MULTI-MODAL FUSION OF SPEECH-GESTURE USING INTEGRATED PROBABILITY DENSITY FUNCTION

Chi-Geun Lee, Mun-Sung Han
Electronic and Telecommunication Research Institute, u-Computing service Research Team, Korea

Chang-Seok Bae, Jin-Tae Kim
Electronic and Telecommunication Research Institute, u-Computing service Research Team, Korea

Keywords: Probability Density Function (PDF), Multi-modal recognition, Classifier level Fusion Recognition, Pattern Recognition.

Abstract: Recently, multi-modal recognition has become a hot topic in the field of Ubiquitous, Speech and gesture recognition, especially, are the most important modalities of human-to-machine interaction. Although speech recognition has been explored extensively and successfully developed, it still encounters serious errors in noisy environments. In such cases, gestures, a by-product of speech, can be used to help interpret the speech. In this paper, we propose a method of multi-modal fusion recognition of speech-gesture using integrated discrete probability density function omit estimated by a histogram. The method is tested with a microphone and a 3-axis accelerator in a real-time experiment. The test has two parts : a method of add-and-accumulate speech and gesture probability density functions respectively, and a more complicated method of creating new probability density function from integrating the two PDF's of speech and gesture.

1 INTRODUCTION

Computing applications are becoming increasingly complex and pervasive as exemplified by ubiquitous microprocessors in every appliances and resulting feature explosion. Recent technologies like speech or gesture recognition can make such systems more natural, easy to use, and robust. With the rapid advancement of computer software and hardware technology, the field of human-to-human interface has been developed. Along with aid of computers, many researchers and engineers have made machines user-friendly. The natural combination of a variety of modalities such as speech, gesture, gaze, and facial expression makes human-to-human communication easy, flexible and powerful (Sharma et al., 1998). Similarly, when interacting with computer systems, users seem to prefer a combination of several modes to a single one alone.

Despite the much interest and extensive research in the last decade, human-to-computer interaction (HCI) is still at an early stage of development. Therefore its ultimate goal of building natural perceptual user interfaces remains a challenging problem.

Two concurrent factors produce awkwardness. First, the current HCI systems make use of both rigid rules and syntax over the individual modalities involved in dialogues. Second, speech and gesture recognition, gaze tracking and other channels are isolated because we do not understand how to integrate them to maximize their joint benefit.

The speech recognition, in particular, needs great improvements on the robustness and accuracy to overcome the noisy effect of ambient environment for commercialization. With growing need and interest on enhancement of multi-modal fusion technology, the research for the fusion method of the various modalities is actively in progress.

The multi-modal fusion method is classified into features vector level fusion and classifier level fusion. Feature vector level fusion method has an advantage of using correlation of the feature data information of each modality, but it suffers from computational complexity due to increase of the number of feature vectors. Another downside of this method is that it requires a large set of training information for a robust recognition rate. On the

Lee C., Han M., Bae C. and Kim J. (2009).
MULTI-MODAL FUSION OF SPEECH-GESTURE USING INTEGRATED PROBABILITY DENSITY FUNCTION.
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*, pages 211-215
DOI: 10.5220/0001432402110215

contrary, Classifier fusion method has a key advantage of simplicity because of recognizers that are independent, but it has a flaw of its inability to utilize feature vectors. The simplest form of classifier fusion method is integrating the weighted values of two recognition results.

In recent studies, several improvements on various methods are proposed. Some of the fusion methods proposed are a method to use each modality information asynchronously (Alissali et al., 1996), a method to calculate a suitable weighted value dynamically according to the confidence of each modality (Heckmann et al., 2001), (Glotin et al., 2001), (Ghosh et al., 2001), a method to add new recognizer in order to combine each recognition result. The following is the usual procedure for fusion recognition process: From each modality, the feature vector is extracted from received input data. Each recognition score is then estimated by various recognition algorithms. Finally weighted value is given to the recognition score of each modality in order to integrate the modalities (Heckmann et al., 2001), (Glotin et al., 2001), (Ghosh et al., 2001), (Kim et al., 2003), (Kwak et al., 2006).

One of the most important problems in the multi-modal fusion recognition is how to integrate two modalities. Usually, the number of sample data has to be large enough to yield an acceptable PDD (Probability Density Distribution) in order for a recognition process to be executed on.

In this paper, a discrete probability density function (PDF) estimated by a histogram for speech-gesture multimodal fusion is used extensively. The discrete PDF by histogram is known to produce a good estimation of a distribution if the number of samples is sufficient, but the demerit of this approach is the inherent discontinuity. To avoid this problem, this paper proposes to use integrated discrete PDF instead of discrete PDF as it is. In this way, a reasonable estimation for all value ranges can be obtained. Furthermore, even in cases that the number of sample is not sufficiently large, a reasonable estimation can be obtained.

The proposed method is tested with microphone and 3-axis accelerator in real-time environment. The test has two parts: a simple method of add-and-accumulate speech and gesture probability density distribution(PDD) separately, and a more complicated method of creating new probability density distribution from integrating the two PDD's of speech and gesture.

The integrated PDF method that proposed in this paper had improvement of performance about 3% compare to the add-and accumulate method.

In section 2, the proposed speech-gesture fusion recognition system is described. Section 3 explains the multimodal fusion algorithm using integrated PDF. Utterance and gesture model list which used in experiment is explained in section 4. The experimental result is given in section 5. Finally, we describe conclusion in section 6.

2 SYSTEM ARCHITECTURE

The Speech-gesture fusion recognition system architecture that implemented for experiment in this paper is shown in figure 1 and figure 2.

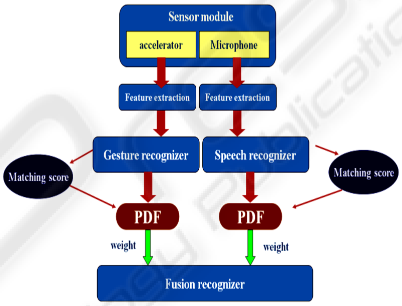

Figure 1: Multi-modal Fusion system using add-and-accumulate PDD's.

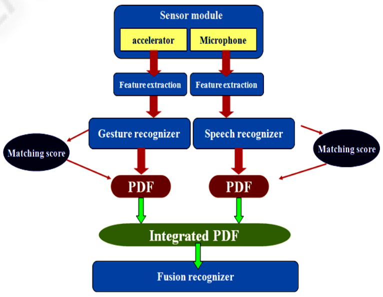

Figure 2: Multi-modal Fusion system using integrate PDD's.

The architecture consists of sensor module, feature extraction module, independent recognition module, and fusion recognition module. The sensor module consists of 3-axis accelerometer and a microphone in order to obtain the speech and gesture input data. The feature extraction module is to extract the feature vector from speech and gesture input data. Speech feature extraction module is comprised of the following two modules: Start and End point Detection module based on the frame energy, and a feature extraction module based on Zero-Crossing with Peak Amplitude (ZCPA) and RelAtive SpecTrAl algorithm (RASTA). The gesture feature extraction module extracts feature vectors from accelerometer-based gesture. This module consists of the following two modules: a Start Point Detection module based on threshold and a Feature Extraction module based on velocity vector.

Speech and gesture recognizer trains the feature vector received from feature extraction module and recognition process is executed. Fusion recognition module then calculates the PDF (Probability Density Function) based on recognition score from the results of speech and gesture recognition followed by creating PDF files. The PDF files are given weighted value to yield the highest fusion performance. The fusion recognition module is then executed either by adding each PDD or by creating a new histogram according to two different PDDs.

3 MULTI-MODAL FUSION RECOGNITION ALGORITHM

In this paper, we calculate the recognition score for the speech and gesture using recognition result of each modality. Subsequently, we make histogram about on speech and gesture using recognition score respectively, and then, we calculate the each Probability Density distribution for speech and gesture using the histogram already created.

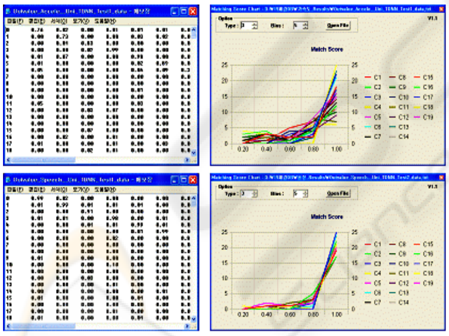

Figure 3: Gesture-Speech recognition score & Histogram.

The recognition score means a degree of confidence in the recognition and we can assume that the larger the recognition score is, the higher the degree of confidence is. In each recognition system, the recognizer decides that the input data is that of the learned gesture-speech model with highest recognition score.

Figure 3 shows the distribution of the recognition scores for each recognition system. The distributions are histograms of recognition scores that are normalized such that the sum of all bins equals 1.

3.1 Probability Density & Fusion Recognition Algorithm

In this paper, we propose to use an integrated discrete probability density function for Speech-gesture fusion recognition. When a certain random variable cannot be modeled efficiently by well-known probability density function, integrated discrete PDF is an effective alternate method to estimate. With the advantage of its simplicity and adoptability, the histogram method yields a reasonable estimate of the original PDF only if there is sufficient number of samples (Heckmann et al., 2001). In our experiment, accumulated and integrated methods of discrete PDF have been compared to show the integrated method is indeed effective.

Accumulated discrete PDF is an accumulation of the discrete PDF by the following equation.

$$P_S(K) = \sum_{n=0}^{k} H_S(n) \tag{1}$$

$$P_G(K) = \sum_{n=0}^{k} H_G(n) \tag{2}$$

In the above equation, $H_S(X)$ and $H_G(X)$ are histograms that are created by recognition score for each Speech and Gesture class. N is the number of bins in the histogram. $P_S(X)$ and $P_G(X)$ are accumulated discrete PDF for Speech and Gesture classes respectively. $P_S(X)$ is the probability of recognition score that belongs to speech class and $P_G(X)$ is of gesture class. Each value ranges from 0 to 1 and is normalized so that same value means same degree of confidence in recognition systems.

In this paper, we implemented accumulated discrete PDF system first. For the Speech-gesture fusion recognition, we add two PDFs, $P_S(X)$ and $P_G(X)$ from each recognition system.

The following equation is speech-gesture fusion method used in accumulated discrete PDF system.

When a recognition score X_S from speech recognition system and X_G from gesture recognition system are present, the fusion probability F is,

$$F = w_S \bullet P_S(X_S) + (1 - w_S) \bullet P_G(X_G) \tag{3}$$

where, $P_S(X)$ and $P_G(X)$ are accumulated discrete PDFs for speech and Gesture recognition systems respectively. w is the weighting factor to yield higher performance of fusion recognition system.

As the Second method that proposed in this paper, we can present following equation based on the equations (1)(2)(3) for the integrated discrete PDF,

$$P_{SG}(K) = \sum_{n=0}^{k} H_{SG}(n) \qquad (4)$$

where, P_{SG} is integrated discrete probability density function for speech and gesture. H_{SG} is a histogram obtained from recognition score for the speech and gesture class. N is the number of bins in the histogram.

4 GESTURE AND UTTERANCE

In this paper, we define 19 kinds of speech and natural gestures model separately, which is happen in daily life concurrently.

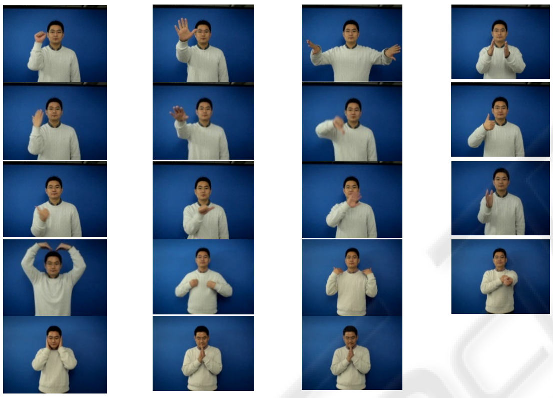

Figure 4: Gesture model.

For the define of gesture and speech model, we selected general speech and symbolic gesture such as grab, release, open, close, come here, bye, up, down, right, left, hot, cold, sweat, etc.

Figure 4 is shown that gesture model which for experiment in this paper.

5 EXPERIMENT AND RESULT

Initially, we analyze the two modalities, namely speech and body gesture separately. Two modalities are asynchronous but temporally correlated. For our concurrent speech and body gesture, classifier fusion method is the most common way of integrating such modalities (Glotin et al., 2001). We have experimented speech-gesture fusion recognition systems with 19 utterance and gesture.

The test has two parts : a method of add-and-accumulate speech and gesture probability density functions separately, and a more complicated method of creating new probability density function from integrating the two PDF's of speech and gesture. Test is executed in real time in a noisy environment, and it tested with the speaker-dependent method.

Speech and gesture feature is extracted by using zero-crossing with peak amplitude and delta values between start and end point of gesture. The experimental data set is collected with a microphone and 3-axis accelerator. Data corpus consists of 200 units per each speech and gestures. 150 training units are used, and 50 units are used as test data.

Table 1: Comparison of fusion recognition rate according to fusion methods.

SNR	Accumulated PDF	Integrated PDF
Clean	96.4	98.6
5dB	94.0	97.0
0dB	81.4	85.2
-5dB	64.8	69.6

Table 1 shows the performance for the two fusion recognition system according to fusion methods. As shown in table 1, our system has some improvements on fusion recognition rate in various noisy environments compared with accumulated PDF fusion method. If the number of sample data or modalities were increased, the fusion recognition performance will be more improved because of the characteristics of probability density function.

Therefore, the result of this experiment demonstrates that our fusion method is more efficient than the accumulated PDF fusion method.

6 CONCLUSIONS

Multi-modal fusion recognition has been tried by many studies to compensate poor performance in recognition rate of speech in a noisy condition. Especially, Speech and gesture are the most important modalities of human to machine interaction. For the Multi-modal fusion recognition, it is essential to use efficient integration method.

In this paper, we proposed the new method that integrates the Probability Density Function. Our experiment has shown that the fusion recognition performance is improved by using the integrated PDF compared to the legacy accumulated PDF method. We expect that our proposed method is useful in the classifier level fusion recognition area.

ACKNOWLEDGEMENTS

This work was supported by the IT R&D program of MKE/IITA, (2006-S032-02, Development of an Intelligent Service technology based on the Personal Life Log), (2008-P1-15-07J31, Research on Human friendly Next Generation PC Technology Standardzation on u-Computing).

REFERENCES

Rajeev Sharma, Vladimir I. Pavlovic. Thomas S.Huang, "Toward Multimodal Human-Computer Interface", Proceedings of the IEEE Vol. 86. No5. May 1998.

D. Alissali, P. Deleglise, A. Rogozan, "Asynchronous integration of visual information in an automatic speech recognition system", in Proc. of Fourth International Conference on Spoken Language(ICSLP 96), Vol. 1, pp. 34 -37, 1996

M. Heckmann, F. Berthommier, K. Kroschel, "Optimal weighting of posteriors for audio-visual speech recognition", in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 161 -164 , 2001

H. Glotin, D. Vergyr, C. Neti, G. Potamianos, J. Luettin, "Weighting schemes for audio-visual fusion in speech recognition", in Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 173 -176, 2001

A. Ghosh, A. Verma, A. Sarkar, "Using likelihood L-statistics to measure confidence in audio-visual speech recognition", in Proc. of IEEE Fourth Workshop on Multimedia Signal Processing, pp. 27-32, 2001

S. Lucey, S. Sridharan, V. Chandran, "Improved speech recognition using adaptive audio-visual fusion via a stochastic secondary classifier", in Proc. of International Symposium on Intelligent Multimedia, Video and Speech Processing, pp. 551 -554, 2001.

D. Kim, J, Lee, J. Soh, Y. Chung. "Real-time Face Verification using Multiple Feature Combination and SVM Supervisor", ICASSP, 2003, Vol. 2, pp 353-356.

Keun-Chang Kwak, Kyu-Dae Ban, Kyung-Suk Bae, "Speech-based Human-Robot Interaction Components for URC Intelligent Service Robot", IEEE/RSJ International Conference on Intelligent Robots and Systems, Video Session, 2006.

U. Meier, R. Stiefelhagen, J. Yang, A. Waibel, "Towards Unrestricted Lipreading", *International Journal of pattern Recognition and Artificial Intelligence*, Vol. 14, No. 5, pp. 571-785, 2000.

Byung-Jun Yoon and P.P. Vaidynathan, "Discrete PDF estimation in the presence of noise", Proc. International Symposium on Circuits and Systems (ISCAS), Vancouver, May 2004.