International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.8 (2015), pp.125-138
http://dx.doi.org/10.14257/ijmue.2015.10.8.13

A Dynamic Interactive Pooling Based Human-Computer Interaction

Yulong Li[1], Yepeng Guan[1, 2, *] and Jinfeng Ma[1]

1. School of Communication and Information Engineering, Shanghai University
2. Key Laboratory of Advanced Displays and System Application, Ministry of Education
ypguan@shu.edu.cn (**Corresponding author*)

Abstract

Human-computer interaction (HCI) has great potential for applications in many fields. A HCI method has been developed based on audio-visual perceptive features and dynamic interactive pooling. A mode analysis based fusion strategy is proposed to fuse the interactive responses from audio-visual modalities especially for the case there are deviations from the modalities at the same time. The developed approach can be applied with a better performance even in single modal with multiple interactive users in the scenario. The users in the scenario are endowed with a fair chance to perform HCI by their interactive priorities no matter how many modalities used, and the interactive authority and priority are updated dynamically according to the variation of HCI content. The scenario and multiple users altogether are regarded as a dynamic HCI pooling. The diversity of interactive habits among multiple users is considered in an ordinary hardware from a crowded scene without any hypothesis for the scenario contents in advance. The pooling is updated dynamically to meet the user demands with the most intentions preferentially. Comparative study with state-of-the-arts has indicated the superior performance of the proposed method.

Keywords: *Human-computer interaction; Dynamic interactive pooling; Audio-visual perceptive features; Interactive response; Multi-modality*

1. Introduction

HCI has been one of the hot topics in artificial intelligence and computer vision fields due to its great potential for interactive applications [1]. There are many HCI modalities including facial expression, body posture, hand gesture, speech recognition and so on [2-3]. Among these modalities, hand gesture is intuitive and easy to be learnt. Elmezain *et al.* [4] proposed hand gesture spotting and recognition simultaneously based on hidden Markov models. Suk *et al.* [5] proposed a HCI method based on hand gestures in a continuous video stream using a dynamic Bayesian network model. Van den Bergh *et al.* [6] developed a hand gesture interaction approach based on time-of-flight. Due to the diversity of hand gesture, there exist some limitations for recognition based HCI. Comparatively, pointing gesture, one of the simplest hand gestures which can be easier to be recognized, has been attracted much attentions. Park and Lee [7] presented a 3D pointing gesture recognition based on a cascade hidden Markov model and a particle filter for interaction with mobile robots. Kehl and Gool [8] proposed a multi-view method to estimate 3D directions of one or both arms. Michael *et al.* [9] set up two orthogonal cameras to detect hand regions, tracked the finger pointing features, and estimated the pointing directions in 3D space. Pointing gesture recognition based HCI methods mentioned above uses the results of face and hand tracking to recognize the pointing direction. However, the HCI is limited by the unreliable face and hand detection. Another

ISSN: 1975-0080 IJMUE

difficult problem is how to recognize some small pointing gestures which usually results in the wrong direction. To overcome the problem, Pan and Guan [10] proposed a HCI way based on an adaptive virtual touch screen which is constructed instead of pointing direction estimation.

Since people naturally interact with the world multi-modally [2-3] both in parallel and sequential multiple perceptual modalities, multimodal HCI has sought for decades to endow computers with similar capabilities as human being. Dias *et al.* [11] developed an HCI method based on modalities of hand gestures and speech commands. Lee *et al*. [12] proposed a smart TV interaction approach based on fusion of face orientation and hand recognition. The face orientation is used in viewer authentication, and the hand gesture is adopted to control the volume and channel of the TV. Zhang *et al.* [13] integrated gaze and speech to select the interactive object based on distance comparison. Agrawal *et al.* [14] used head and hand to communicate with computers and other electronic devices. The proposed approach is useful in near distance interaction. However, it is time-consuming to compute optical flows during detecting head motion. Tu *et al.* [15] combined hand gestures and head pose to control the movement of robot's head. Since skin color has been used to locate the hand, the detection is sensitive to illumination changes. Carrino *et al.* [16] presented a HCI method based on three modalities including pointing gestures, symbolic gestures and speech. The proposed method needs the user to wear cameras and microphones on his or her arm, which reduces the interactive comfort and makes the interaction less natural. Stiefelhagen *et al.* [17] developed components for speech recognition, pointing recognition and head pose estimation. Li *et al.* [18] proposed a HCI based on multimodality recognition including hand gestures, body pose, head gestures as well as gaze. It is complicated to integrate a special range camera, an ordinary web camera and a pair of stereo cameras on a mobile robot platform.

While many HCI methods have been developed so far including using more than two modalities, most methods assume that there is only one user in the scenario and he (or she) is regarded as the interactive one. Moreover, most of the method does not consider the diversity of interactive habit among different users. A HCI method has been developed based on audio-visual perceptive features and dynamic interactive pooling. A mode analysis based fusion strategy is proposed to fuse the interactive responses from audio-visual modalities especially for the case there are deviations from the modalities at the same time. The developed approach can be applied with a better performance even in single modality with multiple interactive users in the scenario. The HCI scenario and multiple users with different interactive habits in the scene are altogether taken as a dynamic HCI pooling. The interactive pooling is updated dynamically to meet the users' demands according to their respective interactive intentions and priorities. In order to effectively perform HCI only in a camera and an ordinary microphone as the sensor to capture visual and audio information, respectively, a strategy based on the interactive priority is developed to choose an interactive user. The interactive authority and priority are updated dynamically according to the variation of HCI pooling. So every user has been offered a fair chance to perform HCI. The main contributions are as follows. The first contribution is that audio-visual perceptive features based HCI has been developed. A fusion strategy is developed to fuse the interactive responses from audio-visual modalities especially for the case there are deviations from the modalities at the same time. A mode analysis is employed to determine the final interactive response according to the modal priority. It can be applied with a better performance even in single modality with multiple interactive users in the scenario. The second contribution is that the user in the scenario is offered a fair chance to perform HCI by his interactive priority no matter how many modalities used, and the interactive authority and priority are updated dynamically according to the variation of HCI content. The third contribution, maybe the most important, is to take the scenario and multiple users together as a dynamic HCI pooling. The diversity of interactive habits among multiple users is considered in an

ordinary hardware from a crowded scene without any hypothesis for the scenario contents in advance. The pooling is updated dynamically to meet the user demands with the most intentions preferentially. Comparative study with state-of-the-arts has indicated the superior performance of the proposed approach.

The organization of the rest paper is as follows. In Section 2, an audio-visual perceptive features and dynamic interactive pooling based HCI is described in detail. Experimental results and analysis are shown in Section 3, and followed by some conclusions in Section 4.

2. Visual-Audio Perceptive Features Based HCI

2.1 Visual Features Based HCI

2.1.1 Face orientation estimation based HCI: People interact naturally with each other by using their face to convey visual information. For example, a person's face orientation is an indication of what he or she is the most interested in, or with which he or her interacts. Face orientation can be applied to perform HCI intuitively.

ASM is a statistical approach for shape modeling and feature extraction, which was originally proposed by Cootes *et al.* [19] and developed by other researchers over the past few years. ASM [19] is adopted to locate the facial feature points for face orientation based HCI instead of AAM [20] due to its less time in fitting.

An OpenNI SDK is utilized to get skeleton data, depth images and color images captured by Microsoft Kinect at first. The skeleton map consists of 25 distinct and major skeletal joints of human body. The head joint is employed to roughly estimate the face region by detecting the face only in the head joint region of color images in the method developed by Viola and Jones [21]. The size and position of face region is used to initialize the ASM model parameters. We use the ASM [19] to locate the facial contour feature point which is less sensitive to changes in illumination conditions, and more stable than other facial feature ones. Besides, the maximum geometrical distance between every two facial contour feature points can be gotten to improve the stability of face orientation estimation.

We take advantage of all the N facial contour feature points extracted by ASM in color map captured by Kinect to build a 2D facial geometric model. After the Kinect is calibrated by the method proposed in [22], the 2D geometric model is mapped to a 3D coordinate space as follows.

Each facial contour feature points in color image is presented by a 3D vector $P_i = [X_i, Y_i, Z_i]$ based on calibration procedure. Number the contour feature points according to the counterclockwise from 0 to (N-1). Based on the N numbered feature points, an initial triangulation divides the contour feature points into numerous triangles.

A chart of modeling facial geometry is given in Figure 1.

For each triangle, three vertices (A, B, C) in a triangle are employed to estimate the facial normal direction as following:

$$\vec{n}_f = \overrightarrow{P_A P_B} \otimes \overrightarrow{P_A P_C} = (X_A - X_B, Y_A - Y_B, Z_A - Z_B) \otimes (X_A - X_C, Y_A - Y_C, Z_A - Z_C) \quad (1)$$

Where the symbol $\otimes$ is a cross operator.

The facial interactive point is constructed as the origin with the nasal tip point for HCI as following:

$$T_{ft} = P_{nose} + c_0 \cdot \vec{n}_f = [X_{nose}, Y_{nose}, Z_{nose}] + c_0 [X_{nf}, Y_{nf}, Z_{nf}] \quad (2)$$

Where P_{nose} is the 3D coordinate of nasal tip point; c_0 is a constant to make $Z_{T_{ft}}$ to zero.

Since the depth map generated by the Kinect has some deficiencies including occlusion areas and small black holes, there exist some random incorrect depth data [23]. Meanwhile, the position of facial contour feature points will fluctuate with the rotation of face. The facial interactive points in (3) will be divergent from each other due to the cases

mentioned above. It is difficult to perform HCI steadily. To alleviate this problem, a temporal median filter strategy is developed as follows. Several successive frames are employed to perform temporal median filter to determine the facial interactive point as follows:

$$T_f = \frac{1}{F \cdot N_t} \sum_{i=0}^{F-1} \sum_{j=0}^{N_t-1} T_{ft}(i, j) \tag{3}$$

Where $T_{ft}(i, j)$ is an interactive point in the j^{th} triangle of i^{th} frame; F is frame number selected for the temporal median filtering (discussed later); N_t represents the whole triangles.

An example of face orientation estimation based HCI is shown in Figure 2.

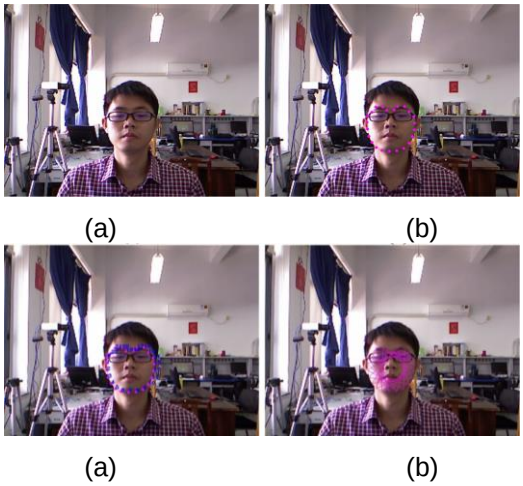


Figure 1. An Example of Facial Geometric Modeling: (a) An Original RGB Image. (b) Some Contour Feature Points Extracted. (c) Number the Points in an Anticlockwise Way. (d) Feature Points Triangulation

Figure 2. An Example of Face Orientation Estimation Based HCI: A Lamp on the Wall is Turned on According to the Face Orientation of One User

2.1.2 Hand pointing based HCI: Apart from face orientation as a user interface input modality, users are able to communicate with computer using their gestures that best suits their current request. Among the set of gestures intuitively performed by humans, pointing gesture are especially interesting for applications [24-25]. It does not require any a priori skills or training, and is a perfect candidate for the design of natural HCI based on computer vision [24-26].

As the same processing in the head region location mentioned above, we detect and track the joints of hands and shoulders in skeletal joints map to get the pointing direction.

In order to reliably find out the pointing fingertip, both RGB and depth information is combined to segment the pointing hand blob region by distance on the Kinect images stream as follows:

$$H_p = \begin{cases} 1, \text{if} \left| d - D_p \right| < \varepsilon \\ 0, \text{else} \end{cases} \tag{4}$$

Where d is the pixel in the depth image; D_p is depth value in the pointing hand joint; $|\square|$ is an absolute operator; ε is a threshold.

The pixel in the contour with the most significant gradient angle change would be the potential fingertip. The contour of the binary hand blob is used to compute the gradient angle as follows:

$$\theta = arccos \frac{\vec{a} \cdot \vec{b}}{\left\| \vec{a} \right\| \left\| \vec{b} \right\|} \tag{5}$$

Where

$$\begin{aligned} \vec{a} &= (X_{i+c} - X_i, Y_{i+c} - Y_i) \\ \vec{b} &= (X_{i-c} - X_i, Y_{i-c} - Y_i) \end{aligned} \tag{6}$$

Where the subscript c is a constant.

It is taken as a potential fingertip when $\sigma_0 < \sigma_1$, where σ_0 and σ_1 are thresholds, respectively.

The final outstretched fingertip is obtained according to the maximal distance between the hand center and the contour points. One example of outstretched fingertip location is shown in Figure 3 where the red point in Figure 3(d) indicates the extracted fingertip.

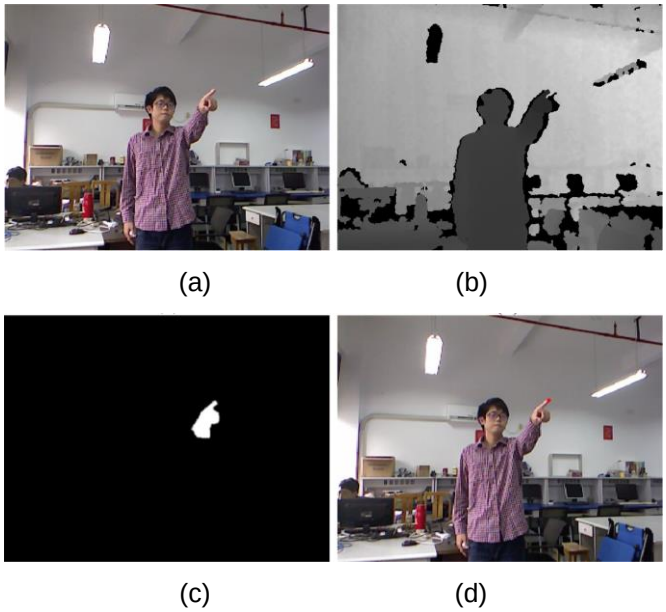


Figure 3. An Outstretched Fingertip Location: (a) An Original Color Image. (b) Corresponding Depth Image. (c) The Outstretched Hand Region Segmentation. (d) The Outstretched Fingertip Location

One key problem is how to recognize the pointing gesture for HCI. The whole motion of hand can be decomposed into three distinct phases: *point-to*, *move-to*, and *non-gesture*. Only in *point-to* phase, the motion of hand namely pointing gesture is meaningful for HCI.

When a user performs hand pointing based HCI, he will outstretch his hand and remain still for a certain period. Based on the location of outstretched fingertip, the pointing gesture P_{ing} is distinguished as follows:

$$P_{ing} = \frac{1}{N_p} \sum_{t=0}^{N_p-1} V_{hand} < V_p \tag{7}$$

Where N_p is a frame number to get the average velocity of pointing hand, V_p is a threshold, respectively, and

$$V_{hand} = \left\| P_{hand}^{t} - P_{hand}^{t-1} \right\| / T_t \tag{8}$$

Where T_t is time-consuming at each frame; P^t_{hand} = [X_{hand}, Y_{hand}, Z_{hand}] is 3D position of hand taken at time *t*.

We seek to interpret its intended pointing target after a pointing gesture has been detected. There are two different methods to estimate the direction of hand gesture. The first approach uses the line of sight between eye and hand while the second one employs the orientation of hand-shoulder. The pointing direction is determined in the latter case to differentiate the hand pointing from the face orientation as two different modalities for HCI.

The 3D position of fingertip $P_{fingertip}$ and the shoulder position $P_{shoulder}$ are used to calculate the final interactive point as follows:

$$T_p = P_{shoulder} + c_1 \cdot \vec{n}_p = [X_{shoulder}, Y_{shoulder}, Z_{shoulder}] + c_1 \cdot [X_{np}, Y_{np}, Z_{np}] \tag{9}$$

Where c_1 is a constant to make the value of Z_{T_p} to zero, and

$$\vec{n}_p = P_{fingertip} - P_{shoulder} \tag{10}$$

One example of hand pointing based HCI is given in Figure 4.

Figure 4. An Example of Hand Pointing Based HCI: A Lamp on the Wall is Turn on According to the Hand Pointing

2.2 Audio Features Based HCI

The human speech, which carries abundant information, is regard as an indispensable communication way. It can be taken as a modality for HCI intuitively. One key issue is how to recognize the speech. We use the speech recognition engine developed in the Hidden Markov Model (HMM) toolkit to recognize the isolated words. The main steps are as follows. Some participants are asked to speak different numbers repeatedly. These speech files are taken as raw data corpus used both for training and testing. The *HCopy* in the toolkit is employed to convert the speech files into feature vectors in the Mel-frequency cepstral coefficient. Some HMM parameters are initialized by the *HInt* in the

toolkit before training. The global means and variances are computed by the *HInt*, and all the Gaussian kernel parameters in the HMM are set as the same means and variances, respectively. The *HRest* in the toolkit are adopted to estimate the optimal values for the HMM parameters including transition probability, mean and variance vectors for each observation function. The *HVite* in the toolkit is designed to match the input acoustical observation with the Markov models of recognizer that is processed by a Viterbi algorithm.

It helps to improve the recognition based on audio-visual speech recognition especially in high noise environments [27-29]. We employed the visual lip feature whether the lip is open or closed to improve speech based HCI performance as follows.

The ASM [19] is used to extract 12 feature points in the outer lip contour. To distinguish whether the lip is open or closed, a geometric feature analysis is performed to measure the height and width between each feature points. For this purpose, we first divide the lip feature points into four groups including the upper lip with 5 points, the lower lip with 5 points, one left corner point, and one right corner point. Make use of the geometric lip features mentioned above to distinguish whether the lip is open or closed as following:

$$\lambda = \frac{\sum_{i=1}^{5} \left\| y_{upper}^{i} - y_{lower}^{i} \right\|}{\left\| x_{left} - x_{right} \right\|} \tag{11}$$

Where *x*, *y* are pixel coordinates of the lip feature points, respectively.

2.3 Dynamic Interactive Pooling Based HCI

To offer the computer with similar capabilities as human beings to a great extent, especially in a HCI scenario where multiple users with some different interactive ways want to perform HCI at the same time, a dynamic interactive pooling based HCI strategy has been developed. The interactive scenario is taken as a pooling, and each user in the scenario is offered a fair chance to perform HCI. As a result, all interactive users and scene together form a pooling. The interactive pooling is updated dynamically to meet the users' demands according to their respective interactive intentions and priorities among the three modalities mentioned above. The interactive priority is defined as follows. The pointing gesture is recognized only when the user raises his hand and points at an object statically, which shows more intense interactive intents compared to that of another two modalities. The pointing gesture is taken as a modality with the highest priority for HCI. Due to the fact that the sound from different source will be mixed and interfered with each other, it is difficult to locate sound source efficiently. The face orientation conveys fast interactive visual information compared to that of speech. The face orientation is taken as a modality with the secondary while the lip-speech is taken as a modality with the lowest priority for HCI.

Based on the interactive modalities and priorities, a HCI scenario pooling with multiple interactive users at the same time can be categorized into interaction with the same or different modal(s). The former indicates that multiple users all want to perform HCI in the same modal(s) at the same time. The latter indicates that multiple users use different interactive modalities in the pooling. In order to effectively perform HCI only in a camera and an ordinary microphone as the sensors to capture visual and audio information, respectively, a strategy for selecting an interactive user is developed as follows. A user is selected randomly and only he has the authority to perform HCI for the former case. A user is chose according to the interactive priority no matter interactive modality numbers for the latter case. When the chosen user stops interacting or the user's interactive priority

become lower, a user with the highest interactive priority will be selected dynamically from the pooling.

Another key issue is how to distinguish the user with the interactive authority from multiple users in the pooling. The Kinect sensor used has the ability to capture the skeletons of different users in the interactive pooling at the same time. The user in the pooling can be detected and tracked based on the skeletal information captured by the Kinect. In the meantime, users with different audio-visual interactive modalities can be distinguished according to the skeletons.

Once an interactive user is selected, his interactive target points (T_p, T_f, T_s) are determined, respectively as mentioned above. The issue is how to fuse the interactive responses from different modalities especially for the case there are deviations from the modalities at the same time. A mode analysis is employed to determine the final interactive point according to the priority defined previously as:

$$T_{final} = F(T_p, T_f, T_s) \tag{12}$$

Where F is a mode operator.

When the interactive responses are different from each other, the final response is given preference to the one with the highest priority.

It is necessary to point out that since there is fluctuation problem during face orientation estimation based HCI as mentioned above, the interactive target is determined only if the user's face orientation towards the target at several continuous frames (discussed later).

3. Experimental Results and Analysis

In order to test the performance of the proposed method, we have done experiments on the video sequences captured by a Kinect sensor and audio streams obtained by an ordinary microphone simultaneously, where the platform as shown in Figure 5. The nine lamps are taken as the interactive targets, and the experience can be understood intuitively by the user. The video frames captured by the Kinect are of 640×480 pixels. The experiment is done in a computer equipped with Pentium(R) 2.6GHz CPU, 4G RAM memory, C/C++ with OpenCV, OpenNI, and HTK under Microsoft Visual Studio 2010 IDE. The user stands in front of the Kinect or the microphone at a distance from 2m to 4m in the experiment. 10 volunteers with different heights and interactive habits participate in the experiment. In order to evaluate the performance of HCI, a correct percentage is defined as follows:

$$P = \frac{N - W}{N} \times 100\% \tag{13}$$

Where N is the total number in the whole test; W is the number of wrong interactions.

3.1 Parameter Analysis

Since there is fluctuation problem in face orientation estimation based HCI, a temporal median filtering is developed in (3) to get a steady interactive point. For F in (3), a large value will magnify the processing time while a small value will cause more fluctuation. A curve map that the variance of the facial interactive point T_f varies with F is shown in Figure 6.

It can be noted that the variance of T_f changes gently after F is more than 8 from Figure 6. F in (3) is set to 8, and remain unchanged in the experiment.

For ε in (4) is set to 120 to segment the pointing hand, and c in (6), σ_0, σ_1 are set to 5, -20, and 20, respectively to get potential fingertip. N_p and V_p in (7) are set to 55, 0.5, respectively to distinguish pointing gesture. λ in (11) is set to 4 to determine the lip shape. These parameters are kept the same in the whole experiment.

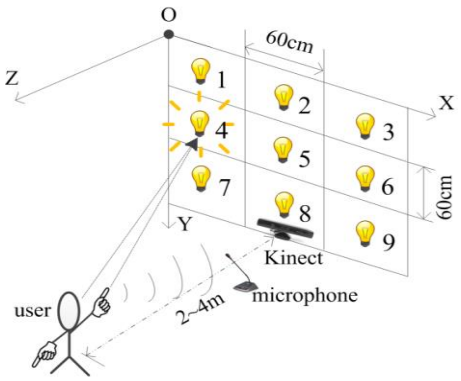


Figure 5. An Audio-Visual Perceptive Features and Dynamic Interactive Pooling Based HCI Skeletal Map

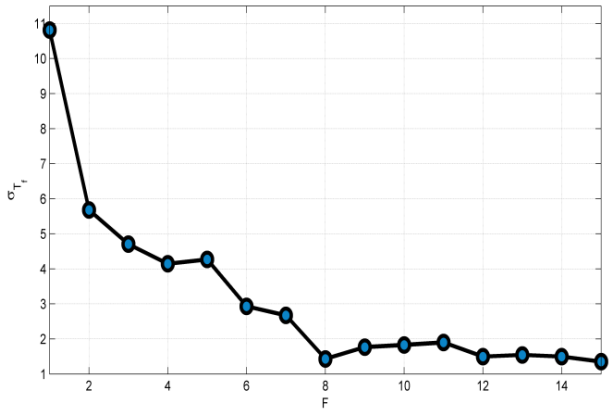


Figure 6. The Variance Variation of the Interactive Points T_f with F

3.2 Experimental Results

Some quantitative HCIs in confusion matrix based on the single interactive modality mentioned above are given in Table 1~4, respectively. L_1, L_2, L_3, L_4, L_5, L_6, L_7, L_8, and L_9 in the Tables represent the nine interactive lamps on the wall.

Table 1. Face Orientation Estimation Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**95.2%**	2.4%	0	1.8%	0.6%	0	0	0	0
*L*2	1.0%	**94.8%**	1.8%	0.4%	1.4%	0.6%	0	0	0
*L*3	0	1.6%	**95.6%**	0	0.8%	2.0%	0	0	0
*L*4	0.6%	0.4%	0	**96.6%**	1.0%	0	0.8%	0.6%	0
*L*5	0.4%	0.8%	0.4%	1.2%	**93.8%**	1.4%	0.2%	1.2%	0.6%
*L*6	0	0.6%	1.0%	0	1.6%	**95.4%**	0	0.4%	1.0%
*L*7	0	0	0	2.2%	1.6%	0	**93.2%**	3.0%	0
*L*8	0	0	0	0.2%	0.6%	0.4%	0.6%	**97.4%**	0.8%
*L*9	0	0	0	0	1.0%	1.8%	0	1.8%	**95.4%**

Table 2. Hand Pointing Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**97.6%**	0.6%	0	1.2%	0.6%	0	0	0	0
*L*2	0.6%	**96.2%**	0.6%	0.8%	1.0%	0.8%	0	0	0
*L*3	0	0.6%	**97.6%**	0	0.6%	1.2%	0	0	0
*L*4	0.4%	0.2%	0	**96.8%**	0.6%	0	1.2%	0.8%	0
*L*5	0.2%	0	0	0.2%	**97.6%**	0.2%	0.4%	0.8%	0.6%
*L*6	0	0.2%	0.4%	0	0.8%	**96.8%**	0	0.6%	1.2%
*L*7	0	0	0	0.8%	1.0%	0	**96.6%**	1.6%	0
*L*8	0	0	0	0.2%	0.4%	0.8%	1.4%	**95.4%**	1.8%
*L*9	0	0	0	0	0.4%	0.8%	0	2.0%	**96.8%**

Table 3. Speech Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**95.2%**	0	1.2%	1.2%	0	0	2.4%	0	0
*L*2	0	**96.2%**	0	1.2%	0	0	0	0.8%	1.8%
*L*3	0	0	**96.6%**	0	1.6%	0	1.8%	0	0
*L*4	0	1.2%	0	**95.2%**	0	2.4%	0	1.2%	0
*L*5	0	2.4%	0	0	**94.8%**	0	1.2%	0.6%	1.0%
*L*6	1.0%	0	0.6%	1.2%	0	**96.0%**	0	0.8%	0.4%
*L*7	3.2%	0	2.6%	0	0	0	**94.2%**	0	0
*L*8	0%	1.0%	0	1.4%	1.8%	0	0	**95.4%**	0.4%
*L*9	1.0%	0	0.8%	0	0	0	1.6%	0	**96.6%**

Table 4. Lip-Speech Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**97.6%**	0	0.6%	0.4%	0	0	1.4%	0	0
*L*2	0	**97.8%**	0	0.6%	0	0	0	0.6%	1.0%
*L*3	0	0	**98.2%**	0	0.8%	0	1.0%	0	0
*L*4	0	0.8%	0	**96.8%**	0	1.8%	0	0.6%	0
*L*5	0	1.6%	0	0	**96.2%**	0	0.8%	0.6%	0.8%
*L*6	0.6%	0	0.4%	1.0%	0	**97.4%**	0	0.4%	0.2%
*L*7	1.2%	0	0.6%	0	0	0	**98.2%**	0	0
*L*8	0	0.8%	0	1.0%	1.8%	0	0	**96.2%**	0.2%
*L*9	0.2%	0	0.6%	0	0	0	1.0%	0	**98.2%**

To further test the performance in multiple interactive modals at the same time, some quantitative results are given in Table 5 based on visual multimodalities including face orientation and hand pointing.

Table 5. Visual Multimodalities Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**97.8%**	0.6%	0	1.0%	0.6%	0	0	0	0
*L*2	0.2%	**96.8%**	0.4%	0.8%	1.2%	0.6%	0	0	0
*L*3	0	0.6%	**97.8%**	0	1.0%	0.6%	0	0	0
*L*4	0	0.2%	0	**98.2%**	0.2%	0	0.8%	0.6%	0
*L*5	0	0.2%	0	0.6%	**97.8%**	0.2%	0.4%	0.6%	0.2%
*L*6	0	0.2%	0.2%	0	0.8%	**97.0%**	0	0.8%	1.0%
*L*7	0	0	0	0.6%	0.8%	0	**97.2%**	1.4%	0
*L*8	0	0	0	0.4%	0.6%	0.2%	1.2%	**96.4%**	1.2%
*L*9	0	0	0	0	0.6%	0.8%	0	1.4%	**97.2%**

One can notice from Table 5 that the performance based on the both visual modalities is better than that of based on the single visual modal during HCI.

Another experiment result is given in Table 6 based on audio-visual multimodalities.

Table 6. Audio-Visual Modalities Based HCI

	*L*1	*L*2	*L*3	*L*4	*L*5	*L*6	*L*7	*L*8	*L*9
*L*1	**98.4%**	0.6%	0	1.0%	0	0	0	0	0
*L*2	0.2%	**98.6%**	0.4%	0	0.8%	0	0	0	0
*L*3	0	0.4%	**99.0%**	0	0	0.6%	0	0	0
*L*4	0.2%	0	0	**98.8%**	0.4%	0	0.6%	0	0
*L*5	0	0	0	0	**99.4%**	0.2%	0	0.4%	0
*L*6	0	0	0	0	0.4%	**98.8%**	0	0	0.8%
*L*7	0	0	0	0.4%	0	0	**98.4%**	1.2%	0
*L*8	0	0	0	0	0.2%	0	0.4%	**98.8%**	0.6%
*L*9	0	0	0	0	0.4%	0	0	1.0%	**98.6%**

One can further note from Table 6 that the performance based on the audio-visual modalities is the best by comparisons.

Moreover, one can find that the developed method has better HCI performances even in single modal with multiple interactive users at the same time from Table 1 to Table 6. It indicates that the developed method can be applied in a real HCI scene with a better performance.

To further evaluate proposed method, some approaches [10, 15] are selected to test the performance at the same conditions. The experimental results are shown in Table 7.

Table 7. Comparisons Among Different Methods

Method	Method in [10]	Method in [15]	**Proposed**
Average *PC*	95.6%	93.2 %	**98.8%**
Time-consuming	98ms	116ms	**103ms**

One can find that the average *PC* of proposed method is the highest among the investigated methods. In the time-consuming, the proposed method is equivalent to one in

[10] while it is superior to one in [15]. As a whole, the developed approach has the best performance by comparisons.

4. Conclusions

An audio-visual perceptive features and dynamic interactive pooling based HCI method has been proposed. A fusion strategy is developed to fuse the interactive responses from audio-visual modalities especially for the case there are deviations from the modalities at the same time. A mode analysis is employed to determine the final interactive response according to the modality priority. It can be applied with a better performance even in single modality with multiple interactive users in the scenario. The user in the scenario is offered a fair chance to perform HCI by his interactive priority no matter how many modalities used, and the interactive authority and priority are updated dynamically according to the variation of HCI content. The scenario and multiple users are altogether taken as a dynamic HCI pooling. The diversity of interactive habits among multiple users is considered in an ordinary hardware from a crowded scene without any hypothesis for the scenario contents in advance. The pooling is updated dynamically to meet the user demands with the most intentions preferentially. Comparative study with state-of-the-arts has indicated the superior performance of the proposed approach.

In the future, we would do further work in some more robust modalities to improve the HCI performance in dynamic cluttered scenarios.

Acknowledgements

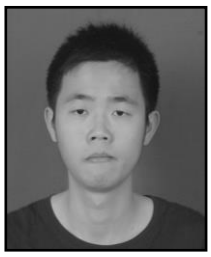
This work is supported in part by the Natural Science Foundation of China (Grant No.11176016, 60872117), and Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20123108110014).

References

[1] F. Karray, M. Alemzadeh, J. A. Saleh and M. N. Arab, “Human-computer interaction: overview on state of the art,” International Journal on Smart Sensing and Intelligent Systems, vol. 1 no. 1, **(2008)**, pp. 137-159.
[2] N. Sebe, “Multimodal interfaces: challenges and perspectives,” Journal of Ambient Intelligence and smart environments, vol. 1 no. 1, **(2009)**, pp. 23-30.
[3] M. Turk, “Multimodal interaction: a review,” Pattern Recognition Letters, vol. 36 no. 15, **(2014)**, pp. 189-195.
[4] M. Elmezain, A. A. Hamadi and B. Michaelis, “Hand trajectory-based gesture spotting and recognition using HMM,” Proceedings of International Conference on Image Processing, **(2009)**, pp. 3577-3580.
[5] H. I. Suk, B. K. Sin and S. W. Lee, “Hand gesture recognition based on dynamic Bayesian network framework,” Pattern Recognition, vol. 43 no.9, **(2010)**, pp. 3059-3072.
[6] M. V. D. Bergh and L. V. Gool, “Combining RGB and ToF cameras for real-time 3D hand gesture interaction,” Proceedings of 2011 IEEE Workshop on Applications of Computer Vision, **(2011)**, pp. 66-72.
[7] C. B. Park and S. W. Lee, “Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter,” Image and Vision Computing, vol. 29 no. 1, **(2011)**, pp. 51-63.
[8] R. Kehl and L. V. Gool, “Real-time pointing gesture recognition for an immersive environment”, Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition, **(2004)**, pp. 577-582.
[9] J. R. Michael, C. Shaun and J. Y. Li, “A multi-gesture interaction system using a 3-D iris disk model for gaze estimation and an active appearance model for 3-D hand pointing,” IEEE Transactions on Multimedia, vol. 13 no. 3, **(2011)**, pp. 474-486.
[10] J. Pan and Y. P. Guan, “Human-computer interaction using pointing gesture based on an adaptive virtual touch screen,” International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6 no. 4, **(2013)**, pp. 81-92.
[11] M. S. Dias, R. Bastos, J. Fernandes, J. Tavares and P. Santos, “Using hand gesture and speech in a multimodal augmented reality environment,” Proceedings of International Conference on Gesture-Based Human-Computer Interaction and Simulation, **(2009)**, pp. 175-180.

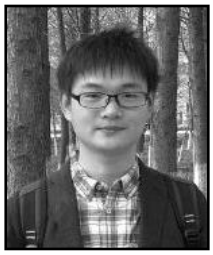
[12] S. H. Lee, M. K. Sohn, D. J. Kim, B. Kim and H. Kim, "Smart TV interaction system using face and hand gesture recognition," Proceedings of IEEE International Conference on Consumer Electronics, **(2013)**, pp. 173-174.
[13] Q. Zhang, A. Imamiya, K. Go and X. Mao, "A gaze and speech multimodal interface," Proceedings of IEEE International Conference on Distributed Computing Systems, **(2004)**, pp. 208-213.
[14] A. Agrawal, R. Raj and S. Porwal, "Vision-based multimodal human-computer interaction using hand and head gestures," Proceedings of IEEE International Conference on Information and Communication Technologies, **(2013)**, pp. 1288-1292.
[15] Y. J. Tu, C. C. Kao and H. Y. Lin, "Human computer interaction using face and gesture recognition," Proceedings of IEEE International Conference on Signal and Information Processing, **(2013)**, pp. 1-8.
[16] S. Carrino, A. Péclat, E. Mugellini, O. A. Khaled and R. Ingold, "Humans and smart environments: a novel multimodal interaction approach", Proceedings of ACM International Conference on Multimodal Interfaces, **(2011)**, pp. 105-112.
[17] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel and A.Waibel, "Natural human-robot interaction using speech, head pose and gestures," Proceedings of IEEE International Conference on Intelligent Robots and Systems, vol. 3, **(2004)**, pp. 2422-2427.
[18] Z. Li and R. Jarvis, "A multi-modal gesture recognition system in a human-robot interaction scenario," Proceedings of IEEE International Conference on Robotic and Sensors Environments, **(2009)**, pp. 41-46.
[19] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active shape models-their training and application," Computer Vision and Image Understanding, vol. 61 no. 1, **(1995)**, pp. 38-59.
[20] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23 no. 6, **(2001)**, pp. 681-685.
[21] P. Viola and M. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol. 57 no. 2, **(2004)**, pp. 137-154.
[22] D. Herrera, J. Kannala and J. Heikkila, "Joint depth and color camera calibration with distortion correction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34 no. 10, **(2012)**, pp. 2058-2064.
[23] F. Qi, J. Han, P. Wang, G. Shi and F. Li, "Structure guided fusion for depth map inpainting," Pattern Recognition Letters, vol. 34 no. 1, **(2012)**, pp. 70-76.
[24] M. Blumendorf, S. Feuerstack and S. Albayrak, "Multimodal user interfaces for smart environments: the multi-access service platform," Proceedings of the Working Conference on Advanced Visual Interfaces, **(2008)**, pp. 478-479.
[25] Y. P. Guan, "Non-wearable pointing gesture recognition based on single optimal view camera," Proceedings of International Conference on Computer Science and its Applications, **(2009)**, pp. 1-5.
[26] J. L. Sibert and M. Gokturk, "A finger-mounted, direct pointing device for mobile computing," Proceedings of ACM Symposium on User Interface Software and Technology, **(1997)**, pp. 41-42.
[27] P. Sujatha and M. R. Krishnan, "Lip feature extraction for visual speech recognition using Hidden Markov model," Proceedings of International Conference on Computing, Communication and Applications, **(2012)**, pp. 1-5.
[28] Y. Kashiwagi, M. Suzuki, N. Minematsu and K. Hirose, "Audio-visual feature integration based on piecewise linear transformation for noise robust automatic speech recognition," Proceedings of IEEE Workshop on Spoken Language Technology, **(2012)**, pp. 149-152.
[29] J. Gui and S. Wang, "Shape feature analysis for visual speech and speaker recognition," Proceedings of International Conference on Applied Informatics and Communication, **(2011)**, pp. 167-174.

Authors



Yulong Li, was born in Yueyang, Human Province, China. He is now a M.D Candidate of School of Communication and Information Engineering at Shanghai University, China. His general interests lie in human-computer interaction, pattern recognition and their application to computer vision.

Yepeng Guan, received his PhD degree from Central South University, China in 2000. He did his first postdoctoral research in Department of Electronic Engineering, Southeast University, China from 2000 to 2002. He did his second postdoctoral research in Department of Information Science and Electronics Engineering, Zhejiang University, China from 2003 to 2004. He is now a professor and doctoral supervisor of Department of Communication and Information Engineering, Shanghai University, China. He is a Shu Guang scholar of Shanghai and embedded by Who's Who in the World. His research interests include digital image processing and analysis, computer vision and so on.



Jinfeng Ma, was born in Zhengzhou, Henan Province, China. He is now a M.D Candidate of School of Communication and Information Engineering at Shanghai University, China. His main interests are computer vision and pattern recognition.