16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

COMPARISON OF TWO DIFFERENT SIMILAR SPEECH AND GESTURES MULTIMODAL INTERFACES

Alexey Karpov[1], Sebastien Carbini[2], Andrey Ronzhin[1], and Jean Emmanuel Viallet[3]

[1]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
SPIIRAS, 39, 14-th line, 199178, St. Petersburg, Russia
[2]Ifremer/LASAA, Technopôle Brest-Iroise, BP 70, 29280 Plouzané, France
[3]Orange Labs France Telecom, 2 avenue Pierre Marzin, BP40, 22307 Lannion, France
{karpov,ronzhin}@iias.spb.su; sebastien.carbini@ifremer.fr; jeanemmanuel.viallet@orange-ftgroup.com

ABSTRACT

The paper presents two different multimodal interfaces based on automatic recognition and interpretation of speech and gestures of user's head and hands, developed within the framework of the SIMILAR European Network of Excellence. The architectures of ICANDO and MOWGLI multimodal interfaces, modalities recognition, information synchronization and fusion as well as qualitative comparison and quantitative evaluation using Fitt's law experiments are described. The comparison of both contactless interfaces shows that in spite of the differences in computer vision and ASR techniques applied, they provide similar performances in contactless human-computer interaction.

1. INTRODUCTION

Mouse and keyboard are the most common input interfaces of human-computer interaction within the WIMP (Window, Icon, Menu, Pointing) paradigm. However during last decade, scientists and engineers have been actively looking for and developing more intuitive, natural and ergonomic interfaces. Of particular interest are multimodal interfaces that use speech and gestures to create human-machine interfaces similar to human-human communication. To organize effective and natural HCI, such interfaces should interpret not only manual gestures, but also head, hands, full-body gestures and eyes gaze.

Both speech and vision-based gesture recognition allow to build contactless interaction devices that offer the advantage of distant interaction without having to be in physical contact with an input device to interact. Our goal is to build multimodal interfaces that benefit from the advantages of both modalities to interact as easily and naturally as possible using a representation of the world with which we are familiar, where objects are described by their name and attributes and locations indicated with hand gesture or facial gesture in case of hands disabilities.

Two SIMILAR multimodal interfaces, developed by the Russian and French teams independently, are presented in the paper. They both provide contactless HCI with speech and gestures. ICANDO hands-free user interface combines speech commands with head gestures, and MOWGLI interface proposes interaction with a large display either with speech and hand gestures or with both hands gestures.

2. ICANDO MULTIMODAL INTERFACE

ICANDO (Intellectual Computer AssistaNt for Disabled Operators) multimodal user interface is intended mainly for assistance to persons without hands or with disabilities of their hands or arms, but it could be useful for contactless human-computer interaction by ordinary users too. A user can manipulate a pointer by moving his/her head and giving speech commands instead of using standard input devices. ICANDO combines the module for automatic recognition of voice commands in English, French and Russian as well as the head tracking module in one multimodal interface (Figure 1). The system processes human's speech and head motions in parallel and then fuses both informational streams in a joint multimodal command for operating with GUI. Each of the modalities transmits its own semantic information: head position indicates the coordinates of the pointer, while speech signal transmits the information about the meaning of the action, which must be performed with an object selected with the cursor or irrespective of the pointer position.

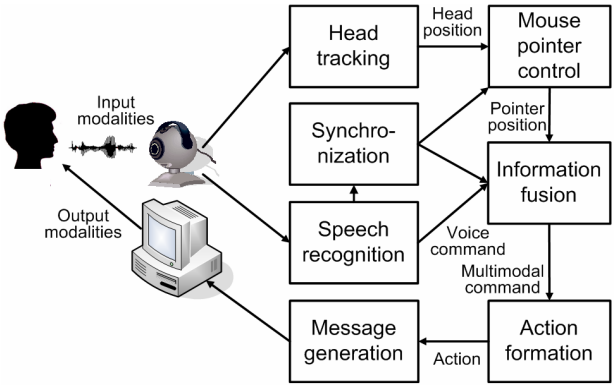

Figure 1 - The architecture of the ICANDO multimodal interface

ICANDO is positioned as a low-cost multimodal user interface available for most potential users so a standard

web-camera with a price under 50 € is employed. A USB web-camera Logitech QuickCam for Notebooks Pro was used for the experiments; it captures both video in 640x480x25fps mode and audio signals in 16 KHz sampling rate and mono format with acceptable SNR level via the built-in microphone.

SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) speech recognition engine is applied to recognize user's voice commands. For the speech parameterization MFCCs with 1st and 2nd derivatives are used. The modeling and recognition of phonemes and words are based on HMMs. The system is trained to understand 30 voice commands of 3 languages in speaker-independent mode. All the voice commands can be divided into four classes according to their functional purpose: pointer manipulation commands ("Left (button click)", "Double click", "Right (button) down", "Scroll up", etc.), keyboard manipulation commands ("Enter", "Escape", etc.), MS Windows GUI commands ("Start", "Next", "Previous", etc.), as well as the special command "Calibration" in order to start the process of the head tracker tuning. However, the commands for the pointer manipulations have a multimodal nature only and they need the information on the coordinates of the pointer cursor.

ICANDO employs optical flow video processing for tracking natural operator's head motions instead of hand-controlling motions. The system was trained for vision-based tracking of 5 natural facial points: the center of the upper lip, the tip of nose, the point between the eyebrows, the left eye (iris) and the right eye points. A pointer controlled by a user's nose was proposed earlier [1]; however the system of 5 points used improves robustness of face tracking. The tracking method [2] uses the iterative Lucas-Kanade algorithm for the optical flow processing.

The synchronization of two information streams is performed by the speech recognition module, which sends messages for saving the pointer coordinates, calculated by the head tracking module as well as for multimodal information fusion. Coordinates saving has to be done at the moment of triggering an algorithm for speech endpoint detection..Figure 2 illustrates the process of modalities synchronization and information fusion. This figure shows a fragment of fulfillment of the test scenario for hands-free work with MS IE for obtaining and copying some text fragment from Internet.

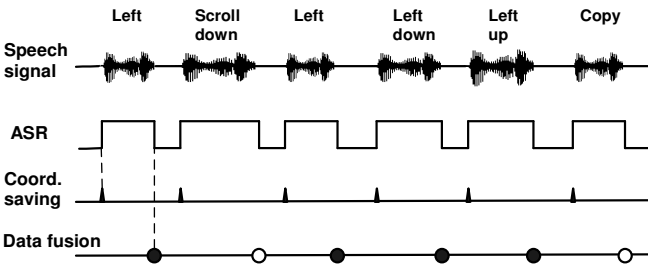

Figure 2 - Mechanism of modalities synchronization and fusion

A black circle on the figure means that a recognized command is multimodal (for instance, "Left down") and a white circle denotes that a command has a unimodal speech-only nature (for instance, "Copy"). ASR module operates in the real-time mode (recognition delay is less than 0.05xRT); since the vocabulary of voice commands is small, there are minor delays between an utterance and fulfillment of the recognized multimodal command and these delays may be excluded from the consideration. If a speech command has the multimodal nature (pointer manipulation commands) then it has to be combined with saved coordinates of the pointer and then a message to the mouse virtual device is sent. If a voice command is unimodal, the coordinates are not taken into account and a message to the keyboard device is posted.

3. MOWGLI MULTIMODAL INTERFACE

MOWGLI (Multimodal Oral With Gesture Large display Interface) provides a system allowing two users to simultaneously interact with a very large display, thus allowing the users to collaborate, being aware of what the other user is doing, sharing a common view [3], and interacting with the computer using the same speech and gesture modalities as in everyday life human-human communication.

Within the multimodal MOWGLI system (Figure 3), gesture recognition and speech recognition results are fused and interpreted by MOWGLI depending on the application context. Pointing and selection are involved in most applications, whereas commands and context are application dependent. Therefore speech vocabulary for addressing commands and context fusion has to be trimmed for each application. The MOWGLI system has been used for different kind of applications including a chess game application, virtual manipulation of 3D objects and navigation [4].

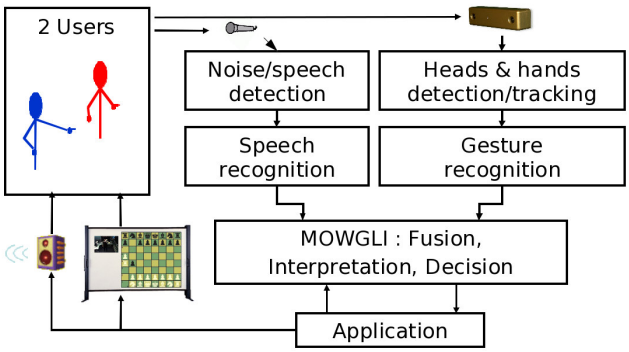

Figure 3 - Architecture of the MOWGLI multimodal interface

The automatic speech recognition uses HMMs and the recognized sentences syntax is described in a grammar. The vocabulary, used by the chess game, is made of some 50 speech items (for example, "take the pawn", "move the queen here", etc.). Each word is obtained by concatenation of context dependent phonetic units. The system outputs the n-best results and is speaker independent.

The pointer is controlled by automatic gesture recognition based on body parts tracking. Detection and tracking of body parts rely on skin-color, disparity and movement information. Skin color is obtained by a broad filter look-up table obtained on different users and under different lighting conditions. Obtained from the two images of a stereo camera, the disparity gives the 3D position of a pixel with respect to the camera coordinates. A filtered disparity image is computed for each user for which the observations too far away (> 1.3m) from the head center are discarded (once the head has been detected by a neural network). The movement is obtained by background subtraction which is updated every image so that a still person quickly fades in the background. Once a head is detected, its 3D position is used to define the body space involve in the detection of hands and to detect the pointing purposiveness of the user.

Biometric constraints limit the search space to a centroid centered on the face. Furthermore, it is reasonable to admit that, when interacting with the display, the user moves its dominant hand towards the display and sufficiently away from its face (more than 30 cm). Thus the hand search space is restricted to a volume delimited by a sphere and a plane, a volume called the 'action area' (Fig. 4, left). Behind this plane, lies the rest area which allows the user not continuously interacting. The action area allows a user expressing its intentionality to interact through gesture or speech. Owing to morphological constraints, the hand non detection area cannot be reach by a hand. These areas are face referenced, so that their absolute positions change as a user moves but the system behavior remains the same. The first detected hand control the pointer whereas selection is triggered either by second hand or by speech. Both hands functions are enable only when the corresponding hand is in action area (Fig. 4, right).

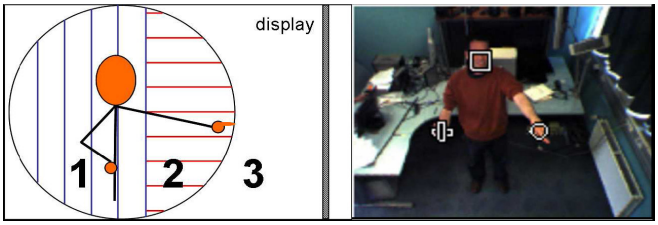

Figure 4 - Left: Schematic user related-space with 1-vertical stripes: user rest area, 2-horizontal stripes: user action area, 3-beyond sphere: hand non detection area. Right: a view of a user with both hands in the action area (rectangle: head, circle: pointing hand, cross: second hand).

Pointing gesture is a permanent process (20 Hz). Gesture selection event is a discrete event occurring on one frame, whereas speech is an event with a beginning and an end. Synchronization is triggered by a speech event or a gesture selection event. In the case of action triggered by speech, we make the hypothesis that speech and gesture are temporally aligned and we consider the pointing position that occurs during speech. The mean pointing position between time $[t_b,t_e]$ with $t_b = t_{bs} - 240$ ms and $t_e = t_{es} + 240$ ms, with t_{bs} and t_{es} the beginning of speech and end of speech times. The 240 ms is the delay used to isolate speech from silence. Indeed it has been experimentally observed that users first stabilize their pointing process before uttering and that they usually wait for the answer of the ASR before beginning pointing to another location.
Gesture or speech synchronization allows to establish a reference for the cursor (and thus the object selected behind the cursor) but also for hands reference position for subsequent bi-manual gesture. Thus, fusion occurs between gestures or between speech and gesture.

Oral commands which do not need to be complemented with a gesture information are taken into account only if performed together with a pointing gesture anywhere on the screen in order to make the difference between an effective oral command addressed to the machine and an oral comment, addressed to a third party, that would be similar to an oral command.

4. INTERFACES EVALUATION AND COMPARISON

A summary of both interfaces is presented in Table 1 which shows a qualitative comparison of 12 parameters for ICANDO and MOWGLI multimodal interfaces. It can be seen from the table that the systems have many differences; however the purpose of both systems is the same: contactless human-computer interaction.

Table 1 - Qualitative comparison of ICANDO and MOWGLI

Parameter	ICANDO	MOWGLI
Target group of users	hand-disabled persons	ordinary users
Number of users	single user	one or two users
Kind of interaction	Human-computer interaction	HCI, human-human interaction
Hardware equipment	low cost web camera	high cost stereo camera
Speech recognition	small vocabulary	small vocabulary
Languages	English, French, Russian	French
Gesture recognition	2D vision-based	3D vision-based
Recognition objects	head and facial objects	hands and head
Operating area	short distance interaction	long distance interaction
Synchronization kind	by speech modality	by speech or gesture
Context awareness	no support	in chess game
Main applications	assistive systems for hands-free HCI, games	edutainment, cooperative HCI, games, virtual reality

Quantitative evaluation of the multimodal input devices was carried out based on Fitts' law experiments and related works [5, 6]. The evaluation of performances of ICANDO and MOWGLI was independently made by Russian and French teams, correspondingly. During the experiments the same software providing Fitt's law experiments was used by both groups. The experiments with MOWGLI device were carried out with 10 adults (7 males and 3 females). Among

the 10 persons, the two French authors of the paper have a long practice of the interface, two have a two day practice, and six beginners had no practice with these devices. For the full gesture and speech and gesture MOWGLI devices, subjects stood up, behind a line two meters away from the wall display 2.21x1.66 meters with a 1280x1024 pixels resolution. ICANDO was evaluated with 6 adult testers including two of the authors who can be considered as experienced users. Working with ICANDO interface the subjects seated at a table about 0.5 meters away from the 17" notebook screen. Prior to the experience, subjects are shown a short demonstration of the task to be performed. Then, the subjects are allowed a short training period for contactless devices, instructed to hit the target as quickly as possible (in order to comply with Fitts' law hypothesis). The users were also instructed to point and to select 16 different targets, with a circular layout so movements are carried out in different directions. When selection occurs, the former target is shadowed and the next target is displayed.

The experiments with several pairs of targets' width-distance, corresponding to different indexes of difficulty were carried out by each subject. The index of difficulty *ID* of the task, measured in bit, is given, according to the Shannon formulation, by $ID = \log_2\left(D/W + 1\right)$, where *D* is the distance between targets and *W* is the target's width. *ID* range is 1.32-4.4 for ICANDO and contact-based devices and 1.32-4.64 for MOWGLI (Table 2). However the location where selection occurs influences both effective distance and effective width. The effective index of difficulty is $IDe = \log_2\left(De/We + 1\right)$, where *We* is the effective target width, given by $We = 4.133\sigma$, where σ is the standard deviation of the coordinates of the point of selection, projected on the axis between the centers of the origin and destination targets, according to [5]. *De* is the effective distance between the first and last point of an inter-target trajectory. The obtained *IDe* differ from *ID* sought for, greater IDe being obtained for the ICANDO contactless device and smaller IDe both for the MOWGLI with gesture selection (GS) of targets and for MOWGLI with speech selection SS. Figure 5 shows the data averaged over all the testers. A data above (below) curve of slope 1 indicates that an accomplished task is harder (easier) than expected.

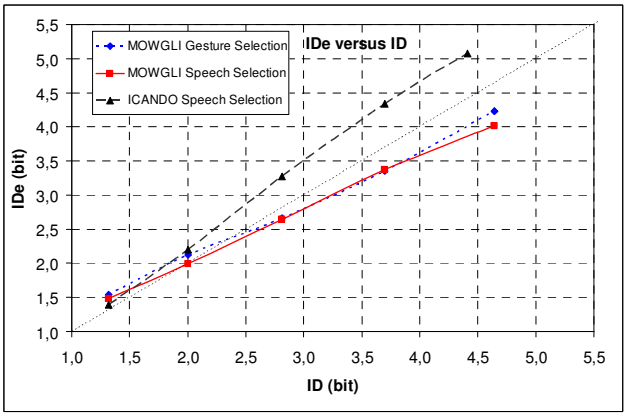

Figure 5 - *IDe* as a function of *ID* for the contactless devices

Fitts' law states that the movement time *MT* between two targets is a linear function of *ID* of the task related to the targets' characteristics. Figure 6 shows the movement time *MT* values for MOWGLI and ICANDO contactless input devices. For each trial, the inter-target movement time is defined and measured as the time between two successive selection events, whether selection occurs inside or outside a target (and counted as a selection error). Movement times for all the users are given in box plot in Figure 6. Boxes indicate the first and third quartile. Bars above and below a box indicate the 90th and the 10th percentile, and bars in the middle mark the medians. It can be seen that more time is needed to hit a target by ICANDO.

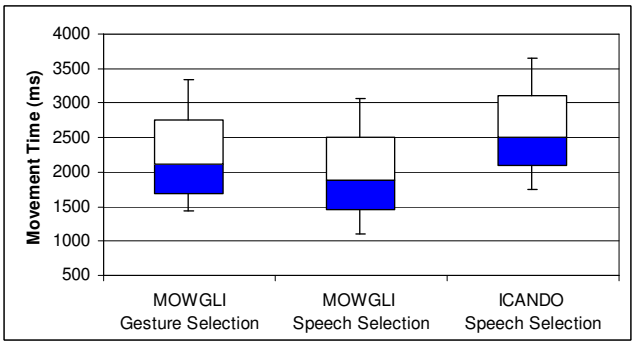

Figure 6 - Movement time *MT* values for MOWGLI and ICANDO

In contrast to Figure 6, Figure 7 shows *MT* versus *IDe* independently for beginners (6 for MOWGLI and 4 for ICANDO) and experienced users (2 for both systems). The clear difference between experts and beginners demonstrates significance of a learning effect using new pointing devices. Also one can notice a high scatter of the data for all devices and scatter is higher for beginners and the data slope for the experienced participants is smaller than the one of beginners.

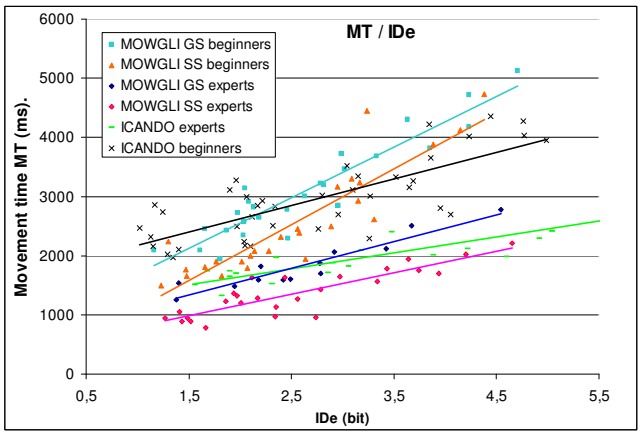

Figure 7 - *MT*/*IDe* with linear estimates for beginners and experts

Effective throughput *TPe* is the ratio between *IDe* and *MT* measured in bits per second. Mean *TPe* (and *TP*) allows comparison between the performance of different devices, performances determined by different studies with different participants. Throughputs for the contactless devices are

given in box plot in Figure 8. It can be seen that MOWGLI and ICANDO have comparable throughput values.

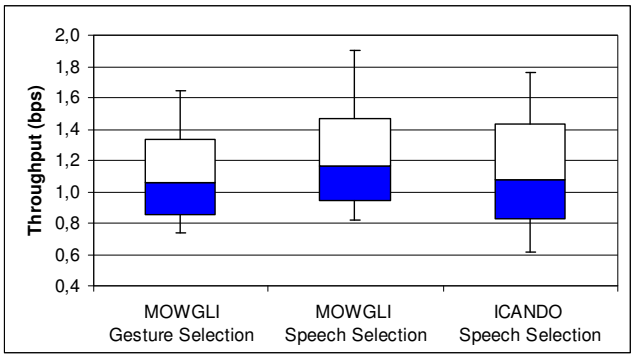

Figure 8 - Effective throughputs *TPe* of the contactless devices

Standard contact-based input devices such as mouse, touchpad, trackball, joystick and 17" touchscreen were also tested to compare their performances with those of the proposed contactless devices. Table 2 shows averaged values of movement time *MT* and effective throughput *TPe*, which shows a tradeoff between pointing speed and target selection precision.

Table 2 - Quantitative comparison between ICANDO and MOWGLI Gesture Selection and Speech Selection interfaces

Input device	W (pixel)	D (pixel)	MT (s)	Selection Error (%)	TPe (bit/s)
Joystick	32-128	96-650	2.17	7.53	1.49
Trackball			1.12	4.07	3.36
Touchpad			0.98	4.36	3.55
Mouse			0.44	3.03	6.50
Touchscreen			0.48	6.30	7.42
ICANDO (2 experts)			1.98	7.33	1.59
ICANDO (4 beginners)			2.95	12.11	0.91
ICANDO (all 6 testers)			2.63	10.51	1.14
MOWGLI GS (2 experts)	32-256	96-768	1.66	9.18	1.48
MOWGLI GS (6 beginners)			2.69	17.08	0.95
MOWGLI GS (all 10 testers)			2.32	14.57	1.12
MOWGLI SS (2 experts)			1.43	12.55	1.84
MOWGLI SS (6 beginners)			2.69	19.32	1.03
MOWGLI SS (all 10 testers)			2.03	17.79	1.26

The best *TPe* results were obtained with a contact-based touchscreen and a mouse device, taking into account that a touchscreen is not so precise for small *W*. Fitts' law results obtained with MOWGLI and ICANDO interfaces are similar, moreover it was found that speech&gesture interaction is faster than HCI by two hands. In some experiments *TPe* for speech&gesture devices outperformed a Thrustmaster joystick used as a mouse pointer and contactless devices have performance quite close to trackball and touchpad pointing devices. However the key advantage of the devices developed is that they provide a natural contactless human-computer interface that is similar to human-human communication.

5. CONCLUSION

Two different contactless multimodal speech and gesture interfaces were presented. They were compared by Fitt's law experiments. Although ICANDO and MOWGLI interfaces function in very different conditions and rely on different computer vision and speech recognition techniques, they have rather similar performances, suggesting that performances may be limited on one hand by the low frequency rate of cameras involved in computer vision processes and on the other hand by time scale involved in human speech. MOWGLI device was better in the pointing speed, whereas ICANDO has shown more precision of target selection. An importance of a learning process for using contactless devices was demonstrated. The results obtained allow us to conclude that both speech&gesture devices have outperformed both contactless HCI with both hands and standard joystick device (for trained users in contactless HCI) according to the effective throughput parameter, having acceptable target selection error.

6. ACKNOWLEDGEMENTS

This work has been supported by the EU-funded SIMILAR Network of Excellence (project # 507609) and by the Russian Foundation for Basic Research (project # 07-07-00073).

REFERENCES

[1] D. Gorodnichy and G. Roth, "Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces," *Image and Vision Computing*, Vol. 22, pp. 931–942, 2004.

[2] A. Karpov and A. Ronzhin, "ICANDO: Low Cost Multimodal Interface for Hand Disabled People," *Journal on Multimodal User Interfaces*, Vol. 1, No. 2, pp. 21–29, 2007.

[3] M. Morris, A. Huang, A. Paepcke and T. Winograd, "Cooperative Gestures: Multi-User Gestural Interactions for Co-located Groupware," in *Proc. CHI 2006 Conference,* Montreal, Canada, 2006, pp. 1201–1210.

[4] S. Carbini, L. Delphin-Poulat, L. Perron and J.E. Viallet, "From a Wizard of Oz Experiment to a Real Time Speech and Gesture Multimodal Interface," *Signal Processing*, Vol. 86, No. 12, pp. 3559–3577, 2006.

[5] R.W. Soukoreff and I.S. MacKenzie, "Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI," *Int. Journal of Human Computer Studies*, Vol. 61, No. 6, pp. 751–789, 2004.

[6] G.C. De Silva, M.J. Lyons, S. Kawato and N. Tetsutani, "Human Factors Evaluation of a Vision-Based Facial Gesture Interface," in *Proc. CVPRHCI IEEE Workshop*, Madison, USA, 2003.