Multimodal Smart Interactive Presentation System

Hoang-An Le[1,2], Khoi-Nguyen C. Mac[1,2], Truong-An Pham[1],
Vinh-Tiep Nguyen[1], and Minh-Triet Tran[1]

[1] Faculty of Information Technology, University of Science, VNU-HCMC, Vietnam
[2] John von Neumann Institute, VNU-HCMC, Vietnam
{lhan.ict,mcknguyen.ict}@jvn.edu.vn, ptan@apcs.vn,
{nvtiep,tmtriet}@fit.hcmus.edu.vn

Abstract. The authors propose a system that allows presenters to control presentations in a natural way by their body gestures and vocal commands. Thus a presentation no longer follows strictly a rigid sequential structure but can be delivered in various flexible and content adapted scenarios. Our proposed system fuses three interaction modules: gesture recognition with Kinect 3D skeletal data, key concepts detection by context analysis from natural speech, and small-scaled hand gesture recognition with haptic data from smart phone sensors. Each module can process in realtime with the accuracy of 95.0%, 91.2%, and 90.1% respectively. The system uses events generated from the three modules to trigger pre-defined scenarios in a presentation to enhance the exciting experience for audiences.

Keywords: Smart environment, presentation system, natural interaction, gesture recognition, speech recognition.

1 Introduction

User interfaces aim to provide users with the most convenient ways to use computing systems. To enhance the usability of a system, various approaches have been studied and proposed to mimic natural inter-personal communications: ZeroTouch [11] enhances regular systems with multifinger interaction; a multi-user interaction system [10] allows users to control both desktop and wall-sized environment using depth-sensing techniques like Kinects or Wii remote controllers; a gaze-based interaction system [3] predicts users' behavior based on their looking. With a multimodal approach, Human Computer Interaction (HCI) provides users with not only ergonomic interfaces but also exciting user experiences.

Presentation is a popular activity in daily life such as in lectures, group discussions, or marketing campaigns. However, presenters usually stay away from computers (near the screen or walk around) during their talks, which may interrupt presentations when they go back to their computers for manipulation. Although different kinds of remote controls can be used, it would be more convenient for presenters to deliver their presentations simply by their body motions,

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 67–76, 2013.

gestures, and speech. This motivates our development of a smart environment that can understand human's behaviors and provide the most inspired and comfortable presentation with high naturalness and flexibility.

Our proposed system fuses three kinds of interaction: users action recognition with Kinect 3D skeletal data, key concepts identification from presenters' natural speech, and hand gesture recognition with smart phones' haptic data. The three kinds of information are captured simultaneously, analyzed in realtime and integrated to trigger certain pre-defined events in a presentation.

The gesture recognition module with 3D skeletal data is based on logistic regression and Dynamic Time Warping and can recognize users' actions with the accuracy of 95.0%. The speech recognition module uses hidden Markov model to recognize speech context automatically with the accuracy of 91.2%. The third module is proposed to overcome difficult situations when sophisticated small hand gestures (with boundary of movement of about 10 to 20 cm) are performed but cannot be recognized with 3D data from a Kinect. The authors use Dynamic Time Warping to process haptic data captured from smart phones' accelerometers to recognize accurately 90.1% of performed gesture.

The content of the paper is as follows. In Section 2, the authors briefly review the development of different methods for interaction with computers. Our proposed system and experimental results are presented in Section 3 and 4 respectively. Conclusions and future work are discussed in Section 5.

2 Related Work

To provide users with ergonomic experience, HCI technologies advance toward intelligent adaptive interfaces with not only unimodal but also multimodal approaches [9]. Each modality is a communication channel that connects the input and the output in an HCI design [6]. In the paper, our proposed system combines three modalities: action recognition using depth data and haptic data, and context recognition based on natural speech data.

There are two main approaches for the action recognition, vision-based or haptic-based, with different difficulties. The vision-based approach depends on the environmental lighting condition while the haptic-based approach requires expensive configuration [9]. The release of Microsoft Kinects with depth sensing technology provides a new trend to solve the difficulties of traditional camera computer vision [16].

For haptic data, it is possible to captured with accelerometer sensors. There are several studies about accelerometer-sensor-built systems for different purposes: Wii controller in presentation system [5], accelerometers in fall detection [15], and culture investigating system [13]. In this paper, dynamic time warping (DTW) is used for gesture recognition (in template matching phase) due to its simplicity, accuracy, and processing speed [12].

Speech recognition systems transform human's spoken speech into digital signal and transcribe the content based on learned data [1]. Since the first application built by Homer Dudley (1930), speech recognition has been developed and used in different applications: automatically call processing systems, query-based information systems, etc [7, 14]. With the capability to recognize hidden sequences of events, hidden Markov model (HMM) is a typical method in speech recognition [8]. Thus we follow this approach to recognize key concepts in a presentation in realtime from a presenter's natural speech.

3 Proposed System

Our proposed system shown in Figure 1 includes three main phases. In the first phase, the system captures a presenter' gesture data from a Kinect and smart phone sensors, and speech data from a microphone, then dispatchs them to appropriate processors in the second phase (c.f. Section 3.1, 3.2, and 3.3). In the last phase, the output gestures and events from the second phase are integrated to trigger corresponding pre-defined scenarios of visualization.

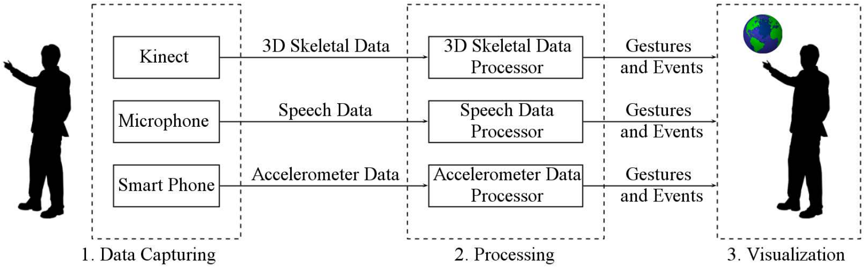


Fig. 1. Overview of the system for the smart presentation environment

3.1 Skeletal Data Based Action Recognition

Figure 2 illustrates 7 typical types of actions that a presenter usually performs during a talk. Although in the current system we focus only on these classes of actions, new categories can be added into the system using the same method to meet users' needs.

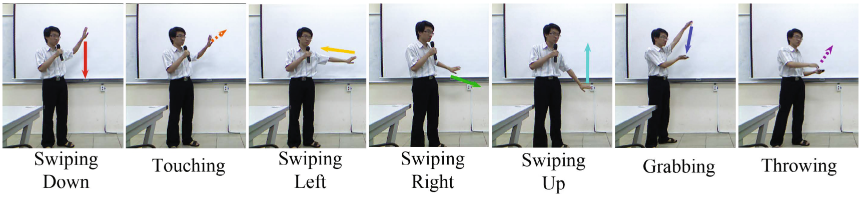


Fig. 2. Proposed action types for action recognition model

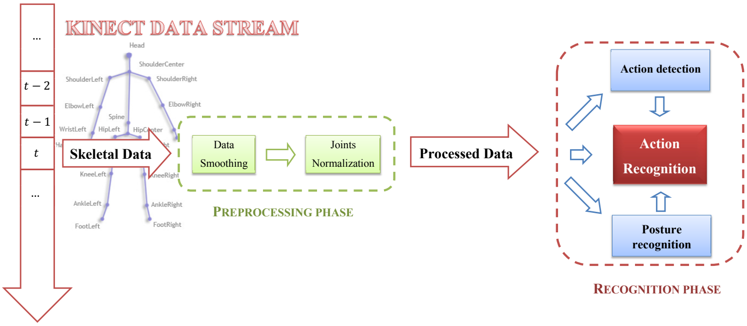


Fig. 3. The presentation action recognition using Kinect skeletal data

The action recognition process consists of 2 main phases (c.f. Figure 3): skeletal data with 20 joints are passed to the *preprocessing* phase, then to the main *recognition* phase.

Preprocessing Phase. The authors use Holt's double exponential smoothing method to correct the noise by the inference process of Kinect sensors [2]. The joint data is then normalized to be independent from the presenter's current position and orientation. The spine joint is selected as the origin of the users coordinate, the x-axis parallels to the shoulder, the y-axis points upward, and the z-axis is computed by the cross product of y- and x-axis.

Recognition Phase. As shown in Figure 4, during a presentation, the movements of a presenter's hands can be one of the two states: idle (slightly move with respect to the presenter) or active (when he or she is performing an action). By detecting the presenter's state, the recognition task can be narrowed down to only sequences of active frames. The detection is performed by logistic regression based on the *stability* measurement calculated by the standard deviation of the presenter's wrists and elbows in a frame sequence.

However, not all the detected active frames belong to meaningful actions but usually initiate special postures. For instance, a swiping down gesture (Figure 2) is usually started by first moving a hand up (frames 1-2) then moving it down (frames 5-7). The learning model used to classify each posture is the logistic regression and one-vs-all technique [4]. The model is trained and tested with different degrees of the feature space to choose the combination that give best prediction.

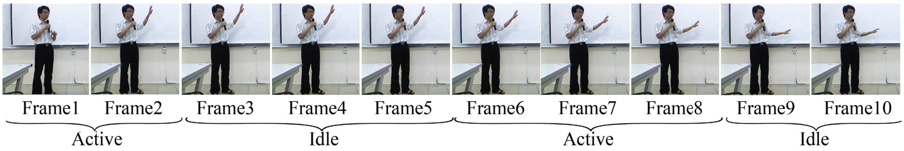


Fig. 4. The sequence of idleness

To recognize patterns of hand movements, features used for classification not only characterize joints' positions but also retain their temporal motion characteristics. The former can be solved by utilizing the information enhanced from the posture recognition subphase; for the latter, the authors construct reference sets consisting of collections of action sequences as a standard to which a given sequence is compared. The similarity between a given sequence and each reference set is used in the feature for recognition. The comparison process between two sequences is done by the Dynamic Time Warping algorithm [12].

3.2 Automatic Speech Recognition

The system records a speaker's commands as spoken speeches then recognizes key terms in the recorded speeches based on the hidden Markov model (HMM). The process has three main stages: data preprocessing, training, and recognition (c.f. Figure 5).

Data Preprocessing Stage. The stage prepares the audio files together with the dictionary and grammar. As Vietnamese language is monosyllabic [14], the phone of a word is kept as the word itself. Although in this paper, the training data cover only Vietnamese digits from zero to nine and solar system's planets, the method can be used with other topics and languages. To increase system performance, users can specify their own set of topics and language.

Training Stage. The training process includes several iterations in which the latter HMM$_{i+1}$ is computed from the former HMM$_i$. However, a model computed solely on training data are weak to match real life's odds. To increase system's performance, the authors add Gaussian mixtures at every three iterations during the training process. The training process is halted when system performance, which is recomputed after having a new model, does not significantly increase.

Recognition Stage. The new speech signals are transformed into MFCCs before being matched with the training models. The model with highest recognition accuracy is selected. The recognized words, however, might not be the desired ones. Wrong recognition appears because of unlearned words or noisy environment.

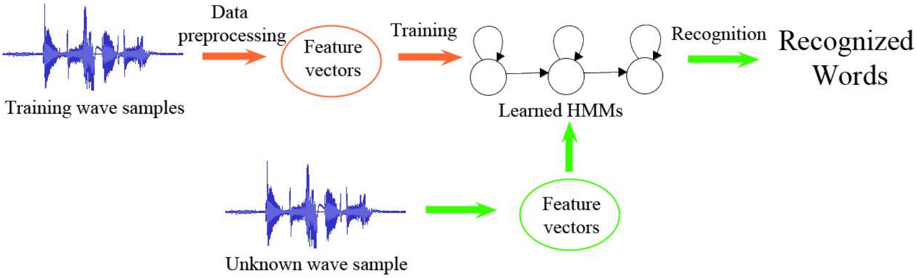


Fig. 5. Overview of the automatic speech recognition module

The authors use a filter to remove noisy words. After the system recognizes the training data, the authors can obtain the distribution of accuracy probabilities and the regular lasting duration of each key term. The filter is a matrix with each row having two thresholds (each is the subtraction of respective mean and standard deviation) of a term's accuracy and lasting duration. The recognized term is removed if either its accuracy or duration is less than the corresponding threshold.

3.3 Mobile Accelerometer Gesture Recognition

The gesture recognition with data captured from mobile devices' accelerometers is used to detect delicate gestures that are not appropriate to be recognized with 3D data from a Kinect, e.g. small or occluded hand movements. Accelerometer data is quantized and matched with a template library by dynamic time warping (DTW). Then a gesture can be recognized with the minimum distance (c.f. Figure6).

Data Preprocessing. A sequence of data captured from a mobile device's accelerometer as a 3D-vector is classified into five classes: moving leftward, rightward, upward, downward, and forward. Because the number of data points collected from a mobile device's accelerometer varies among samples, linear interpolation is used to sample data at a the sampling rate of 32Hz. Orientation values (tilt, yawn, and roll angles) obtained by a compass sensor are used to normalized the 3D vectors in term of world's coordinates.

Template Matching. Because DTW can match two series with temporal dynamics [12], it is used in our system to fix nonidentical time interval of gestures without normalizing the data points. DTW is used to calculate the distance between the quantized data and each template in the template library. A sequence is classified into the class of a template with the minimum DTW distance if the distance is less than a certain threshold.

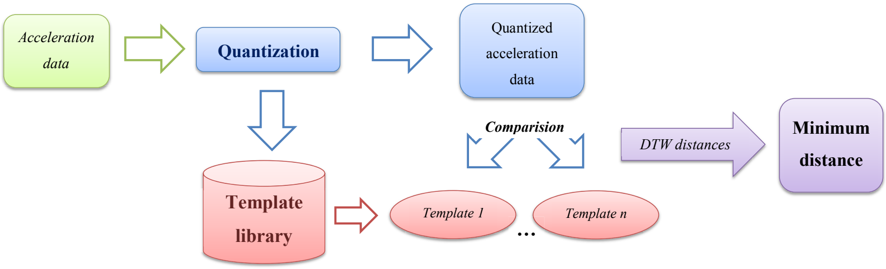


Fig. 6. Overview of the mobile accelerometer gesture recognition module

4 Experiences and Results

4.1 Skeletal Data Based Action Recognition

The data for the action recognition process are collected from 35 clips of 7 action types (c.f. Figure 2). All clips have the same length of 500 frames and the same frame rate of 20fps. To overcome the issue of underfitting and overfitting, the logistic regression models are trained with different degrees (chosen from 1 to 5) and tested with both the training and cross validation sets.

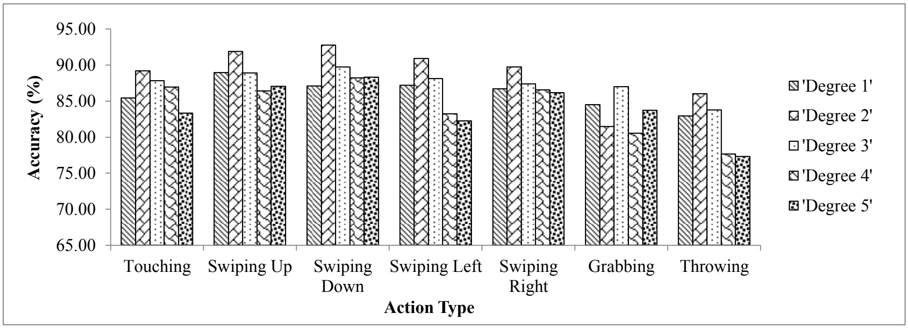


Fig. 7. Recognition accuracy for each action type with different model degrees

As shown in Figure 8, the average accuracy of each action is about 85 – 95%. Although there are some actions, such as swiping up and swiping down, with high accuracy (about 95%), the accuracy of the grabbing and throwing actions are quite low (about 85%). The deficiency of the action's accuracy could be explained as a grabbing or a throwing action (as shown in Figure 2) requires a user to do more rotation to one side in which hand movements are occluded by the user's body and thus can confuse the Kinect skeletal tracking system.

4.2 Automatic Speech Recognition

The training data contains 970 samples of two topics: solar system's planets (550 samples) and digits from zero to nine (420 samples). Data is recorded in waveform audio format (WAV) with monochannel and sampling rate of 11,025Hz. Figure 8 shows the system's recognition accuracy over different number of Gaussian mixtures and data portions. The experiment suggests that more training samples can give better recognition accuracy as they can cover more real life scenarios. Besides, with a certain data portion, increasing the number of Gaussian mixtures improves the accuracy. The highest accuracy in our experiment is 91.2% and corresponds to the case of 30 mixtures, the highest number of mixtures in experiments.

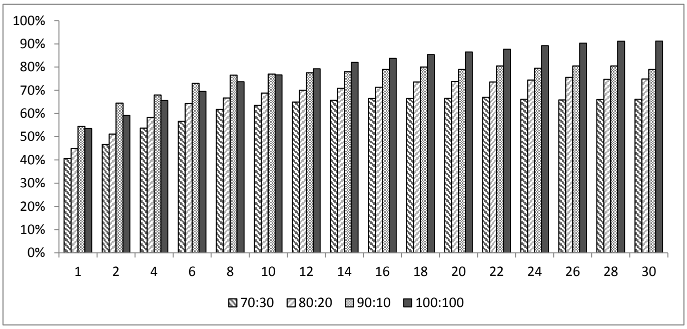


Fig. 8. Recognition accuracy corresponding to different number of Gaussian mixtures (horizontal axis) over four scenarios, each with the ratio between training and testing data respectively be 70:30, 80:20, 90:10, and 100:100.

4.3 Mobile Accelerometer Gesture Recognition

Figure 9 illustrates system's accuracy using DTW. The data with 200 samples (with the sampling rate of 32Hz) are divided into training and testing sets with the ratio of 1:1. The testing samples are modified so that they are 15 degrees tilted, compared to the training samples. The experiment, however, shows that tilting does not affect much on the system's accuracy as the average accuracy by DTW can be 90.1%. Besides, data with orientation have higher accuracy than the other case. Therefore, the suitable settings are those with data orientation using DTW method.

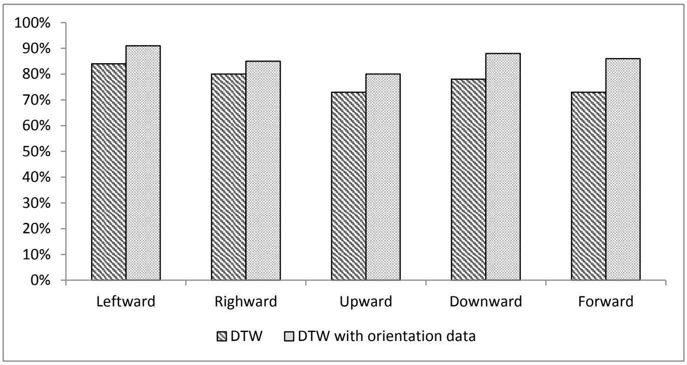
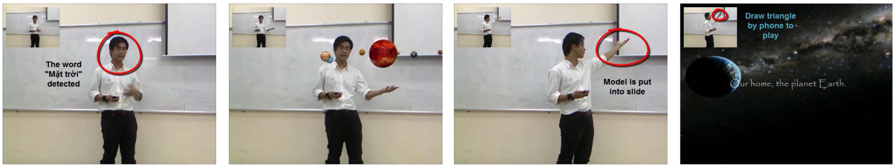


Fig. 9. System's accuracy using DTW matching method with normal and orientation data



(a) (b) (c) (d)

Fig. 10. An experimental scenario of the system: (a) speech recognition, (b) gesture recognition and visualization, (c) putting model into slide, and (d) using mobile accelerometer

4.4 An Experimental Scenario

Figure 10 shows the scenarios originated from a demonstrative presentation. The demonstration window consists of two screens, the small upper left shows the presenter in real life captured by a camera while the rest is what being currently displayed to the audience on the presentation screen. The annotations are interted into the four frames to illustrate triggered events.

Figure 10a indicates the recognition of the term "mat troi" ("the sun" in English) from presenter's speech, which starts the astronomy lecture. When the lecturer puts his right hand as if he were holding a 3D model of the solar system, the Kinect recognizes the pose and the system displays an augmented 3D model of the solar system above his right hand (Figure 10b). In Figure 10c, the action of moving a hand upward to the presentation screen is recognized as putting the augmented model into the screen and thus opens a full-screen video clip about the solar system. To start the clip, the presenter draws a triangle with a smart phone (Figure 10d), which triggers the "play" event via the phone's accelerometer.

5 Conclusion and Future Work

This paper introduces a smart interaction system that can react to presenters common and natural behaviors, using their body gestures and spoken speeches. The system has three main modules: gesture recognition using Kinect's 3D skeletal structure, speech recognition using normal microphones, and hand gesture recognition using mobile devices' accelerometer.

By experiment, our system can run with high accuracy in real time: 95.0%, 91.2%, and 90.1% for the three modules respectively. Besides, each module can run independently and can be trained with user's personal data to match demanded reactions, i.e. one can make the presentation of mathematics, physics, or chemistry with different gesture and command set.

For future works, the authors propose to upgrade the system into an interactive framework that allows users to use under several scenarios: meetings, seminars, education, etc. It allows users to integrate different devices (clients) with the framework (server) and provides them with several kinds of interaction and high comfort.

Acknowledgement. This research is supported by research funding from Advanced Program in Computer Science, University of Science and John von Neumann institute, Vietnam National University - Ho Chi Minh City.

References

1. Adams, R.: Sourcebook of automatic identification and data collection. Van Nostrand Reinhold (1990)
2. Azimi, M.: Skeletal Joint Smoothing White Paper (accessed August 13, 2012), `http://msdn.microsoft.com/en-us/library/jj131429.aspx`
3. Bednarik, R., Vrzakova, H., Hradis, M.: What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In: ETRA, pp. 83–90 (2012)
4. Hilbe, J.: Logistic regression models. CRC Press, Boca Raton (2009)
5. Holzinger, A., Softic, S., Stickel, C., Ebner, M., Debevc, M.: Intuitive e-teaching by using combined hci devices: Experiences with wiimote applications. In: Stephanidis, C. (ed.) UAHCI 2009, Part III. LNCS, vol. 5616, pp. 44–52. Springer, Heidelberg (2009)
6. Jaimes, A., Sebe, N.: Multimodal Human-Computer Interaction: A survey. Computer Vision and Image Understanding 108(1-2), 116–134 (2007)
7. Juang, B.H., Rabiner, L.R.: Automatic speech recognition - a brief history of the technology development. Elsevier Encyclopedia of Language and Linguistics (2005)
8. Jurafsky, D., Martin, J.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall (2009)
9. Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N.: Human-Computer Interaction: Overview on State of the Art. International Journal on Smart Sensing and Intelligent Systems 1(1), 137–159 (2008)
10. Lou, Y., Wu, W., Zhang, H., Zhang, H., Chen, Y.: A multi-user interaction system based on kinect and wii remote. In: ICME Workshops, p. 667 (2012)
11. Moeller, J., Kerne, A.: Zerotouch: an optical multi-touch and free-air interaction architecture. In: CHI, pp. 2165–2174 (2012)
12. Myers, C., Rabiner, L.: A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition. The Bell System Technical Journal 60(7), 1389–1409 (1981)
13. Rehm, M., Bee, N., André, E.: Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, BCS-HCI 2008. British Computer Society, vol. 1, pp. 13–22 (2008)
14. Vu, Q., Demuynck, K., Van Compernolle, D.: Vietnamese automatic speech recognition: The fLavor approach. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (eds.) ISCSLP 2006. LNCS (LNAI), vol. 4274, pp. 464–474. Springer, Heidelberg (2006)
15. Zhang, T., Wang, J., Liu, P., Hou, J.: Fall detection by embedding an accelerometer in cellphone and using kfd algorithm. International Journal of Computer Science and Network Security 6(10) (October 2006)
16. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia 19(2), 4–10 (2012)