during storytelling, and help the audience in VR have a more enjoyable and involved experience via the social presence of the presenter. Instead of evaluating the performance of each individual module, we explore the system for its potential to help live storytelling in VR.

Before the user evaluation, we had regular group meetings to discuss and refine the user interface and system iteratively. We also conducted a pilot study with two members from our research group (aged 23 and 26) to identify potential problems of the system and the user evaluation procedure. We then conducted two studies in chronological order: 1). interviews with three experts on related fields, and 2). a workshop study with six participants. Study 1 is used to obtain feedback from experts about the system design. Study 2 is a formal study to obtain insights into the system.

For study 2, we adopt a pre-test and post-test study method [38] and ask participants to self-report their engagement to explore the effect of the cartoon-style emotion embodiment. The pre-test is our system with only the VR scene and verbal story, which can be seen as similar systems like Google VR Tour. In the post-test, we add the emotion embodiment module. The comparison of the pre-test and the post-test can provide us insights into our emotion embodiment module qualitatively.

1) *Pre-set Stories*: We select eight stories from Google Tour Creator Templates [4], and download their corresponding assets. The stories contain: 1) Jiuzhaigou National Park, 2) Mount Everest, 3) Chichén Itzá, 4) Earth Timeline, 5) Solar System, 6) Petra, 7) Antarctica, and 8) Seven New Wonders of the World. For each story, we use the first six scenes to make the workload of telling each story similar. We also use the captions from Google Tour Creator Templates as initial storylines for users to construct their storytelling structure and content. To help participants familiarise themselves with the visual programming tool, for each story, we build-in six emotional auxiliary effects, and participants can create their own effects on the basis of them.

A. Expert Interview

We conducted semi-structured interviews with three experts (E.1, E.2, and E.3) from related fields to get feedback on the system design. We recruited the experts via word of mouth. E.1 has about four years of live audio-visual performing experience, which is an art form that is similar to our storytelling scenario [39]. E.2 has been an industrial visual designer for three years. E.3 has been working on cartoon making with graphics software for four years.

1) *Procedure*: We first asked participants to try out our system, and then collected their suggestions, including pros and cons of the design and user experience. Each interview lasted from 30 minutes to one hour. The interview with E.1 is on-site, and the interviews with E.2 and E.3 are remote via a voice messaging application. We used the thematic analysis method [40] to analyze the interview scripts.

2) *Feedback and Analysis*: As a proof-of-concept, our system receives positive feedback from E.1, E.2, and E.3.

Feedback on Emotion Expressing. Both E.1 and E.3 acknowledged the usefulness of the multimodal method for expressing emotions in VR. E.1 described using “*sound*” and “*motion*” as “*effective ways*” for evoking audiences’ emotions in live stage performances, which is similar to cartoon animation and comics “*in principle*”. E.3 also confirmed that “background music” and “texture” play an important role in expressing emotions in cartoons. The feedback shows that the adoption of emotion expressing techniques from cartoon animation and comics is viable for live VR storytelling. E.1 was particularly fond of the use of emojis for the texture effect. “*In current live streaming platforms, people love to use emojis, or kaomojis, to directly express emotions. The movements of emojis can evoke audiences’ emotional resonance easily.*”-E.1.

Feedback on Visual Programming. The design of the block-based visual programming tool aims to lower the bar for

ordinary users to customize different effects. E.1, E.2, and E.3 all acknowledged its usefulness, as E.1 described it as “convenient” and “flexible”. “I always use touchdesigner to pre-edit audio-visual effects before my performance. Compared to writing code manually, node programming (block-based visual programming) has a shorter development cycle, and is more convenient and flexible for people who are not coding experts. For example, when I construct some functions, I can easily modify previously edited modules by adding or removing blocks, and it would be a disaster if I have to modify it by typing lines and lines of code.”-E.1. In addition, different from professional users who are eager to explore new effects with the tool, E.1 indicated that it is better to provide pre-defined effect libraries for an ordinary user. “For general users, like live streamers, they may not want to learn new things. They just demand something that is ready to work. So you’d better prepare some built-in effects for them.”-E.1. Following the experts’ suggestion, we provide six default emotional auxiliary effects for users to create their own effects on top of them.

Feedback on Face Tracking. E.1 found the facial emotion recognition function a nice feature, especially for ordinary users. “I will use the emotion recognition function. I think other people would too. We always like to try things with minimal mental efforts. Controlling things with facial expressions is a perfect example of that.”-E.1.

B. Workshop Study

To further gain insights into the efficacy of emotion embodiment in VR-enabled live storytelling, we conducted a workshop study to explore our system’s impact on user experiences from both the presenter’s and the audience’s perspectives. On the presenter’s side, we evaluate his/her sense of presence and joy during storytelling. On the audience’s side, we assess his/her level of enjoyment and involvement when exposed to embodied emotional expressions.

1) *Apparatus:* We use Amazon Web Service (AWS)¹ to host the system. Each participant has a unique user account, with which they can edit and store their own customized effects on the server. We run the live storytelling client interface on a laptop with an Intel CPU i7 2.8GHz, 16GB RAM, and Nvidia GeForce 1070 GPU. We use an Oculus Rift² for audiences to view the story scenes in VR. We also use four projectors to broadcast the VR scene to other viewers without HMDs.

2) *Participants:* We organized a workshop titled *emotion embodiment for live storytelling in VR*, and sent advertising posters to local schools. Five students (aged 16) under the supervision of one teaching assistant (TA, aged 22) from a local secondary school joined the workshop voluntarily. All participants (P.1~P.6) reported that they had played with VR devices several times before.

3) *Procedure:* After welcoming the participants, we first obtained their consent (and also their guardians’ consent for

Fig. 3. User ratings from the presenter’s perspective.

students under 18). After that, we provided an overview of the workshop and introduced potential usage scenarios of the system, e.g., VR storytelling for tourism and education. Each participant then got a randomly assigned unique ID that each corresponds to a different story from the pool.

The workshop then proceeded to three sessions in chronological order: 1) storytelling without emotion embodiment (pre-test), 2) emotion embodiment training, and 3) storytelling with emotion embodiment (post-test). In Session 1, we first introduced the basic usage of the system with the emotion embodiment function hidden from the interface (the avatar and visual editing multimodal effects). The participants took turns to perform their stories, and one of the audiences viewed it in VR while the others watched the live broadcast on four projected screens with Oculus Mirror. Each presenter could have at most 5 minutes for his/her storytelling. At the end of Session 1, every participant got to be the presenter and the VR audience once. After a short break, in Session 2, we introduced the emotion embodiment module in detail, and asked participants to try and create their customized effects for their assigned stories. Each participant used a separate PC to edit and store their results on the server. Two investigators walked around taking questions from the participants. This session lasted for about 30 minutes followed by another short break. In Session 3, we repeated the same storytelling procedure with the emotion embodiment module enabled. After each session, the participants filled in a questionnaire to report their user experience. An exit interview was conducted to collect more detailed feedback. We video-recorded the whole procedure with consent.

4) *Findings: Presenter.* It usually took a similar time (2~5 minutes) for both pre-test and post-test storytelling, which shows that our emotion embodiment module does not bring about any extra heavy workload for presenters. With the emotion embodiment module, we find that presenters are more likely to actively tell the story emotionally. For example, presenters try to pose different facial expressions, change voice tone (although we do not support voice emotion recognition), and trigger pre-edited effects (3~9 times per person).

¹<https://aws.amazon.com>

²<https://www.oculus.com/rift>

Fig. 4. User ratings from the audience's perspective.

After pre-test and post-test performances, we asked each presenter to report their feeling of presence on a 7-point Likert scale. In particular, we asked participants to evaluate their agreement level on statements: 1). I can express my emotions easily, and 2). I can interact with the audience easily. The divergent percent results are shown in Figure 3, from which we can see that generally with the emotion embodiment module, presenters are more likely to have a better feel of expressing emotions and interacting with audiences. Participants think that with the emotion embodiment module, they can influence what the audience see better. Another interesting observation is that equipped with the emotion embodiment, we find one presenter actively asked the audience questions during the storytelling, which shows that enabling emotion embodiment actually makes the presenter engage more actively in the performance. However, we also find that the emotion recognition accuracy does influence the option of one participant. *"The avatar looks funny when I present my fearful emotion for introducing dinosaurs. I think it cannot understand fear intelligently."*—P.2 (as presenter).

We further asked the participants to report their perceived enjoyment on a 7-point Likert scale. The result is shown in Figure 3. In general, the participants find that with our emotion embodiment module, they have a more interesting storytelling experience. Please note that for pre-test and post-test, each participant presented the same story. In other words, having the emotion embodiment capability, even presenting the same story another time can bring fun experience to presenters. *"I think it is fun and enjoying to have the emotion embodiment. I love to use it."*—P.4 (as presenter).

Audience. On a 7-point Likert scale, we asked the participants to evaluate the perceived enjoyment and involvement. In particular, we asked the participants to evaluate their agreement level on statements: 1) Another person in the VR is present with me. 2) I am emotionally involved. 3) I perceive the presenter's emotions in the VR scene. The results are shown in Figure 4. In general, with the emotion embodiment module, audiences can feel the social presence of the presenter during storytelling, which ultimately boosts their involvement

and enjoyment experience. *"It (the emotion embodiment) can directly express the feeling of the presenter just like a direct view of facial expressions and emojis."*—P.4 (as audience). *"With the avatar and the effects, I feel I am actively interacting with the presenter. It is fun and interesting."*—P.2 (as audience).

VI. LIMITATIONS

Despite the encouraging results, there are several limitations of the current proof-of-concept system. First, we only run a small scale workshop study to evaluate the system qualitatively. Although the study provides us some insights into the system design, it is also important for us to further investigate the system more systematically and quantitatively through controlled experiments. Second, the current system does not support customization of the avatar. We can enrich it by providing users with more personalized facial features to create their own avatar and expressions. Third, non-verbal communication, such as body language, also plays an important role in human-human communication. It is interesting to further investigate the design and control of an upper body or even full body avatar, and how to embody emotions on it.

VII. CONCLUSION

We propose a new form of emotion embodiment by combining facial expression and auxiliary multi-modalities for live storytelling in VR. We design a proof-of-concept system to support authoring and performing with an emotionally embodied avatar. Specifically, we design a block-based visual editing tool to help users customize multimodal effects, and develop a simple face tracking module to map presenters' facial states to a virtual avatar. We demonstrate the potential of the VR storytelling system through qualitative user evaluation, which consists of interviews with three experts in related fields and a workshop study with six participants. In the future, we plan to improve the embodiment model [41], and extend the system to support multiple audiences and remote storytelling, which can better empower real-world applications.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank Zhenhui Peng for developing the first versions of emotion recognition and blink detection algorithms, and Meng Xia and Michelle Wong for designing the artworks. This work is supported by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grant No.: T44-70716-N.

REFERENCES

- [1] R. Pausch, J. Snoddy, R. Taylor, S. Watson, and E. Haseltine, "Disney's Aladdin: first steps toward storytelling in virtual reality," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1996, pp. 193–203.
- [2] D. A. Guttentag, "Virtual reality: applications and implications for tourism," *Tourism Management*, vol. 31, no. 5, pp. 637–651, 2010.
- [3] P. Hodgson, V. W. Lee, J. C. Chan, A. Fong, C. S. Tang, L. Chan, and C. Wong, "Immersive virtual reality (IVR) in higher education: Development and implementation," in *Augmented Reality and Virtual Reality*. Springer, 2019, pp. 161–173.
- [4] G. V. Tour, "Create a virtual tour: Tour creator makes it easy to build immersive, 360° tours right from your computer," Retrieved March 25, 2019. [Online]. Available: <https://vr.google.com/tourcreator>

- [5] X. Pan, M. Gillies, and M. Slater, "Virtual character personality influences participant attitudes and behavior—an interview with a virtual human character about her social anxiety," *Frontiers in Robotics and AI*, vol. 2, p. 1, 2015.
- [6] X. Pan, M. Slater, A. Beacco, X. Navarro, A. I. B. Rivas, D. Swapp, J. Hale, P. A. G. Forbes, C. Denvir, A. F. d. C. Hamilton *et al.*, "The responses of medical general practitioners to unreasonable patient demand for antibiotics—a study of medical ethics using immersive virtual reality," *PloS one*, vol. 11, no. 2, p. e0146837, 2016.
- [7] L. Avila and M. Bailey, "Virtual reality for the masses," *IEEE Computer Graphics and Applications*, vol. 34, no. 05, pp. 103–104, sep 2014.
- [8] C. Timmerer, M. Waltl, B. Rainer, and N. Murray, "Sensory experience: Quality of experience beyond audio-visual," in *Quality of Experience*. Springer, 2014, pp. 351–365.
- [9] M. Craven, I. Taylor, A. Drozd, J. Purbrick, C. Greenhalgh, S. Benford, M. Fraser, J. Bowers, K.-M. Jää-Aro, B. Lintermann *et al.*, "Exploiting interactivity, influence, space and time to explore non-linear drama in virtual worlds," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2001, pp. 30–37.
- [10] Z. Deng, P. Benckendorff, and J. Wang, "Blended tourism experiencescape: A conceptualisation of live-streaming tourism," in *Information and Communication Technologies in Tourism 2019*. Springer, 2019, pp. 212–222.
- [11] N. Saquib, R. H. Kazi, L.-Y. Wei, and W. Li, "Interactive body-driven graphics for augmented video performance," in *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, 2019.
- [12] R. M. Baños, C. Botella, M. Alcañiz, V. Linao, B. Guerrero, and B. Rey, "Immersion and emotion: their impact on the sense of presence," *Cyberpsychology & Behavior*, vol. 7, no. 6, pp. 734–741, 2004.
- [13] B. Lufkin, "From japanese anime characters to barbie, virtual YouTubers talk and act just like people and they could change the way we all interact forever," Retrieved March 13, 2019. [Online]. Available: www.bbc.com/capital/story/20181002-the-virtual-vloggers-taking-over-youtube
- [14] S. Hayden, "Facebook artist makes a case for VR manga in homage to iconic 'dragon ball' scene," Retrieved October 10, 2019. [Online]. Available: <https://www.roadtovr.com/facebook-artist-makes-case-vr-manga-homage-iconic-dragon-ball-scene>
- [15] British Museum, "Chinese painted scroll," Retrieved October 10, 2019. [Online]. Available: <https://www.facebook.com/britishmuseum/videos/10156035794244723/>
- [16] S. McCloud, "Understanding comics: the invisible art," *Northampton, Mass*, 1993.
- [17] F. Thomas, O. Johnston, and F. Thomas, *The illusion of life: disney animation*. Hyperion New York, 1995.
- [18] X. Ma, J. Forlizzi, and S. Dow, "Guidelines for depicting emotions in storyboard scenarios," in *8th International Design and Emotion Conference*, 2012.
- [19] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motivation*. University of Nebraska Press, 1971.
- [20] M. W. Easterday, D. G. R. Lewis, and E. M. Gerber, "The logic of design research," *Learning: Research and Practice*, vol. 4, no. 2, pp. 131–160, 2018. [Online]. Available: <https://doi.org/10.1080/23735082.2017.1286367>
- [21] A. Tse, C. Jennett, J. Moore, Z. Watson, J. Rigby, and A. L. Cox, "Was I there?: impact of platform and headphones on 360 video immersion," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '17. New York, NY, USA: ACM, 2017, pp. 2967–2974. [Online]. Available: <http://doi.acm.org/10.1145/3027063.3053225>
- [22] N. De la Peña, P. Weil, J. Llobera, E. Giannopoulos, A. Pomés, B. Spanlang, D. Friedman, M. V. Sanchez-Vives, and M. Slater, "Immersive journalism: immersive virtual reality for the first-person experience of news," *Presence: Teleoperators and Virtual Environments*, vol. 19, no. 4, pp. 291–301, 2010.
- [23] F. Lu, F. Tian, Y. Jiang, X. Cao, W. Luo, G. Li, X. Zhang, G. Dai, and H. Wang, "Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 1919–1928. [Online]. Available: <http://doi.acm.org/10.1145/1978942.1979221>
- [24] M. Sakashita, T. Minagawa, A. Koike, I. Suzuki, K. Kawahara, and Y. Ochiai, "You as a puppet: evaluation of telepresence user interface for puppetry," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '17. New York, NY, USA: ACM, 2017, pp. 217–228. [Online]. Available: <http://doi.acm.org/10.1145/3126594.3126608>
- [25] M. Howard, "Avatars: the art and science of social presence," Retrieved April 4, 2019. [Online]. Available: <https://tech.fb.com/avatars-the-art-and-science-of-social-presence>
- [26] R. J. Dolan, "Emotion, cognition, and behavior," *Science*, vol. 298, no. 5596, pp. 1191–1194, 2002.
- [27] M. Hassib, M. Pfeiffer, S. Schneegass, M. Rohs, and F. Alt, "Emotion actuator: Embodied emotional feedback through electroencephalography and electrical muscle stimulation," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 6133–6146. [Online]. Available: <http://doi.acm.org/10.1145/3025453.3025953>
- [28] J. Tewel, J. Bird, and G. R. Buchanan, "The heat is on: A temperature display for conveying affective feedback," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 1756–1767. [Online]. Available: <http://doi.acm.org/10.1145/3025453.3025844>
- [29] L. Bartram, A. Patra, and M. Stone, "Affective color in visualization," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 1364–1374. [Online]. Available: <http://doi.acm.org/10.1145/3025453.3026041>
- [30] P. Bucci, X. L. Cang, A. Valair, D. Marino, L. Tseng, M. Jung, J. Rantala, O. S. Schneider, and K. E. MacLean, "Sketching cuddlebits: coupled prototyping of body and behaviour for an affective robot pet," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 3681–3692. [Online]. Available: <http://doi.acm.org/10.1145/3025453.3025774>
- [31] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikinen, "Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 563–577, Oct 2018.
- [32] Y. Chen, Q. Gao, and P.-L. P. Rau, "Watching a movie alone yet together: understanding reasons for watching danmaku videos," *International Journal of Human Computer Interaction*, vol. 33, no. 9, pp. 731–743, 2017. [Online]. Available: <https://doi.org/10.1080/10447318.2017.1282187>
- [33] D. Weintrop, A. Afzal, J. Salac, P. Francis, B. Li, D. C. Shepherd, and D. Franklin, "Evaluating coblox: a comparative study of robotics programming environments for adult novices," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18. New York, NY, USA: ACM, 2018, pp. 366:1–366:12. [Online]. Available: <http://doi.acm.org/10.1145/3173574.3173940>
- [34] Kaggle, "Challenges in representation learning: facial expression recognition challenge," Retrieved March 30, 2019, <https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge>
- [35] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," *CoRR*, vol. abs/1710.07557, 2017.
- [36] F. Song, X. Tan, X. Liu, and S. Chen, "Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients," *Pattern Recognition*, vol. 47, no. 9, pp. 2825 – 2838, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320314001228>
- [37] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," vol. 50, no. 2. Ieee, 2002, pp. 174–188.
- [38] J. L. Green, G. Camilli, and P. B. Elmore, *Handbook of Complementary Methods in Education Research*. Routledge, 2012.
- [39] M. Faulkner, *VJ: Audio-Visual Art and VJ Culture: Includes DVD*. Laurence King Publishing, 2006.
- [40] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.
- [41] E. J. Carter, M. Mahler, M. Landlord, K. McIntosh, and J. K. Hodgins, "Designing animated characters for children of different ages," in *Proceedings of the The 15th International Conference on Interaction Design and Children*, ser. IDC '16. New York, NY, USA: ACM, 2016, pp. 421–427. [Online]. Available: <http://doi.acm.org/10.1145/2930674.2930681>