Technical Correspondence

A Text-Driven Conversational Avatar Interface for Instant Messaging on Mobile Devices

Mario Rincón-Nigro and Zhigang Deng

Abstract—**In this letter, we investigate the use of conversational avatars as a means to improve the user experience on instant messaging (IM) for mobile devices. We describe the design and implementation of an interface for IM featuring a 3-D facial avatar that is driven by the text messages being exchanged between chatting participants. Our design is affordable under the limited computational capacity of current generation mobiles. We evaluate user acceptance and reaction via user studies, by comparing it to a more conventional IM interface, and provide recommendations for the effective design of conversational avatar interfaces for mobile applications.**

Index Terms—**Affective interfaces, conversational avatars, mobile devices, text messaging, user experience, user modeling.**

I. Introduction

Arguably, mobile phones are one of the most important communication tools available these days. In this context, instant messaging (IM) stands out among the most prominent mobile phone applications. About a trillion messages are being exchanged each year through applications such as Facebook Messenger, WhatsApp, WeChat, Line, Kik, and BlackBerry Messenger, based on the Global Mobile Statistics 2012 (http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats).

Although IM has been both widely and successfully adopted, its conventional text format lacks the support for many of the nonverbal hints that are required for effective communication. In face-to-face communication, facial expressions, voice prosody, and gestures are vital elements for conveying meaningful information and establishing the context of the communication [1].

To this effect, conversational avatar interfaces can be an alternative and effective way to cover these needs for nonverbal expression in IM for mobile phones. Well-known applications, such as video conference (e.g., Skype, iOS FaceTime, Tango), have the potential to cover the same needs. However, in some situations highly realistic avatars are preferable than video chatting. For instance, if a participant wants to influence the perception of other participants by changing the image he/she projects [2] (e.g., the participant is sick and wants to appear as being healthy), or the participant wants to remain anonymous, then an avatar would be preferable. In terms of entertainment, conversational avatars also offer an advantage over video conferences as users can easily change their features (i.e., hair style and color, jewelry, skin tone, freckles) to correspond with the self-image they wish to project [3]. High realism in avatars leverages effectiveness in communication [4], and symbolism in avatars leverages expressiveness and enjoyability of user experience [5]. Realistic 3-D avatars can be an attractive option as they compromise between the aesthetic and behavioral realism of the high-fidelity representations (e.g., video conference), and the semantic and emotional flexibility of abstract representations (e.g., 2-D avatars/cartoons).

In this letter, we investigate novel ways to improve the user experience on IM for mobile devices by using realistic conversational avatars. The goal of our approach is to imprint the avatars with a high level of aesthetic and behavioral realism. We describe the design and implementation of an IM client for mobile phones that features an animated conversational 3-D facial avatar interface. The avatars are driven by the text messages exchanged between chatting participants, which exhibits realistic lip-synchronized facial animation as well as head and gaze motions. We adopt the eFASE framework [6] for lip-synchronized animation synthesis. The framework that we describe is affordable under the limited computational capacity offered by current generation mobile devices. We evaluate the conversational avatar interface system and investigate user acceptance and reaction, by comparing it with a more conventional IM interface via user studies. From our experience developing the application, as well as lessons learned from our user study, we provide suggestions and guidelines that we consider to be useful for the effective design of application interfaces that features multimodal realistic conversational avatars on mobile devices.

II. Related Work

Avatars in virtual environments [7], their control [8], and their personalization [3] have been extensively studied. Internet chat spaces where symbolic 2-D avatars are used for social interactions are described by Smith *et al.* [5]. They report that control features requiring complex actions from the user tend to decline in time. Following this suggestion, we opt for full text-driven control of avatars as it requires minimal effort from the users. Low-realism symbolic avatars for IM, and related applications, as a medium to enhance user experience, and for the expression of nonverbal communication cues, have also been widely studied. User acceptance and emotional bonding for symbolic avatars on mobile phone applications are investigated in [9] and [10]. A system for concatenating animations of symbolic 2-D avatars and attaching them to messages is described in [11]. The use of highly realistic communicational avatars remains as an active ongoing research area [4], [12]. We ground our design decision of imprinting high realism to the conversation avatars on the previously reported evidence that higher aesthetic and behavioral realism leverages engagement, and leads to a higher sense of "social copresence" [4], [12]. That is, realistic representations help participants feel communication is taking place within a shared space at present time, and increase the sense that they have access to each others' thoughts and intentions, which improves the effectiveness of computer-mediated communication.

III. Text-Driven Avatar Interface

The proposed avatar interface stands in the ground between the highest fidelity representation of users (e.g., video chat) and low-realism/symbolic avatars. It is capable of automatically creating animations that feature high-fidelity lip-synchronized animation and speech, and nonverbal cues such as head motion and eye gaze, from the exchanged messages. An architecture overview of our text-driven avatar messaging system is shown in Fig. 1. The whole process starts with the

Manuscript received August 17, 2012; revised January 18, 2013; accepted February 1, 2013. Date of current version April 15, 2013. This work was supported in part by NSF IIS-0914965 and in part by unrestricted research gifts from Google and Nokia. This paper was recommended by Associate Editor Y. Xiao of the former IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans (2011 Impact Factor: 2.123).

The authors are with the Department of Computer Science, University of Houston, Houston, TX 77204-3010 USA (e-mail: mario.rincon.nigro@gmail.com; zdeng@cs.uh.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2013.2250498

2168-2291/$31.00 © 2013 IEEE

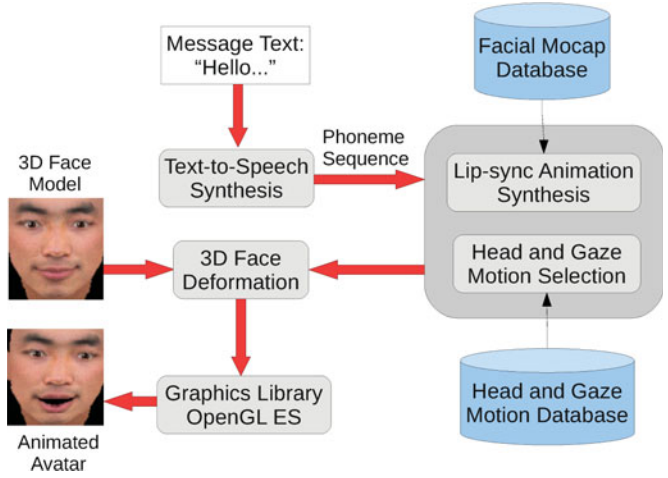
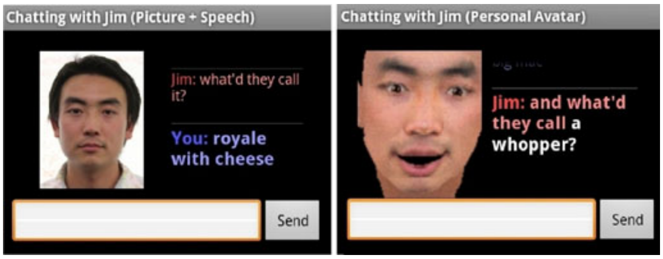


Fig. 1. Semantic illustration of the architecture of the designed chatting avatar application/interface.



Chatting with Jim (Picture + Speech)

Jim: what'd they call it?

You: royale with cheese

Chatting with Jim (Personal Avatar)

Jim: and what'd they call a whopper?

Fig. 2. Screenshots of the avatar messaging interface. (Left) Conventional interface. (Right) Conversational avatar interface.

texts inputted by the user. All of the words in the input texts are fed into a module for text-to-speech synthesis (TTS) that produces a phoneme sequence with timing information, and an audio track for the corresponding speech. The phoneme sequence is then fed into a module for lip-synchronized facial animation synthesis. This module makes use of motion capture data for face, head, and gaze motion. The database for face motion holds sequences of facial marker frames (i.e., marker positions) that correspond to the utterance of every possible phoneme. The database for head and gaze motion holds sequences of rotation angles for pitch, roll, and yaw of head and eyes. The animation synthesis produces a new sequence of frames by concatenating and smoothing the motion capture data, as well as a sequence of rotation angles for head and gaze motion. The marker positions on the new sequence are used as control points to drive the deformation of a face model during the animation, and the rotation angles are used to transform the head and eyes of the model. The animation frames are then rendered and displayed in the device screen in a synchronized manner along with the audio track. Screenshots of the graphical user interface for our IM client are shown in Fig. 2.

A. *Lip-Synchronized Facial Animation Synthesis*

We adopt the eFASE framework [13] for lip-synchronized animation synthesis. The eFASE technique is a data-driven technique that makes use of facial motion capture data to select an optimal sequence of motion nodes that represent the observed facial motion caused by the utterance of a given phoneme sequence (for algorithm details, refer to [6] and [13]). A motion node is a segment of facial marker frames, which in our case represents facial motion for a phoneme. The eFASE framework is based on a dynamic programming algorithm that minimizes a penalty function that supports the synthesis of natural and smooth facial animation.

We adapt the eFASE framework to cope with the device limitations of mobile devices, and responsiveness constraints implied by IM applications, by clustering and reducing the employed facial motion dataset. In addition, we adopt the linear shell algorithm [14] for facial mesh deformation. To animate the head and the eyes on our avatar, we also make use of motion capture data. Alongside with the facial animation, sequences of head and gaze motion are selected to make our avatars move their heads and eyes. The sequence of rotation angles is selected from the dataset based only on the length of the corresponding sequence of frames for the facial motion. We randomly select one of the sequences within the dataset, play it until half of the face sequence, and then repeat it backward. For every utterance, our avatar head starts and ends at a rest pose after performing the motion. Eye motion is created in a similar fashion. Although the motion is not directly correlated to the avatar speech, the use of motion capture data enforces the realism of the chosen head and gaze motion.

B. *Implementation on Mobile Devices*

Our application is targeted and tested on the Nexus One Android mobile phones. Most of our program has been implemented in Java using the Google Android SDK. The Flite library, written in C, is used for TTS synthesis, and our eFASE implementation is written in C++ for performance reasons. Java native interfaces (JNI) are created by using the Android NDK for both the libraries, in order to be used from the Java application. For the Flite library, we have both a generic female and a generic male voice with neutral prosody tone for the English language. Chat clients communicate with each other by establishing a TCP connection through sockets, and IP addresses from other clients are received from a logging server. Model rendering is done using Android SDK support for OpenGL ES. A small diffuse light shader is written in GLSL and used to render the models. Model geometry is stored in Vertex Buffer Objects and every animation frame is rendered offline to an OpenGL render buffer object, and later redisplayed to satisfy animation synchronization and timing requirements. It exhibits a small delay of about 3 seconds for an average sentence consisting of 25 phonemes.

The main technical challenge when adapting eFASE to the mobile platform is related to its performance. The synthesis algorithm has $O(N^2T)$ time complexity, where N is the number of nodes in the motion capture database, and T is the length (in phonemes) of the input text. Since IM can take place in both synchronous and asynchronous modes, we want our framework to be able to handle both cases. We speed up the animation synthesis in order to comply with the response-time constraints of synchronous messaging by reducing the motion capture database. We normalize all motion nodes within the database to have the same length in terms of frames and cluster the database through K-means. There is a tradeoff at this point between the synthesis speed and the quality of the resulting animation. The quality of the animation is proportional to the size of the space of face motion that is covered by the database. We reduce a database to 10% of its original size in terms of motion nodes (the reduced database contains 1076 motion nodes), which makes sure that at least one motion node corresponding to each phoneme is present. Though the original database contains samples for phonemes uttered using different emotions, the reduced database does not offer enough coverage of the facial expression space for synthesizing high-quality expressive facial motions. Another issue with including emotion in this letter is the TTS synthesis with emotions, which is actually an actively ongoing research problem in speech synthesis. The voice databases in the Flite library as well as other existing TTS systems do not commonly support this feature. For rendering and mesh deformation, the resolution of the mesh had to be adjusted as well. A high-resolution mesh would be too slow to render on the device, along with synthesis time. Undersampling the mesh alleviates this problem, but at the cost of reducing the quality of deformation of the mesh. A correspondence between the vertices of the mesh and the control marker points in our database needs to be established, and they should correspond closely for better results. This requires careful editing of the meshes, which can be time consuming.

IV. User Study Design

We probe user acceptance and reaction to our avatar interface, to obtain information from potential users about desirable and undesirable features, as well as to obtain sound guidelines for the design of similar applications. For this purpose, we recruit 20 participants. Their ages range between 18 and 35 (average: 25.80; standard deviation: 6.98), 15 males and 5 females. Some of our participants previously knew each other. All of our participants are regular users of IM, and reported an average of 10.85 weekly hours spent in messaging activities. Most of them are students at a university (both undergraduates and graduates) with balanced ethnic backgrounds. Each participant was rewarded a Starbucks $10 gift card to compensate his/her time spent in the study.

The user study takes place during the course of two weeks and consists of two sessions for each pair of participants. During the first session, pictures of the front, right side and left side of the faces of participants are taken, under appropriate lighting conditions. Then, by using the FaceGen software, we create a personalized (i.e., individual-specific) avatar for every participant, holding a high resemblance to the actual participant in terms of both geometry and texture. The models are textured using the taken pictures and include skin details (e.g., correct color tone, freckles, marks), face jewelry, and the same eye color. However, our models do not include hair, teeth, tongues, or head props such as glasses. After the first session, each model is then set up to be used for animation including manually establishing a one-to-one correspondence between model vertices and facial marker positions in our motion capture dataset. A constructed avatar example is shown in Fig. 2.

In the second session, each pair of participants is asked to chat between each other for a lapse of 12 min. The participants sit in small isolated cubicles in order to provide privacy and reduce conversational pressure on their sides. Headphones are also provided to avoid the participants from hearing speech synthesized from their own sent messages. Thus, during the chatting part of the session, the participant pair could not see each other and could not hear the audio of the other phone. In order to faithfully resemble a real-world IM communication, the participants are told to carry a regular conversation as they would normally do when using IM, instead of asking them to achieve a certain goal or to play a role in an artificial task. Before the study, participants are explained the functionality of the application. Two Nexus One HTC phones are used in the chatting sessions. The Nexus One mobile phone features a multitouch capacitive touchscreen, with 3.7 in (94 mm) diagonal length and a resolution of 480×800 pixels.

Data in our user study are collected in two different ways: 1) participants are asked to answer a 5-scale Likert-type questionnaire, and 2) a short poststudy interview is conducted to obtain user feedback and free-form comments. The purpose of our questionnaire is to capture the impressions of the participants on the usability of the avatar messaging interfaces, as well as to obtain their comparative preferences with respect to the enjoyability of each messaging mode and their perception of interface features.

A. Instant Messaging Modes

The IM client features two different messaging modes. Screenshots of the messaging modes are shown in Fig. 2. The chatting modes in our experiments are selected by taking into account the following two criteria: 1) the difficulty of implementation and 2) the level of expression that it allows. The goal is to assess the value of each mode by taking into account the enjoyability of the mode versus its implementation cost.

1) The picture and TTS mode (PS) displays a static picture of the other end user (i.e., the chatting partner) and plays synthesized speech audio track for each message being received. The message history is displayed, with attenuated blue-colored sent messages and attenuated red-colored received messages. This mode enables nonverbal expressions through a photo of the user and TTS. Expression of emotions at this level using TTS would require the creation of phoneme databases for each enabled expression, which is difficult in reality. In our implementation, we use only neutral expression.
2) The personalized avatar (PA) displays an animated 3-D avatar. The avatar is modeled and textured based on the front and side pictures of the chatting participant. As every message is received, a lip-synchronized animation of the avatar uttering the content of the message is played along with TTS synthesized speech. The mode requires the creation of a 3-D PA model for each user as previously described. Setting up the PA model can be performed in a semiautomatic manner.

V. User Study Result Analysis

A. Avatar Messaging Interfaces

Participants were asked to rate on a scale from 1 to 5 (5 means "strongly agree" and 1 means "strongly disagree") in terms of their level of agreement with the following statements.

Perceived Realism

1) *Model: I think the face model of my chatting partner looks realistic enough.*
2) *Animation: I think the avatar animation looks natural.*
3) *Head: I noticed the head motion on the avatar interfaces.*

Engagement

1) *Participate: I felt more like participating in the conversation while using the PA interfaces than the PS interface.*
2) *Empathy: I felt more empathy with my chatting partner while using the PA than the PS interface.*
3) *Distracting: I feel the animated avatars are distracting from the content of the conversation.*

Usability

1) *Every Day: I would use the text-driven avatar chatting application for my everyday chatting if it were available.*
2) *Informal: I would use the avatar interface for informal chatting with my friends and family.*
3) *Formal: I would use the avatar interface for formal conversations (e.g., interview or discussion).*
4) *Delay: I noticed a delay on message delivery while using the text-driven avatar interfaces.*

Descriptive statistics for the questions appear in Table I. Based on the mode values below 3 for the *Model* realism question, the participants do not think the avatars are realistic. Regarding the question about the *Animation* realism, the participants are neutral about the avatar animation looking natural. The Spearmans rank correlation coefficient reveals a statistically significant relationship between the perceived realism for the avatar model and the lip-sync animation

TABLE I
DESCRIPTIVE STATISTICS FOR SCORES RELATED TO PERCEIVED AVATAR REALISM, ENGAGEMENT, AND USABILITY

Question	Mean	S. Dev.	Median	Mode
Model	2.85	1.18	3	2
Animation	3.05	0.89	3	3
Head	3.15	1.63	3.5	5
Participate	3.95	0.94	4	5
Empathy	2.85	0.99	3	3
Distracting	1.95	1.10	2	1
Every Day	3.15	1.18	3	4
Informal	3.9	1.12	4	4
Formal	2.45	1.50	2	1
Delay	2.2	1.10	2	2

$(rs[601.702] = 0.548, p = 0.012)$. This stresses the importance of a good compromise between model quality and animation realism, specially for computational platforms with limited resources such as mobile phones. The mode value above 3 for the head motion question suggests that the participants tend to agree that they noticed the head motion in the avatars.

In general, the participants agree that they felt more like participating in the conversation when using the avatar interfaces. They are neutral about feeling more empathy with the chatting partner, while using the avatar interface than the nonavatar interface (i.e., the PS interface). Based on the mode values below 3 for the question about whether the avatars are distracting from the content of the conversation, the participants agree that the avatars are not distracting.

The participants are also neutral about using the text-driven avatar chatting application for everyday use. Regarding the kind of communication tasks that participants would prefer to use the avatar interface for, a paired sample Wilcoxon signed-rank test showed that the participants provide higher agreement ratings with respect to using the avatar interface for informal conversations (question *Informal*, $Mdn = 4$) as opposed to for formal (question *Formal*, $Mdn = 2$) ones $(W = 102, p = 0.0019)$

B. Poststudy Interview

A poststudy, free-form interview is conducted to the participants of our user study. They are basically asked to provide feedback about desirable and undesirable features of our text-driven avatar messaging interfaces, as well as to give their impressions on their overall user experience during the chatting session. We believe that the comments and suggestions obtained during these short interviews provide interesting insight for the design of applications using avatar interfaces for mobiles, and about user expectations for such applications. Enhanced avatar personalization is the most common suggestion to improve the user experience. Users expect to be able to modify the avatars themselves and imprint them with their own customization (e.g., hair style, face props such as glasses, jewelry, etc.). This seems to be an important feature in terms of entertainment and self-expression. The possibility of selecting features to caricaturize their own avatars, as well as the avatar of chatting partners, is also suggested. Regarding avatar speech, personalization of voice is a very desirable feature they mentioned. Users expect to hear a distinct voice for different chatting partners. Additionally, some of the participants expressed that support for common IM slang words in the avatar speech would add up to the attractiveness of the interfaces. The possibility of expressing emotions through their avatars is another common suggestion among all the participants to make the interface more enjoyable. They expect to be able to use the emoticons that they put in the text to control the emotions expressed by the avatar. Autonomous realistic avatar behavior is another desirable feature they suggested. Some of the participants expressed that they expect to observe some idle motion on the avatar between messages. For example, they suggested that the avatars should exhibit some head and gaze motions, as well as blinking, between received messages.

VI. LESSONS LEARNED AND RECOMMENDATIONS

We stress that the most important factors that we have determined for users to embrace the avatar interfaces are naturalism in lip-synchronization, the visual realism of avatars, and most importantly allowing users to personalize the used avatars themselves. Special attention should be payed to the quality of lip synchronization since this is the main factor influencing user perception on the realism of the animation. Natural motion for head and eye gazes should be provided. The display of mobile phones is, however, too small to fully notice minute details. Additionally, the computational resources on mobile phones are scarce and thus should be used sparely. Perception of head motion and that of gaze motion are correlated, and users most likely will perceive gaze motion, to a certain extent, from only the head movements.

Based on the scores given by the participants in the user study for questions *Formal* and *Informal,* we argue that the value given by users to the avatar interfaces will be focused on entertainment and fun activities. Although users look for visual realism in the avatars, their interest is more focused on being able to make them look like as they want them to be. Therefore, the possibility of personalizing avatars while still exhibiting some features resembling the appearance of the real person is an important feature to be included. This is specially significant for avatar interfaces used in socialization tasks. The same idea can be applied to the voice of the avatar as well.

Since constructing a synthetic personalized voice for every possible user is a daunting task, we recommend to allow the users to choose the avatar voices from a small voice database with a reasonable level of variety. Additionally, speech synthesis limitations due to misspellings seem to break the illusion of realism in avatar interfaces. A list of common misspellings, abbreviates, and Internet lingo should be handled by the system to avoid these nuances.

The subjective reports by the participants show that the illusion of realism was broken (low score for perceived realism) at some cases, which can be interpreted that the uncanny valley effect [15] showed up on some of our avatars. Interestingly, the participants also gave higher scores to the level of engagement and empathy on the PA interface. This suggests that even when the participants do not find the synthetic talking avatars to be completely realistic, their communication experience is positively affected by adding traits of realism to the avatars. Finally, since the expression of emotions is so important for the effective conveying of meaning and emotional context (the current study does not consider synthesis of emotion), enabling this capability should be of primary concern for future research efforts in this direction.

ACKNOWLEDGMENT

The authors would like to thank the RocketBox Libraries for providing the used high-quality 3-D avatars. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the agencies.

REFERENCES

[1] V. Bruce and A. Young, "Understanding face recognition," *Brit. J. Psychol.*, vol. 77, no. 3, pp. 305–327, 1986.

[2] J. N. Bailenson, A. C. Beall, J. M. Loomis, J. Blascovich, and M. Turk, "Transformed social interaction: Decoupling representation from behavior

and form in collaborative virtual environments," *Presence*, vol. 13, no. 4, pp. 428–441, 2004.

[3] N. Ducheneaut, M.-H. Wen, N. Yee, and G. Wadley, "Body and mind: A study of avatar personalization in three virtual worlds," in *Proc. 27th Int. Conf. Human Factors Comput. Syst.*, Boston, MA, USA, Apr. 2009, pp. 1151–1160.

[4] S.-H. Kang, J. H. Watt, and S. K. Ala, "Communicators' perceptions of social presence as a function of avatar realism in small display mobile communication devices," in *Proc. 41st Annu. Hawaii Int. Conf. Syst. Sci.*, Washington, DC, USA, Jan. 2008, p. 147.

[5] M. A. Smith, S. Farnham, and S. M. Drucker, "The social life of small graphical chat spaces," in *Proc. Int. Conf. Human Factors Comput. Syst.*, 2000, pp. 462–469.

[6] Z. Deng and U. Neumann, "Expressive speech animation synthesis with phoneme-level controls," *Comput. Graph. Forum*, vol. 27, no. 8, pp. 2096–2113, 2008.

[7] T. Erickson, N. S. Shami, W. A. Kellogg, and D. W. Levine, "Synchronous interaction among hundreds: An evaluation of a conference in an avatar-based virtual environment," in *Proc. Int. Conf. Human Factors Comput. Syst.*, Vancouver, BC, Canada, May 2011, pp. 503–512.

[8] B. Francesca and C. John, "Cursive: A novel interaction technique for controlling expressive avatar gesture," in *Proc. ACM Symp. User Interf. Softw. Technol.*, 2001, pp. 151–152.

[9] B. Timothy and M. Daniel, "Modalities for building relationships with handheld computer agents," in *Proc. ACM Conf. Human Factors Comput. Syst.*, 2006, vol. 2, pp. 544–549.

[10] W. L. Johnson, C. Labore, and Y. chun Chiu, "A pedagogical agent for psychosocial intervention on a handheld computer," in *Proc. AAAI Fall Symp. Health Dialog Syst.*, 2004, pp. 22–24.

[11] P. Persson, "Exms: an animated and avatar-based messaging system for expressive peer communication," in *Proc. Int. ACM SIGGROUP Conf. Support. Group Work (GROUP-03)*, Nov. 2003, pp. 31–39.

[12] S. H. Kang, J. H. Watt, and S. K. Ala, "Social copresence in anonymous social interactions using a mobile video telephone," in *Proc. Conf. Human Factors Comput. Syst.*, 2008, pp. 1535–1544.

[13] Z. Deng and U. Neumann, "eFASE: Expressive facial animation synthesis and editing with phoneme-isomap controls," in *Proc. ACM SIGGRAPH/Eurograph. Symp. Comput. Animat.*, Vienna, Austria, 2006, pp. 251–259.

[14] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Trans. Vis. Comput. Graph*, vol. 14, no. 1, pp. 213–230, Jan./Feb. 2008.

[15] T. Geller, "Overcoming the uncanny valley," *IEEE Comput. Graph. Appl.*, vol. 28, no. 4, pp. 11–17, Jul./Aug. 2008.