Interactive Multimodal Platform for Digital Signage

Helen V. Diez, Javier Barbadillo, Sara García,
Maria del Puy Carretero, Aitor Álvarez,
Jairo R. Sánchez, and David Oyarzun

Vicomteh-IK4, Paseo Mikeletegi 57, 20009 Donostia-San Sebastián, Spain
{hdiez,jbarbadillo,sgarcia,mcarretero,aalvarez,jrsanchez,doyarzun}@vicomtech.org
http://www.vicomtech.org

Abstract. The main objective of the platform presented in this paper is the integration of various modules into Web3D technology for Digital Signage systems. The innovation of the platform consists on the development and integration of the following technologies; 1) autonomous virtual character with natural behaviour, 2) text-to-speech synthesizer and voice recognition 3) gesture recognition. The integration of these technologies will enhance the user interface interaction and will improve the existing Digital Signage solutions offering a new way of marketing to engage the audience. The goal of this work is also to prove whether this new way of e-commerce may improve sales and customer fidelity.

Keywords: Multimodal Platform, User-Interface-Interaction, Digital Signage

1 Introduction

In the latest years, technology and especially new media has empowered marketing and commerce areas with new tools that go towards ubiquity and more and more faithful virtual representations of real products.

Nowadays, HTML5 and Web3D technologies are strongly pushing to web standardization of new media. Good and serious examples of efforts that are being done in this direction are the low level WebGL specification [1] and the high-level X3DOM architecture [2].

These new approaches could provide the basic platform for creating innovative marketing and e-commerce applications, which take advantage from potentiality of all technological channels and devices in a standardized way.

With this premise, the work presented in this paper consists on the development and integration of a web-based 3D engine and software modules that enable natural communication channels.

This technical work is built over three pillars:

- Coherent coexistence and communication among technologies coming from different disciplines and with different levels of maturity.

- Strong focus of usability, providing new interaction channels that make the human/computer communication more natural.
- Keep the message. That is, build technology that improves the way a message is transmitted to the user, not to condition the own message.

Therefore, these three pillars pretend to improve the channel related modules of the Shannon Weaver communication schema [3], as shown in Figure 1

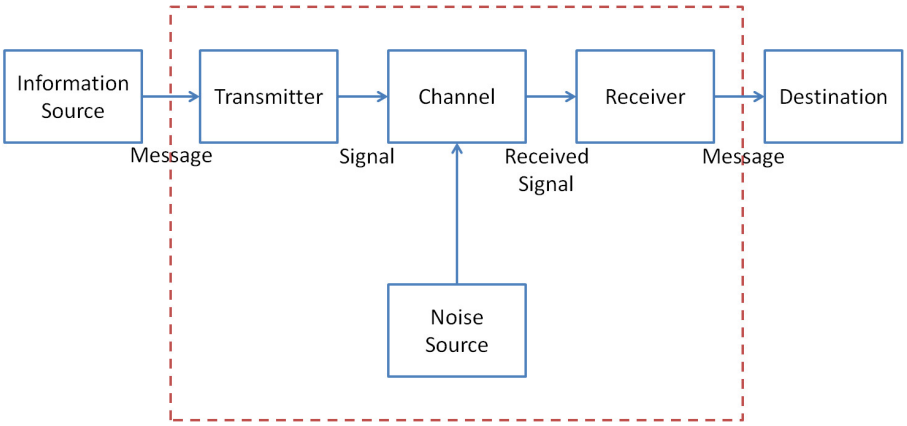

Fig. 1. Shannon Weaver communication schema on message transmission.

A realistic marketing use case has been built over this integration. The use case is designed as an interactive marketing platform to be shown in digital screens in public spaces. The platform allows the end-user to interact through natural channels, such as gestures and voice.

Therefore, the platform developed acts as a testbed to experiment interaction practices that maximizes the way a digital message is sent to the end-user. The use case is considered ideal when the end-user is a potential customer in this case and so, the importance of properly transmiting the digital message is even more critical.

Moreover, the standardized feature of the technologies developed keeps the coherence of the information when it is shown in any additional device, from PCs to smartphones. A modular design allows the content creator to abstract the digital message from the interaction channels, providing a platform that easily adapts to and takes advantage from the interaction capabilities of each device where it is running.

The paper shows the technical work carried out to get a stable version of the whole platform. Beyond the potential technical capabilities of these new technologies, the marketing use case is being validated in a controlled but real environment to check the usability in the marketing area during these months.

The paper is organized as follows; Section 2 analyzes the related work regarding digital signage and user categorization, Section 3 explains the architecture

followed to accomplish the goals of this work, describing in detail each of the modules involved. The final section is about conclusions and future work.

2 Related Work

Research on Human Computer Interaction (HCI) goes back to the 1980s [4], however the growig affordability of the devices using these interfaces and the accessibility to software development kits [5] have led to the evolution of HCI into Natural User Interfaces (NUI), this new way of interaction operates through intuitive actions related to natural, everyday human behaviour such as; touch screen interaction, gesture recognition, speech recognition or brain machine interfaces. These new ways of interaction have gained broad interest within the HCI community [6] and various experiments have been done to prove its benefits.

The use of Microsoft Kinect sensors has been crucial in many of these experiments due to the availability of its open-source and multi-platfom libraries that reduce the cost of algorithm development. Keane, S. et al. [7] present a survey on the Kinect sensor. A gesture recognition module based on the motion sensor included in the Kinect is used in [8] to improve user experience in the management of an office environment.

NUI is also being introduced into digital signage systems. Satho, I. [9] presents a framework for building and operating context-aware multimedia content on digital signage systems in public or private spaces and to demonstrate the utility of the framework, he presents a user-assistant that enables shopping with digital signage.

Chen, Q. et al. [10] describe a vision-based gesture recognition approach to interact with digital signage. Bauer, C. et al. [11] also introduce a conceptual framework for interactive digital signage which allows the development of various business strategies.

Adapting content according to the audience is one of the objectives pursued by the companies that offer digital signage. There are several studies that personalize content according to the audience. For example, Müller et al., [12] present a system that automatically learns the audience's preferences for certain content in different contexts and presents content accordingly.

Ravnic, R. and Solina, F. [13] developed a camera enhanced digital signage display that acquires audience measurement metrics with computer vision algorithms. The system also determines demographic metrics of gender and age groups. It was tested in a clothing boutique where the results showed that the average attention time is significantly higher when displaying the dynamic content as compared to the static content.

The introduction of autonomous characters into user interface platforms is also a matter of study. In 2006, Gribaudo, C. and Manfredi, G. [14] patented a modular digital assistant that detects user emotion and modifies its behaviour accordingly.

3 System Overview

During this work a 3D avatar able to interact with the user through different channels has been implemented. It has been designed using a modular schema that includes three main components: a web component, the speech component, and the gesture component.

The web component is responsible for displaying the virtual character along with the content of the signage application. It supports any browser that implements WebGL technology.

The speech component allows the user to interact with the virtual character using voice commands. It integrates speech recognition and synthesis technologies.

The gesture component integrates computer vision technologies for face and hands tracking. It allows the user to interact directly with the content using hand gestures, at the same time allowing the system to estimate the emotional state of the user through his face.

All the modules are integrated in a HTML5 compliant application that is used as the frontend of the signage system. However, part of the core of the speech and gesture components are native applications and must be executed in a desktop environment.

Following sections describe each module in detail.

3.1 Virtual Character

A main aspect of this work is the introduction of an autonomous virtual character into digital signage systems to act as a natural interface between the user and the content offered by the device. The role of the avatar will be to ease the communication between the audience and the digital information provided. Thus, users will experience a more natural and intuitive interaction emulating the one between real people. Likewise, the user can customize this interaction by accessing information according to his interests or preferences.

To achieve this virtual character with natural behaviour an animation engine based on WebGL technology as the one presented in this work [15] has been developed. This animation engine allows realistic simulation of both the avatar's face and body expressions. The engine is capable of real-time rendering of the lips when the avatar is speaking and it also interpolates the facial expressions depending on the avatar's mood.

As for the body language the avatar performs gestures and movements as humans do when communicating with others. A thorough study regarding natural hand, arm and body gestures has been done and animations emulating these movements have been designed.

Figure 2 represents the introduction of a virtual character with natural behaviour into a WebGL compatible browser. Nowadays most commonly used browsers support this technology (Firefox, Chrome, Opera, Safari).

Fig. 2. Integration of the virtual character into a WebGL compatible browser.

3.2 Speech Synthesis and Recognition

The platform includes technologies for both automatic speech recognition and speech synthesis.

Regarding speech recognition, the Google Speech Recognizer for Spanish was integrated adapting the publicly available java API to the needs of the platform. During the recognition process, the audio is collected from the microphone. It is then encoded to FLAC and passed via an HTTPS POST to the Google speech web-service, which responds with a JSON object with the transcription. The Google Speech Recognizer is speaker-independent and provides two language models to be used, based on (1) web searches for short phrases and (2) a generic language model for dictation. Considering the needs of the project, the generic language model was used to allow continuous speech recognition.

The integration with the web platform has been done using a regular text file. The recognition software writes into the file the transcription which is consumed by a script using long pooling techniques. This integration forces the component to be deployed on the same machine as the virtual character, but in the future it could be done using the new HTML5 standards for audio input.

Two possible solutions were included in the platform for speech synthesis in Spanish. Like for speech recognition, the Google Speech Synthesis was integrated for text-to-speech conversion. In this case, the text is sent to the servers of Google via an HTTP REQUEST and a speech file in MP3 format is returned through an HTTP RESPONSE. Since the Google Synthesizer is limited to a maximum of 100 characters, the API was modified to enable the platform to synthesize longer texts. For this purpose, the input text is previously splitted on the full stops. Each sentence is then synthesized and all the returned audios are concatenated in a unique WAV file at the end.

As an alternative to Google, the Microsoft Speech Synthesizer was integrated in the platform. This technology is provided through the Microsoft.Speech. Synthesis namespace, which contains classes that allow user to easily integrate functionalities for speech synthesis. In order to extend the voicebank of the platform, a module for voice transformation was also included. This module transforms the synthesized voices modifying some prosodic features like the fundamental frequency, the speech rhythm and the energy. As a result, this module is able to modify the source speakers speech to make it sound like that of a different speaker.

3.3 Gesture component

The platform implements a method for detecting and tracking the user's facial emotions and hand gestures. The system is composed by a Kinect device which captures video and a depth map, and it also includes a face detector for emotion recognition. The goal of the gesture component is to allow the interaction of the user with the avatar in both directions, resulting in a more natural experience. The avatar behaves according to the user's emotions and the user can perform gestures to communicate with the avatar.

The Kinect device captures video with an integrated camera and a depth map using infrared sensors. With the help of OpenNI and Nite APIs the system is able to detect and track human body parts and perform gesture recognition. Our system first detects the human body and then gets the head and the right and left hand positions. The 3D position is projected to 2D screen coordinates and the distance of the user with respect to the camera is obtained. This way the interaction is restricted to users that are facing the camera and close enough to it, avoiding interaction with people passing by.

In order to perform gesture recognition the user's hands are segmented from the rest of the body and tracked. The Nite API allows to track and detect the click gesture, the waving gesture and the rising hand gesture. The coordinates of the hand are also converted to screen coordinates so the user can use the hand as a mouse for selecting or clicking objects in a screen.

For the emotion detection process the face of the user is detected in combination with the head detection of the Kinect and a probabilistic face detector. First, the 2D position of the head is obtained from Kinect. If the distance to the camera is close enough the probabilistic detector is applied. Finally the face is detected and tracked and the system performs the emotion detection.

Our implementation of the emotion detection is based on the method proposed by [16]. When a face is detected in the screen a facial point mask is fitted to the face. This is achieved by first detecting facial features based on local image gradient analysis and then adjusting a deformable 3D face model to those features in the 2D plane. The mask represents the main facial features of a human face an it is able to track facial deformations computational efficiently and under challenging light conditions.

The emotion recognition method is based on the Facial Action Coding System developed in [17]. Every component of a facial movement is represented by an

Action Unit (AU) and therefore every facial expresion can be decomposed into AUs. An AU is independent of any interpretation as they are the result of the contraction or relaxation of one or more muscles. In our program an AU is represented by the movement of a point of the facial mask. For example, the hapiness expression is detected if the threshold of the AUs "cheek raiser" and "lip corner puller" is exceeded. Using the facial point mask makes it trivial to measure AUs and detect if a facial emotion is being performed. A threshold is set to skip low intensity muscle actions. Although there are up to 100 AUs our system just measures a few AUs related to the seven universal emotions: fear, surprise, sadness, anger, happiness, disgust and contempt.

Finallly the system filters the detected emotions to reduce them to three emotions of interest for our application: the user can be interested, neutral or not interested. The avatar will behave differently depending on the emotions recognized on the user.

To avoid sending massive information to the avatar controller, the gestures and emotions are filtered over the frames to generate statistics that are sent every certain number of frames.

The integration has been done in the same way as the speech component. In this case the component is deployed with an executable that writes the gesture and face information in a text file. The web component reads the file using long pooling techniques.

4 Conclusions and Future Work

New multimedia sources are those that blend computer technology, with the audiovisual and telecommunications technology. They support a given language formed by image, sound, voice, written text, gestures and expressions and reach the user in a single marketer message.

Digitization is the universal tool that is profoundly transforming the international markets. This has eased the creation of new forms of marketing and industries dependent on these modes of information. The evolution of technology has changed the environment. The technological society of most developed countries live with new modes of experimental communication based on interactivity and the development of new forms of interdisciplinarity.

These digital technologies have completely changed the way information is transmitted. Due to its interactivity, the medium becomes the message itself. The multimedia models create a new, more powerful way to inform. New audiences are segmented and differentiated by gender, age and other components and the message focuses on this fact. These audiences are very selective with the message they receive because of their multiplicity.

New languages of human-machine communication are generated. Digital communication refers to the use of technology to achieve a certain purpose. The digital format of the medium indicates that the message content has something different and innovative. The format of the content is dependent on the medium, the distributor and the transmission system. The representation by digital screens

in public spaces, allows private conversations and public environments simultaneously. It is a message created for the public, because the public participates by their expressions in the creation of the message. It is based on real-time interaction with an availability of 24 hours a day.

The transmitter, in this case the avatar, is in the same physical space as the receiver, so there is no distancing. A first approach takes place, it does not depend on the receiver who is surrounded by technology, but it depends on a transmitter / avatar searching for his receiver in their natural environment and a receiver that unintentionally becomes such.The figure of transmitter and receiver are constantly exchanged.

The message is not predetermined, it is dynamic, it is being created as the feedback happens. The receiver gives meaning to the message that has been sent massive / selectively. The audience tends to choose their messages, which improves the effectiveness of them. The range of possibilities the content displayed to the user is based on their characteristics and their willingness to receipt of the message, plus the response obtained along the communication between the avatar and the client.

Electronic technologies have a greater impact on the audience than more traditional media, since this support is also the means of reaching people and does not need more intermediaries. This way the message reaches out to the viewer, and the screen becomes the scenery for the reception.

The system has been tested in an academic atmosphere and it has proved to work recognizing and categorizing each user correctly. However, as future work we are planning to set the platform on a real scenario at a public space. This validation will serve the purpose of examining whether personalized marketing as the one proposed by our platform is better than traditional marketing systems.

To validate our system we will perform the following experiment; two groups of volunteers covering various age and gender ranges will be created. These volunteers will be invited to enter a mall in which the two trial systems have been set. Each group will try only one of the systems, once they have tried their corresponding system they will fill in a survey. In each corridor of the mall one of the marketing device systems will be set (Figure 3). Both systems will be monitored by Kinects in order to gather information regarding the amount of time people spend in front of the device. After the volunteers have walked through the corridors they will be invited to fill in a survey with questions such as:

1. did you stop in front of the device?
2. what struck your attention?
3. why did you leave?
4. did you find the experience amusing?
5. did you enjoy talking to the avatar?
6. did you find the communication with the avatar natural?
7. have you entered the store/s proposed by the system?
8. did you buy anything in any of the stores proposed by the system?
9. how would you improve the system?

The conclusions drawn from these surveys will direct further investigations in this field.

As mentioned in section 3.2 we also plan to standardize the voice capture from the microphone using WebRTC API [18], this communications standard developed by the W3C enables the embedding of audio and video in applications and websites. The WebRTC standard solves incompatibilities in real-time communications between browsers. This will also allow to integrate the Kinect device and video processing with HTML, which is currently handled by the Gesture Module plugin.

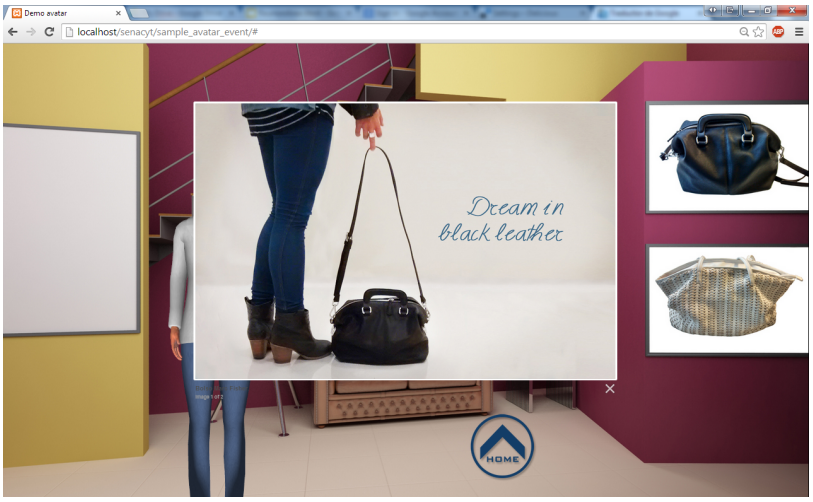

Fig. 3. Interactive Multimodal Platform for Digital Signage.

References

1. WebGL Specification, http://www.khronos.org/registry/webgl/specs/latest/1.0/ retrieved on March 2014.
2. X3DOM Specification, http://www.x3dom.org/x3dom/doc/spec/ retrieved on March 2014.
3. Weaver, W. Recent contributions to the mathematical theory of communication. In C.E. Shannon and W. Weaver (Eds.), The mathematical theory of communication, pages 128. 1949.
4. Myers, B. A. A brief history of human-computer interaction technology. interactions, pp. 44-54, 1998.
5. Goth, G. Brave nui world. Commun. ACM, pp. 14-16, 2011.
6. Seow, S. C., Wixon, D., Morrison, A., and Jacucci, G. Natural user interfaces: the prospect and challenge of touch and gestural computing. In CHI'10 Extended Abstracts on Human Factors in Computing Systems (pp. 4453-4456). ACM. 2010.
7. Keane, S., Hall, J., and Perry, P. Meet the Kinect: An Introduction to Programming Natural User Interfaces. 2011.

8. Re, G. L., Morana, M., and Ortolani, M. Improving user experience via motion sensors in an ambient intelligence scenario. 2013.
9. Satoh, I. A framework for context-aware digital signage. In Active Media Technology. Springer Berlin Heidelberg. pp. 251-262, 2011.
10. Chen, Q., Malric, F., Zhang, Y., Abid, M., Cordeiro, A., Petriu, E. M., and Georganas, N. D. Interacting with digital signage using hand gestures. In Image Analysis and Recognition. Springer Berlin Heidelberg. pp. 347-358. 2009.
11. Bauer, C., Dohmen, P., and Strausss, C. Interactive Digital Signage-An Innovative Service and Its Future Strategies. In Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference. IEEE. pp. 137-142. 2011.
12. Müller, Jörg, et al. "Reflectivesigns: Digital signs that adapt to audience attention." Pervasive computing. Springer Berlin Heidelberg, pp. 17-24. 2009
13. Ravnik, R., and Solina, F. "Audience measurement of digital signage: Quantitative study in real-world environment using computer vision." Interacting with Computers 25.3 pp. 218-228. 2013.
14. Gribaudo, Claudio; Manfredi, Giorgio. Virtual Assistant With Real-Time Emotions. U.S. Patent Application 11/617,150, 28 Dic. 2006.
15. Diez, Helen V., Sara Garca, Jairo R. Snchez, and Maria del Puy Carretero. "3D animated agent for tutoring based on WebGL." In Proceedings of the 18th International Conference on 3D Web Technology, pp. 129-134. ACM, 2013.
16. Unzueta, Luis, Waldir Pimenta, Jon Goenetxea, Lus Paulo Santos, and Fadi Dornaika. "Efficient generic face model fitting to images and videos." Image and Vision Computing 32, no. 5, pp. 321-334. 2014.
17. Ekman, P., Freisen, W.V. and Ancoli, S. "Facial signs of emotional experience." Journal of Personality and Social Psychology, 39(6), pp.1125-1134. 1980.
18. Web Real-Time Communications Working Group, et al. WebRTC 1.0: Real-time Communication Between Browsers. 2012. http://dev. w3. org/2011/webrtc/editor/webrtc. html, 2012.