

# A MULTIMODAL 3D HEALTHCARE COMMUNICATION SYSTEM

*Cem Keskin\**, *Koray Balci†\**, *Oya Aran\**, *Bülent Sankur\*\**, *Lale Akarun\**

Bogazici University\*  
Dept. of  
Computer Engineering  
Bebek, 34342, Istanbul  
Turkey

Bogazici University\*\*  
Dept. of  
Electrical Engineering  
Bebek, 34342, Istanbul  
Turkey

ITC-irst†  
Intelligent Interactive Information  
Presentation Group  
Via Sommarive, 18, 38050, Trento  
Italy

## ABSTRACT

We present a system that integrates gesture recognition and 3D talking head technologies for a patient communication application at a hospital or healthcare setting for supporting patients treated in bed. As a multimodal user interface, we get the input from patients using hand gestures and provide feedback by using a 3D talking avatar.

*Index Terms*— gesture recognition, multimodal user interfaces, 3D facial animation

## 1. INTRODUCTION

Embodied Conversational Agents (ECA) have improved human-computer interaction by providing a sympathetic face and human voice to communication between computers and humans. An ECA is a human-like character embedded in the user interface, mimicking interpersonal communication and creating an illusion of a social environment with the users. It has been observed that users tend to anthropomorphize the ECAs, and in general it is believed that ECAs improve the communication quality and user satisfaction. While anthropomorphism in human-like interfaces is still a long-term debate [1, 2], it has been reported that the presence of an ECA has a positive effect [3]. Substantial evidence shows that well designed human-like interfaces are able to properly model communicative functions and improve usability of computer systems [4, 5], especially in the areas of tutoring systems and edutainment.

On the other hand, gesture-based communication presents a convenient means of interaction with computers. Many people are unable to use a keyboard because of a handicap, or simply because the environment does not permit it. Voice based communication is an alternative. However, some handicaps or environments do not accommodate voice, either. The situation of a bedridden patient is a typical example: The environment does not easily accommodate a keyboard; and the patient is often too

weak to use a keyboard or talk intelligently. 3D gesture based communication via a gesture recognition system presents a convenient solution.

In this study, we present our system built for a hospital or healthcare setting, where patients can communicate with doctors and nurses using a set of multimodal interface modules. Patients can be elderly or hearing impaired. Instead of pressing a button and calling for help, in our typical scenario, patients can perform a set of hand gestures, which are automatically recognized by our system and sent to a particular workstation controlled by a nurse. The patient computer uses an ECA to give feedback to the patient and to inform that the message is communicated to healthcare personnel. On the nurse terminal, the message of the patient can be presented by an ECA as visual speech, or in plain text to let the nurse handle multiple patient requests at the same time. According to the gesture, the nurse takes the correct action, one of which is sending back a message to the patient letting her know her problem is being taken care of. This message is sent back to patient's computer over the network and an ECA acts according to the contents of the message.

In section 2, we present the details of our gesture recognition system, and in Section 3, we describe the ECA implementation. In Section 4, we describe the data collected for this study and report the recognition results. Section 5 describes system integration and we conclude in Section 6.

## 2. GESTURE RECOGNITION

The system is intended for patients in hospitals or healthcare centers, who are possibly unable to speak, but are able to move their hands. This scenario imposes certain constraints on the gesture recognition module. Such gestures, as performed by the patients, might not have the same speeds or scale, and may be performed anywhere on the scene. The camera setup intended for the gesture recognition module may also differ in every hospital or room. Furthermore the default gestures may not be suitable for some patients, or a new gesture may be needed for a different setting.

Therefore, the gesture recognition module should be able to handle the spatio-temporal variability of the gestures without depending on special hardware; the camera setup should be configurable, and the system should allow definition and further training of gestures.

In a previous study, we have developed STARS (Sign Tracking and Recognition System), a 3D hand gesture interface for generic applications (Figure 1). A preliminary version of the system that used discrete hidden Markov models (HMM) to recognize trajectories was presented in [6, 7]. The HMM-based interface allowed real-time recognition of 3D hand gestures, online Kalman filtering, online training and gesture spotting. The main shortcoming of the system for this scenario was its inability to incorporate hand shape information. Therefore we have developed a continuous hidden Markov model based version of STARS that is capable of recognizing complex hand gestures that involve both hand shape and hand motion information.

STARS is developed as an interface for generic applications, where the hand motion replaces the mouse, and communicative gestures can be trained to control target applications. The system is capable of spotting known gestures in continuous video streams, and hence, the patients are relieved from having to initiate the recognizer before performing a gesture. The system automatically segments the gestures, discards the meaningless hand movements and generates an event upon successful recognition of a gesture.

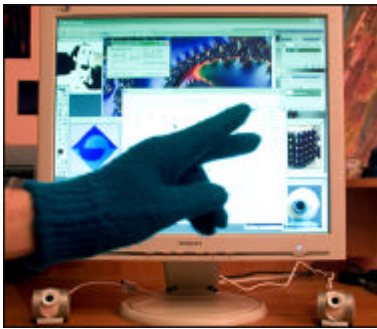


Figure 1 STARS Setup with two cameras

Hand gestures are modeled by STARS, using the spherical angles of the velocity vector of the hand, the angle of the hand image, and the seven Hu moments for each view, which add up to a total of 18 parameters for each frame pair of the stereo camera stream. Instead of modeling the entire information using a single Gaussian distribution, we separate the hand motion and hand shape parameters by modeling each set with a different multivariate Gaussian function. In order to incorporate the information coming from different dimensionalities, we normalize the values of the Gaussian functions by scaling them according to their maximum values. The weighted sum of the scaled values is used as the

observation probabilities of the local states of the gesture models. The experimental results of using this scheme are explained in Section 4.

STARS has the capability of triggering certain events upon recognition of known gestures. The types of events range from mouse events and keyboard shortcuts to sequences of such inputs. Hence, integration with other applications is straightforward. In our scenario, upon a successful recognition of a gesture, a local face animator client responds to the patient to confirm the gestural request, and a remote face animator client processes the request and informs the intended party. The integration of STARS with the face animation system is explained in Section 5.

### 3. 3D FACIAL ANIMATION

Synthetic characters are often integrated in multimodal interfaces as an effective modality to convey messages to the user. They provide visual feedback and engage the user in the dialogue through emotional involvement. However, avatars should not be considered as an individual modality but as the synergic contribution of different communication channels that, properly synchronized, generate an overall communication performance: characters can emit voice and sounds, animate speech with lips and facial expressions, move eyes, head and body parts to realize gestures, express emotions, perform actions, sign a message for a deaf companion, or display listening or thinking postures .

SMIL-Agent scripting language [8] provides all the above modalities required for an ECA implementation with clear definitions of distinctive communicative channels. Here, gestures recognized by the STARS system are mapped to stock SMIL-Agent scripts and sent to our ECA over the LAN where the information is transformed into synthesized speech and facial animation.

Here is a sample SMIL-Agent script that informs the patient and performs a sentence on diagnostics where the first part is articulated in a sad emotion and the final part in happy mood with added eye and head movements.

```
<par system-language="english">
  <speech channel="alice-voice" affect="sorry-for"
    type="inform" id="angina">
    <mark id="*1*" />
    You have been diagnosed as suffering from
    <mark id="*2*" />
    angina pectoris, which appears to be mild.
  </speech>
  <seq channel="alice-face" >
    <speech-animation affect="Sad" content-
      id="angina" content-end="*2*" />
    <speech-animation affect="Happy" content-
      id="angina" content-begin="*2*" />
  </seq>
  <action channel="alice-eyes" action-
    type="turning" intensity="0.5" content-
```

```

id="angina" content-begin="*1*" content-
end="*2*"
<parameter>LookLeft</parameter>
</action>
<action channel="alice-head" action-
type="pointing" content-id="angina"
content-begin="*2*">
<parameter>15</parameter>
<parameter>0</parameter>
<parameter>5</parameter>
</action>
</par>

```

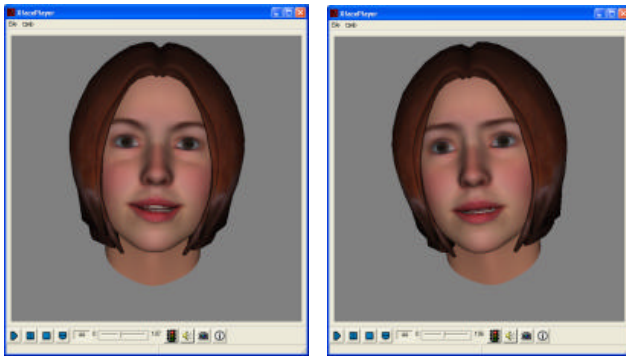


Figure 2 XfacePlayer Snapshot

In this study, as an ECA, we use our open source 3D facial animation toolkit, Xface [9, 10] for providing feedback to patients. With Xface toolkit, one can author 3D avatars that support MPEG-4 Facial Animation (FA) [11] standard and keyframe interpolation using SMIL-Agent scripting language. As XfacePlayer (Figure 2) natively provides control over LAN using TCP/IP, we can conduct our experiments with little effort. For speech synthesis, Xface relies on Microsoft SAPI 5.1 for this study, although one can easily integrate other TTS (text-to-speech) engines to the toolkit.

#### 4. DATA COLLECTION AND RECOGNITION RESULTS

Our system is based on receiving gesture commands from the patient and conveying the inherent message to the responsible person or system. We identified nine gestural commands and their respective actions. The commands are integrated to the system through the gesture recognition module, STARS. The command is recognized and passed to the 3-D facial animation module and then the respective action is taken.

Table 1 shows the gestural commands and the response of our system. Although the gestures are inspired from Turkish Sign Language (TSL), the selected signs are symbolic and intuitive; thus easy to learn and imitate for any patient, and not necessarily a TSL signer (Figure 3).



Figure 3 The gesture for "I am thirsty"

For each command, we collected 30-40 examples from two subjects. For better hand segmentation, we used colored gloves. The videos are recorded with two webcams with 320x240 resolution and 30 frames per second. The two webcams are necessary for high recognition accuracy since the commands are performed in 3D. Following the 3D reconstruction and feature extraction steps, HMM models are trained for each command as described in Section 2.

Table 1 Patient's gestural commands and system actions

Gesture Command	Xface Response
<i>hungry</i>	Patient is hungry
<i>thirsty</i>	Patient is thirsty
<i>bath</i>	Patient needs to go to the bathroom
<i>pain</i>	Patient has pain
<i>cold</i>	Patient is cold – Room is cold
<i>hot</i>	Patient is hot – Room is hot
<i>doctor</i>	Patient is calling the doctor
<i>nurse</i>	Patient is calling the nurse
<i>help / emergency</i>	An emergency situation

Table 2 Recognition Rates

Gesture	Hungry	Pain	Bath	Doctor	Nurse	Hot	Thirsty	Cold	Help	Rate %
1	36	0	0	0	0	0	0	0	0	100
2	1	18	0	0	1	0	24	1	0	40
3	0	0	29	0	0	0	9	0	0	76.3
4	0	0	0	40	0	0	0	0	0	100
5	0	0	0	3	33	0	0	0	0	91.7
26	0	0	0	0	0	40	0	0	0	100
7	0	0	0	0	0	0	27	3	0	90
8	0	0	0	0	0	0	0	35	10	77.8
9	1	0	0	5	0	0	0	0	39	86.7
<b>Total:</b>										<b>84.7</b>

Table 2 lists the results of the recognition tests for the gesture module using continuous HMMs. It should be noted that the particular gesture for ‘Pain’ has a low recognition rate. This is due to the similarity of both hand motion and hand shape information of the gestures ‘Pain’ and ‘Thirsty’. Nevertheless, it is straightforward to distinguish these gestures using high level heuristics, such as the position of the hand relative to the face of the observer. By using similar heuristics the recognition rates can be significantly enhanced for similar gestures.

## 5. SYSTEM INTEGRATION

Xface implements a server-client architecture that allows communication over the network. A client application can send SMIL-Agent scripts to the Xface running as a server. These scripts are processed by the server to generate the corresponding animation. The requests made to the server are queued until the client sends a ‘PLAY’ message to the server, which is processed when the server is idle and able to playback the animation. Finally, the server sends back a notification message for confirmation. This architecture can be easily adapted for our scenario.

The gestural requests need to be processed immediately, since the party to whom the message is to be sent is not known beforehand. Therefore, an instance of STARS application and an Xface client have to run locally, along with an Xface server that will inform the patients about their requests. Also each party that may receive requests should have an instance of an Xface server.

STARS processes the gestures it detects and sends a message to the local Xface client, which directly sends a confirmation script to the local server and a request message to the remote Xface server of the intended party. When the remote server sends the notification message back to the client, another confirmation script is sent to the local server to inform the patient.

## 6. CONCLUSION

In this paper, we present a multimodal communication application framework integrating gesture recognition and 3D talking head avatars designed for healthcare settings. We have run recognition tests with gestures selected for an example scenario. In future studies, we would like to test the effectiveness of this system in a healthcare environment plan to test the efficiency of the system in a hospital setting. We also plan to implement Input/Output HMMs [12] for improving gesture recognition.

## 7. REFERENCES

- [1] Shneiderman B., *Designing the user interface. Strategies for effective human-computer interaction.* A-W, Reading, 1998.
- [2] Reeves B, and Nass C., *The media equation: how people treat computers, television, and new media like real people and places.* Cambridge University Press, 1996.
- [3] E de Vos, *Look at that Doggy in my Windows, on effects of anthropomorphism in human-agent interaction*, Doctoral Thesis, Utrecht University, 2002
- [4] J. Cassell, T. Bickmore, H. Vilhjálmsdóttir and H. Yan, “More than just a pretty face: affordances of embodiment,” *In Proceedings of the 5th ICIUI*, New Orleans, 2000.
- [5] R. Rickenberg, B. Reeves. “The effects of animated characters on anxiety, task performance, and evaluations of user interfaces.” *In Proceedings of CHI*. 2000
- [6] C. Keskin, A.N. Erkan, L. Akarun, “Real time hand tracking and 3D gesture recognition for interactive interfaces using HMM”, *Proceedings ICANN/ICONIP*, Istanbul, 2003.
- [7] C. Keskin, O. Aran, L. Akarun, “Real time gestural interface for generic applications”, *European Signal Processing Conference, EUSIPCO Demonstration Session*, Antalya, 2005.
- [8] E. Not, K. Balci, F. Pianesi, and M. Zancanaro., “Synthetic characters as multichannel interfaces,” in *Proc. ICMI05*, Italy, October 2005.
- [9] K. Balci, “Xface: MPEG-4 based open source toolkit for 3d facial animation,” *In Proc. Advance Visual Interfaces*, Italy, 2004.
- [10] K. Balci, “Xface: Open Source Toolkit for Creating 3D Faces of an Embodied Conversational Agent,” *In Proc Smartgraphics*, Germany, 2005.
- [11] Pandzic, I. and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*, Wiley, New York, 2002.
- [12] Y. Bengio, P. Frasconi, “Input/Output HMMs for sequence processing”, *IEEE Transactions on Neural Networks* vol. 7(5), pp. 1231–1249, 1996.