

Gestures Interactions for Immersive Communications

Emmanuel Marilly, Olivier Martinot, Arnaud Gonguet

Alcatel-Lucent Bell Labs France, Route de Villejust, 91620 Nozay - France
{Emmanuel.Marilly;Olivier.Martinot;Arnaud.Gonguet}@alcatel-lucent.com

ABSTRACT

Video meeting and presentation systems are more and more widely used. Yet interacting or controlling such video systems is far from easy. Immersive communication aims to enable better interactions between end users and video systems. In this paper we propose an end-to-end study on gesture interactions for a video presentation system starting from the algorithms (i.e. algorithms enabling posture recognition) to the user evaluation (i.e. evaluation by users of gestural interactions).

Index Terms— Immersive Communication, Gesture, Posture, Recognition, Interactions, User Evaluation.

1. INTRODUCTION

One of the key challenges of the telecommunication industry is to identify the future of communication. Video is identified for long as the future of telecommunication, however use only pure video and audio will not deeply change the current mode of telecommunication. Therefore, immersive communication has been defined as the way to exploit video and multimedia technologies in order to create new relevant and valuable usages. Regarding the immersive communication aspect, our first focus of interest is the enterprise context where immersion can provide a better experience for meeting at distance, presentation at distance. Our approach can solve current issues of existing solutions such as the travel cost for face to face meeting or the cost for current tele-presence solutions [1].

The “Presentation at Distance” use case was chosen for its capability to illustrate numerous facets of the immersive communication model we are building. Among all the issues we are facing when presenting at distance, the gesture is an important one that is not yet exploited today. Gestures can be a solution for natural interactions solving either pointing issues, control/command issues or contents interaction issues.

Although humans can identify and recognize gestures easily, the implementation of an automatic approach performing these tasks is a challenge due to many constraints and the wide semantic gap [3].

In addition to common problems (e.g. luminosity, background complexity), the hand gesture recognition process has to differentiate the unintentional hand movements from the other hand gestures. A gestural taxonomy which explains the difference between all classes of hand gestures can be found in [4]. Besides, due to the unlimited number of all possible hand gestures, find the most significant gestures that can be used to interact with a video meeting system can be considered as a difficult task and need to put the user in the gesture selection loop.

2. WHAT IS IMMERSIVE COMMUNICATION

Immersion is usually defined by all the technical capabilities to mimic sensorial feelings: it can be the CAVE (Cave Automatic Virtual Environment) [18], the 3D technologies, the Virtual World, the spatial audio or haptic system [17]. Most of the time, the focus is done on the capability of the immersion to be as realistic as possible. This approach is called sensorial immersion.

But in a context where the objective is to improve distant communications, sensorial immersion is not enough. Because communication is made of social interaction, narration, task driven activities, we need to include a new aspect for immersion: attentional immersion. Attentional immersion concerns the cognitive experience to be immersed in a narration, in a task or in a social interaction. Attentional immersion may rely on sensorial immersion to help keeping attention. Nevertheless anyone experienced being immersed in a phone call, or reading a book, thus with few sensorial features: attention mechanisms are not necessary linked to sensorial aspects.

3. A FLAGSHIP USE CASE: PRESENTATION AT DISTANCE

“Presentation at distance” use case has been chosen as it illustrates the two aspects of the immersive communication: displayed contents for the sensorial immersion, and message, story around the contents, as well as social interaction, for the attentional immersion.

Issues encountered when using existing systems for distant presentations have been identified and categorized. For the

remote audience, key issues are the static layout of the video capture leading to boring feeling, the difficulty to have good quality both for the presenter and the slides rendering, the bad audio quality, the lack of capabilities to track interesting topics, or to get key points from a presentation during the presentation itself and also after the presentation. For the presenter, key issues are the technical difficulty to setup a session, the lack of feedback from the remote audience, the difficulty to interact with the contents (i.e. pointing, manipulation and navigation). This paper is focusing on the ways to manage this latter point through hand gestures.

4. HAND GESTURE RECOGNITION

Hand gesture recognition process is a challenging problem due to the deformable nature of the hand [5]. In hand gestures analysis, two concepts have to be distinguished: hand posture and hand gesture. The first one is defined as a static hand gesture which does not change during a period of time. The second one is a dynamic hand movement which can be defined as a temporal trajectory of some estimated parameter over time [6] or as a sequence of hand postures [5]. Although invasive hand gesture recognition techniques provide very accurate results, they tend to impose a burden to many users and they are not common in our daily life. On the other hand, vision-based hand gesture recognition algorithms, which are less intrusive, can be split in two categories: the 3D hand model-based approaches and the appearance-based approaches [4]. The appearance-based approaches extract directly the 2D images features from the video stream to model the appearance of the hand. For hand postures recognition, the appearance-based approaches refers mostly to a pattern recognition problem. Several approaches have been used for posture and gesture recognition such as neural networks [7], elastic graph [8], and statistical approaches [9], HMMs [10].

Taking into account the advantage and disadvantage of each approach and the requirements of our specific application, an appearance-based approach has been developed. The proposed approach is based on 7 processing steps that can be regrouped in 2 main functionalities: hand posture processing and hand gesture processing [19].

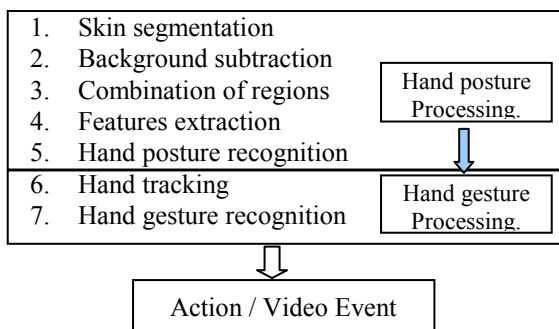


Fig. 1: Gesture & posture processing steps

4.1 Hand posture processing

Firstly, the hand region is extracted using color and foreground information. For the skin segmentation, the RGB (Red, Green, Blue) chromaticity space for skin segmentation [9] has been used. The Approximate median background subtraction method [11] has been used for the background subtraction. The background model is updated every 10 frames.

The hand region is extracted using the combination of the two previous steps. First of all, the holes are filled in the binary image resulted from the skin segmentation. Then the connected components are extracted. For each connected component its intersection with the foreground detected is calculated using the logical operator “And”. Finally, only components which the area of intersection is higher than a predefined threshold are kept. Figure 2 gives an overview of this first processing step.

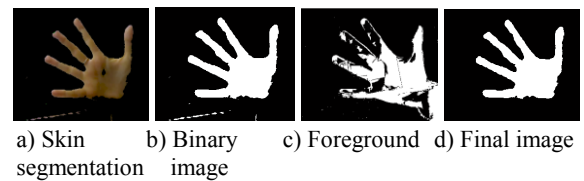


Fig.2: Combination of regions

Secondly, feature vector is estimated for the posture recognition which contains different characteristics of the hand. The feature vector, used for the posture classification, is composed of statistical features (Zernike Moments [12]) and geometrical features (circularity and rectangularity).

Finally, the classification of postures is performed using the Principal Components Analysis [6]. Hand postures recognition is based on the Euclidian distance between the calculated vector and the vectors constituting the “learning set” in the new PCA representation.

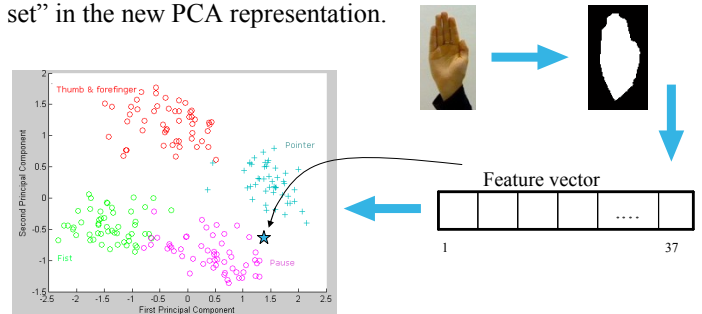


Fig.3: PCA of 205 vectors split in 4 posture sets

The Figure 3 presents the classification of 205 vectors in the 4 selected sets of postures (pointer, thumb & forefinger, palm, clenched fingers). Each vector corresponds to one of the 4 postures sets taken in various conditions of luminance, hand position, hand rotation, and depth. The different sets are clearly differentiated.

4.2 Hand gesture processing

Hand gesture processing starts from previous results. The proposed approach consists in combining a signal similarity process with data mining tools.

To know the trajectory of the hand over a period of time, a Kalman filter [13] is used to track the centroid of hand.

To avoid the confusion between the unintentional hand movements and the explicit hand gestures, the recording of the hand trajectory starts when the hand posture “Record” is recognized. The recorded hand trajectory, which is performed over a period of 15 frames, is compared to pre-defined trajectories models using the intercorrelation function. There is a maximum when significant similarity exists between the two signals.

The data mining “SIPINA” method [2] has been used to determine the threshold value of the intercorrelation function that defines the limit between the non-recognition and recognition.

4.3 Results & Evaluation

Figure 4 presents respectively the selected hand postures and the selected hand gestures with their associated actions.

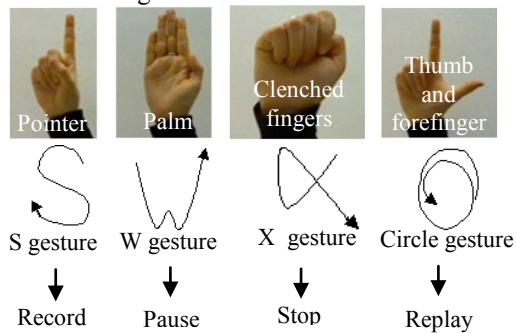


Fig. 4: Mapping Hand gestures-postures / actions

The study was done using a set of 156 different samples composed of: 34 trajectories for the gesture “record”, 40 trajectories for the gesture “pause”, 42 trajectories for the gesture “replay” and 40 trajectories for the gesture “stop”. The proposed approach for posture and gesture recognition has been evaluated and tested on a set of 240 postures (60 samples per class) and a set of 104 gestures (26 samples per class). For postures, the global error rate for the recognition is 8%. For gestures, the global error rate for recognition is 15%.

As the objective is to provide more immersive communication tools, the challenging evaluation that we have to do now, is the evaluation of the technology from the user point of view. In our study, this technology will be evaluated in the context of a video recording system. The user evaluation will provide information on the efficiency of the gesture interactions with the system and how the users perceive this approach. The provided feedbacks will give tracks for improvements.

5. USER EVALUATION

The objective of the user evaluation is to answer three questions:

- Are the proposed postures and gestures easy enough to remember, and is it hard for the users to execute them?
- Which interaction mode (postures versus gestures) is the most efficient to use and/or the most popular among the subjects?
- Which is the most efficient action-trigger mapping solution?

The evaluation of the proposed interactions used a standard usability testing methods [14] to assess the utility, efficiency and satisfaction of the users. More specifically, as we considered the users’ feedbacks were valuable, we collected verbal data using Nielsen’s formative evaluation methodology [15]. The objective is to help designing the next version of the interactions to better fit the user needs. In a perspective of qualitative study, the evaluation will be done on a sample of 10 persons. In a next step, this number will be increased for quantitative study. However, for a pre-test, a sample of 10 enables to extract the main lines and trends.

5.1. Methodology

5.1.1. Environment & material

The materials used for the evaluation was:

- A MORAE activity recorder system [16] to record user behavioral data and to provide instructions to the participants during the tests.
- The video recorder mockup.
- Two booklets used as assistive tools.

5.1.2. Metrics

We monitored the time needed to execute the correct posture/gesture that triggers a specific action, between the instruction given by the experimenter and the effective motion done by the user. It gives the information on the level of affordance of each posture/gesture and the time needed to learn how to perform it properly. We controlled visually the proper execution of the subject’s postures/gestures, in order to detect if bad recognition of a posture/gesture was due to a system’s malfunction or was attributable to subject’s incorrect behavior.

5.2. Protocol & Results

5.2.1. Sample description

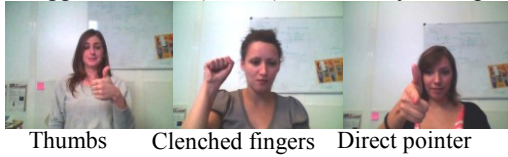
Our sample is distributed as follows: Six females and four males. The average age is of 25.8 years, and the average level of education is Master degree. Every subject has already experienced the use of tactile interfaces, and seven of them are using such interfaces every day. Nine of the ten subjects have experienced the use of gesture-based interfaces.

5.2.2. Intuition test

This first test did not require any software. We presented to the participants a sequence of four actions to do. Without more details, we asked them to trigger the given actions by trying the postures and gestures that came to their mind intuitively. When a subject had done up to ten postures/gestures, or when he thought that he did not had more ideas, we asked him to move on to the next action, without telling him the “correct” posture, thus without influencing his next trials.

Intuitively, the main proposed gestures (example for record & pause) were the following:

- To trigger the *REC* (record) function by hand postures.



- To trigger the *PAUSE* function by hand postures.



For the intuitive gestures test, the general observation is that none of the proposed intuitive gestures was corresponding to any of the four predefined gestures proposed by the video recorder system. Statistically, the mapping of the most popular postures on each function gives table 1:

Function	REC	PAUSE	STOP	REPLAY
Thumbs up	3			
Clenched fingers	3		2	
Palm		3	3/4	
Planar palm		2		

Table 1. Mapping of the most popular postures on each function

5.2.3. Innate preference test

We presented the participants the two booklets (gesture booklet & posture booklet), and asked them to complete the first part of the satisfaction questionnaire. The goal was to collect the subjects’ feeling about using the postures and gestures, with respect to their mental representations solely, without any kind of real use experience.

Posture	Pointer	Palm	Clenched fingers	Thumb and forefinger
Average score	2,7	4,5	3,6	2,5

Table 2. Average innate satisfaction scores of hand postures.

With respect to the postures proposed by the video recording mockup, the users’ innate satisfaction gives table 2.

Significant differences of satisfaction between postures were observed. The *palm* posture is by far the most appreciated, followed by the *clenched fingers*. In overall, the subjects didn’t like very much the *pointer* nor the *thumb and forefinger* postures, mostly because they didn’t mean anything to them as metaphors.

With respect to the postures proposed by the video recording mockup, the users’ innate satisfaction gives table 3.

Gesture	S	W	X	Circle
Average score	3,1	2,1	2,8	3,3

Table 3. Average innate satisfaction scores of hand gestures.

Significant differences of satisfaction between gestures were observed. The *circle* gesture is the most appreciated one, with half the subjects “*linking it*”. The “*W*” gesture is the least appreciated, with four subjects who “*don’t like it at all*”, and three subjects who “*don’t like it*”.

Majority of the subjects found the semantic links between the letters and their associated functions “*inadequately chosen*”, if not “*confusing*”. This was particularly pointed toward the “*S*” gesture, which would more intuitively mean “*Stop*” than “*Start recording*”.

Without predefined context, half of the subjects thought that they would prefer using gestures. In a context of the video recorder mockup, half of the subjects thought they would prefer using postures. Some of them were “*afraid of confusing the gestures between them*” and thus found the postures “*more reliable*”, while other ones thought that the postures would be “*more intuitive*”, “*easier to remember*” and “*easily recognizable*”. Two subjects wanted a mixed interface combining gestures and postures to control the video recorder.

5.2.4. Real use test

The purpose of this test was to immerse the subjects into a real-use experience, in order to submit them the same satisfaction questionnaire as for the innate preference test. We asked the subjects to run a sequence of actions: (Start; Wait; Pause; Resume; Stop; Replay). Finally, after running this sequence, we asked to the participants to define their preferred action-trigger mappings for both hand postures and hand gestures.

The satisfaction study of postures usage done after the use of the mockup gives the following results (table 4):

Posture	Pointer	Palm	Clenched fingers	Thumb and forefinger
Average score	2,7	4,1	3,3	2,8

Table 4. Average real satisfaction scores of hand postures.

Significant differences of satisfaction between postures were observed. The *palm* posture remains the most appreciated, still followed by the *clenched fingers*.

For gestures, the pos-test satisfaction study gives table 5.

Gesture	S	W	X	Circle
Average score	1,6	1,2	1,4	1,3

Table 5. Average real satisfaction scores of hand gestures.

There is no significant difference of satisfaction between gestures. None of the gestures were appreciated.

Users had better satisfaction for posture than gesture. This conclusion has to be moderated taking into account the experimental bias mentioned above (poor performances of the gesture recognition due to the mockup conditions of use).

6. CONCLUSIONS & RESEARCH PERSPECTIVES

Postures are currently the far most efficient way to trigger actions. After using the current version of the video system, every user prefers the postures. According to the sorting done by the users and to the intuition test, the most consensual action-trigger mapping is shown table 6.





Actions	Postures	Gestures
 REC	Thumbs up	Grabbing
 PAUSE	Palm	Advancing palm
 STOP	Clenched fingers	"S"
 REPLAY	Thumb and forefinger	Circle

Table 6. Most consensual action-trigger mapping

The innate preference for postures or gestures is dependant of the context of use and of the user's personal feelings. If the context of use requires fast actions, postures are considered as most efficient, and are then preferred. However, if the context of use requires controlling dynamic objects as a timeline, gestures are considered to offer the most intuitive interaction mechanism. In future works, the gesture interactions should take into account the difficulty for users to move the whole hand instead of drawing gestures with their forefinger.

Therefore, the next steps will consist in improving the dynamic gesture recognition process by modifying the gesture models (i.e. whole hand) and providing more robust tracking approach that can estimate the state and parameters of non linear systems in order to take into account a wide range of possible gestures. It will be done by applying active curve modeling with "new forces" for gesture recognition.

7. REFERENCES

- [1] O. G. Stadt, M. H. Gross, A. Kunz, M. Meier. "The Blue-C: Integrating Real Humans into a Networked Immersive Environment", CVE 2000, Third International Conference on Collaborative Virtual Environments.
- [2] Zighed D., Rakotomalala R., Graphes d'Induction: Apprentissage et Data Mining, Hermès, 2000.
- [3] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, No 12, pp.1349-1380, December 2000.
- [4] V. I. Pavlovic, R. Sharma and T. S. Huang, "Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review", Pattern analysis and machine intelligence, vol. 19, No. 7, pp. 677-695, July 1997
- [5] Qing Chen, "Real-Time Vision-Based Hand Tracking and Gesture Recognition", PhD Thesis, University of Ottawa, 2008.
- [6] Principal Component Analysis, I.T. Jolliffe, Springer-verlag, 1 janvier 2002
- [7] J.J. Stephan, S. Khudayer, "Gesture Recognition for Human-Computer Interaction (HCI)", International Journal of Advancements in Computing Technology, Vol. 2, N° 4, Oct. 2010
- [8] J. Triesch, C. Von der Malsburg, "Robust Classification of Hand Postures against complex Backgrounds", Proceedings of the Second International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, 14-16 Oct 1996.
- [9] J. Kovac, P. Peer and F. Solina, "2D versus 3D colourspace face detection", 4th EURASIP Conference on Video/Image Processing and Multimedia Communications, Croatia, pp. 449-454, 2003.
- [10] H Haberdar and S. Albayrak, "Real Time Isolated Turkish Sign Language Recognition from Video Using Hidden Markov Models with Global Features", ISCIS 2005, LNCS 3733, pp. 677 – 687, 2005.
- [11] S. Cheung and C. Kamath, "Robust techniques for background subtraction in urban traffic video" Electronic Imaging 2004, San Jose, California, January 20-22 2004.
- [12] Chee-Way Chong., P. Raveendranb, R. Mukundanc, "Translation invariants of Zernike moments", Pattern Recognition, pp. 1765-1773
- [13] G. Welch and G. Bishop, "An Introduction to the Kalman Filter", SIGGRAPH 2001.
- [14] Dumas J. S., Redish J. C., A Practical Guide To Usability Testing, Intellect Boobs, 1999.
- [15] Nielsen J., Usability Engineering, Morgan Kaufmann, 1993.
- [16] <http://www.techsmith.com/morae.asp>
- [17] S. J. Gibbs, C. Arapis and C. J. Breiteneder. "TELEPORT-Towards immersive copresence", Multimedia Systems - MMS, vol. 7, no. 3, pp. 214-221, 1999
- [18] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti. "Surround-screen projection-based virtual reality: the design and implementation of the CAVE", SIGGRAPH, pp. 135-142, 1993
- [19] N. Fourati, E. Marilly, "Gestures for natural interaction with video", San Francisco, Electronic Imaging 2012, Jan. 2012, Proceedings of SPIE Vol. 8305.