Recognition of Voice and Hand activities through Fusion of Acceleration and Speech[1]

Young-Giu Jung, ChangSeok Bae and Mun-Sung Han

Electronics and Telecommunications Research Institute,
138 Gajeongno Yuseong-gu, Daejeon, Korea,
{reraj, csbae, msh}@etri.re.kr

Abstract. Hand activity and speech comprise the most important modalities of human-to-agent interaction. So a multimodal interface can achieve more natural and effective human-agent interaction. In this paper, we suggest a novel technique for improving the performance of accelerometer-based hand activity recognition system using fusion of speech. The speech data is used in our experiment as the complementary sensor data to the acceleration data in an attempt to improve the performance of hand activity recognizer. This recognizer is designed to be capable of classifying nineteen hand activities. It consists of 10 natural gestures, e.g., 'go left', 'over here' and 9 emotional expressions by hand activity, e.g., 'I feel hot', 'I love you'. To improve performance of hand activity recognition using feature fusion, we propose a modified Time Delay Neural Network (TDNN) architecture with a dedicated fusion layer and a time normalization layer. Our experimental result shows that the performance of this system yields an improvement of about 6.96% compared to the use of accelerometers alone.

Keywords: multimodal interaction, hand activity recognition, modified TDNN, human cognitive

1 Introduction

Interface technology using activity or gesture is one of the key functions for agent system in ubiquitous computing environment. In general, accelerometers are currently among the most widely studied wearable sensors for activity or gesture recognition, thanks to their accuracy in the detection of human body movements, small in size, and reasonable power consumptions[1]. Ling Bao and others[2] presented algorithms to detect physical activities from data acquired using five small biaxial accelerometers worn simultaneously on different parts of the body.

In the study of Bharatula and others[3], a low power sensor hardware system was presented, including accelerometer, light sensor, microphone, and wireless

[1] This work was supported by the IT R&D program of MIC/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 Standardization on u-Computing]

communication. Accelerometer has been widely used in many pattern recognition methods in order to access physical activities. In the study of Kiani and others[4], the artificial neural networks were used. There have been several research efforts to enhance the performance of accelerometer-based activity recognition.

We use a human cognitive-based technique for improving the performance of activity recognition. Humans do not depend only on their hearing in order to recognize information. This fact is illustrated by the McGurk effect[5], a perceptual phenomenon which demonstrates an interaction between hearing and vision in speech perception. The presentation of an audio /p/ with a synchronized incongruent visual /k/ often leads listeners to identify what they hear as /t/, a phenomenon referred to as 'fusion'

Currently, the recognition of human input using data fusion has been partially achieved in a lip-reading system. The fusion algorithm can be carried out either at the feature-level or the class-level[6]. Figure 1 shows the block diagram of the class-level fusion. Two input signals are separately classified, and the results of each classifier are combined in next step. Fusion module has a set of algorithms to integrate the individual decision of each sensor. Several different methods of class-level fusion have been proposed and studied extensively such as voting method, behavior-knowledge space method and soft-output classifier fusion method[6].

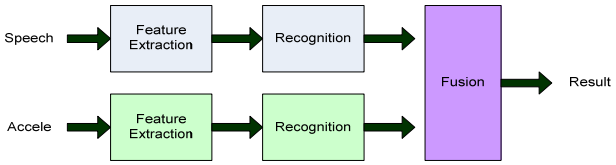

Fig. 1. The fusion at class level

Figure 2 shows the block diagram of the feature-level fusion. Two feature vectors of accelerometer and speech signal are combined into a joint feature vector, and the joint feature vector is used as input vector of fusion classifier. As already mentioned, the feature fusion uses a single classifier to fuse two modalities. Several approach have been proposed : fuzzy logic, Artificial Neural Network(ANN), Hidden Markov Model(HMM), hybrid ANN-DTW(Dynamic time warping), hybrid ANN-HMM, genetic algorithm, Support Vector Machines (SVM) etc[7]. In recent years, ANN based on back propagation(BP) or radial basis function(RBF) network has been widely used as a useful tool to the feature-level fusion modeling.

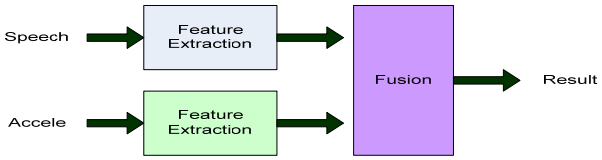

Fig 2. The fusion at feature level

In this paper, we propose a feature fusion method in an attempt to improve the performance of accelerometer-based hand activity recognition. To develop hand activity recognizer with high performance, we present a modified TDNN architecture with a dedicated fusion and time normalization layer.

2 Accelerometer-based Hand Activity and Voice Representation

In this section, details of hand activity and the feature extraction module of speech and acceleration used in this paper are described.

2.1 Speech Feature Extraction

The Speech Feature Extraction(SFE) module extracts feature vectors from the speech signal. This module is comprised of the following components : an End Point Detection(EPD) module based on the frame energy, a Feature Extraction(FE) module based on Zero-Crossing with Peak Amplitude (ZCPA)[8] and RelAtive SpecTrAl algorithm (RASTA)[9].

The ZCPA model is more robust in noisy environments than other popularly used feature extraction methods, such as LPCC or MFCC. It is composed of cochlear bandpass filter, zero-crossing detector and peak detector. And frequency information is obtained by the zero-crossing detector, and intensity information is also incorporated by the peak detector. RASTA processing of speech is a bandpass modulation filtering, operating on the log spectral domain. Slow channel variations should in principle be removed. Finally, the SFE module captures 16 features per frame.

2.2 Accelerometer-based Hand activity Feature Extraction

The acceleration data of the subject was collected using two MTx(Xsens technologies) accelerometers. These 3-axis accelerometers are accurate to ± 1.7G with tolerances within 0.2%. The accelerometers were mounted on wrist and sampled at 100Hz. Figure 3 shows MTx accelerometer with 3-axis and the attached type on the wrist.

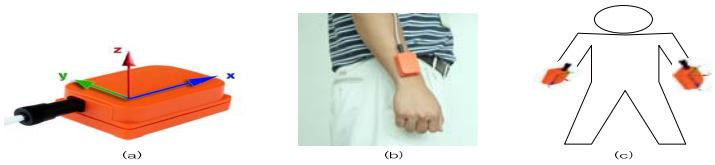

Fig. 3. (a) MTx accelerometer with 3-axis (b) MTx were attached (c) Accelerometers were mounted on wrist

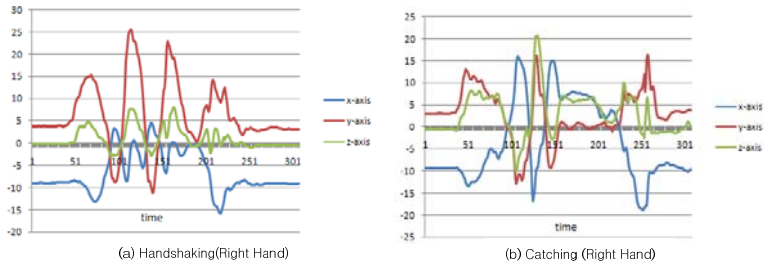

Fig. 4. Acceleration signals from 3-axis accelerometer for right hand moving : (a) Handshaking, (b) Catching

Figure 4 shows acceleration signal from 3-axis accelerometer for the movement of the right hand. As shown in figure 4, (a) is acceleration signal from MTx on right wrist for 'Handshaking' activity and (b) is acceleration signal for 'Catching' activity. In figure 4, the blue, the red, and the green lines are acceleration signal of x-axis, y-axis and z-axis direction respectively. From the graphs, it is trivial to conclude that the acceleration signal of each axis is highly sensitive to any movements of hand.

The Accelerometer-based Activity Feature Extraction (AAFE) module extracts feature vectors from 3-axis accelerometer signal on two wrists. This module is consists of two components : a Start Point Detection(STD) module based on threshold and a Feature Extraction module using the difference between accelerations of each axis. The STD module detects using a base signal on the sum of signal differences over a 10 frame window.

After finding the start point of activity, AAFE module extracts feature vectors using the difference between accelerations of each axis. The six coefficients are computed at every 10ms and fed to the fusion classifier as an input. The feature vectors are obtained as,

$$\begin{aligned} \mathrm{Lx}_t' &= |x_{t+1} - x_t|, \quad \mathrm{Ly}_t' = |y_{t+1} - y_t|, \quad \mathrm{Lz}_t' = |z_{t+1} - z_t| \\ \mathrm{Rx}_t' &= |x_{t+1} - x_t|, \quad \mathrm{Ry}_t' = |y_{t+1} - y_t|, \quad \mathrm{Rz}_t' = |z_{t+1} - z_t| \end{aligned} \tag{2}$$

where x_t, y_t and z_t are acceleration of each axis at time t. R and L is accelerometer sensor on the left and right hand.

2.3 Specific of Hand Activity and Speech Data

In our experiment, we examine natural gestures and emotional expressions by hand activity. A natural gesture is defined as an action that everyone can understand in human-to-human communication. The supported natural gestures are as follows: 'go right', 'go left', 'go up', 'go down', 'over here', 'go away', 'catch', 'release', 'open' and 'close'.

So an emotional expression by hand activity is defined as an action of hand according to the emotion change such as ‘I love you’, ‘I feel cold’, ‘I feel hot’, ‘I feel so-cold’, ‘I feel so-hot’, ‘I feel real-hot’, ‘I feel real-cold’, ‘Handshaking’ and ‘Good-bye’. But the linguistic definitions of emotional expression are often ambiguous. To address ambiguities in hand activity labels, test subjects were provided with image descriptions of each hand activity and short sentence descriptions. Figure 5 shows an example of image descriptions of each hand activity. Table 1 lists Korean utterance of each hand activity along with its short sentence description.

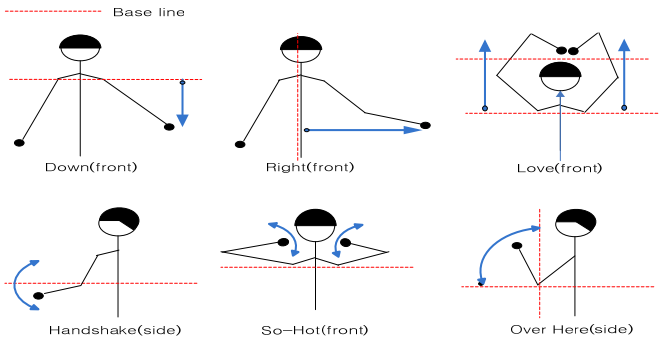

Fig. 5. An example of image descriptions of hand activities

Table 1. Descriptions of utterance and hand activity

Utterance (Korean-English)	Hand activity (short sentence descriptions)
Jap-A (catch)	Catching
Noh-A (release)	Releasing
Dad-A (close)	Closing
Yeol-Eo (open)	Opening
I-Ri-Wa (over here)	Over here
Jeo-Ri-Ka (go aware)	Go away
A-Rea (go down)	Go Down
Wi (go up)	Go Up
O-Reun-Jok (go right)	Go Right
Woen-Jok (go left)	Go Left
Jal-Ga (say good-bye)	Good-bye gesture
Ban-Ga-Ueo-Yo(handshake)	Give a person the right hand of fellowship
Sa-Rang-Hae-Yo(I love you)	Making heart figure by raising two hands
Jin-Ja-Deb-Da(real-hot)	Waving the collar back and forth toward face
Ne-Mu-Deb-Da(so-hot)	Waving two hands back and forth toward face
A-Deb-Da (hot)	Roll up one’s sleeves
O-Chub-Da(real-cold)	Rub one’s ear using hand
Chub-Da(cold)	Rub one’s hands
A-Chub-Da(so-cold)	Stamping one’s feet

3 Modified TDNN Architecture for Data Fusion

In 1989, it was shown that neural network model yields a high performance in the speech recognition. The main goal of TDNN was to have a neural network architecture for non-linear feature classification invariant under translation in time or space. TDNN uses time-delay steps to represent temporal relationships. The translation invariant classification is realized by sharing the connection weights of the time delay steps[10]. The activation of a unit is normally computed by passing the weighted sum of its inputs to an activation function(i.e. a sigmoid function).

We explain modified TDNN architecture to improve performance of hand activity recognition system. One of most difficult challenges in the feature-level fusion is the synchronization between accelerometer and speech data. In our system, speech features are extracted with the dimensions 64x16(where 64 is the number of frames and 16 is the number of coefficients). So the accelerometer-based hand activity features are extracted with the dimensions 120 x 6. Therefore, the method chosen to synchronize between the two feature spaces has a significant effect on the improvement of the recognition rate. We solve the synchronization problem by using a dedicated fusion layer and time normalization layer. Figure 6 is the modified TDNN architecture for data fusion.

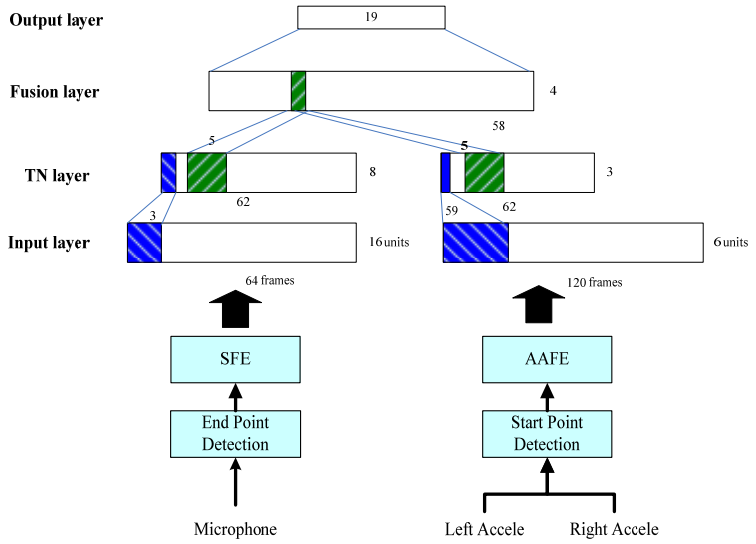

Fig. 6. Architecture of the parameter optimized TDNN for data fusion

As shown in Figure 6, fusion module is comprised of four layers-Input layer, Time Normalization(TN) layer, Fusion layer and Output layer. The TN layer controls the number of input frames to fusion layer. We use the TN layer, the output of each node

at fusion layer is given as.

$$z'_{F_j} = f\Big(\sum_{S_i=F_j}^{(N_S+F_j-1)} w_{S_iF_j} z_{S_i} + \sum_{A_i=F_j}^{(N_A+F_j-1)} w_{A_iF_j} z_{A_i}\Big) \tag{3}$$

where F_j is the index of node at fusion layer, A shows the hand activity features. S denotes the speech feature and f is a sigmoid fusion. So N is the number of windows at TN layer, i is the index of node at TN layer, j is the index of node at fusion layer. z is the output value of TN layer, z` is the output value of fusion layer and w is weight.

In figure 6, The input layer of the Speech Network(SN) has 16 feature values at each 10ms interval, 64 frames, and overlap windows of 3 frames. And the input layer of the Accelerometer-based Activity Network(AAN) has 6 feature values at each 10ms, 120 frames, and overlap windows of 59. So the TN of the SN consists of 62 frames, 8 units per frame and overlap windows of 5 frames. And the TN of AAN consists of 62 frame, 3 units per frame and overlap windows of 5 frame. In the case of the SN, the 48 units of this input layer are fully interconnected to a layer of 8 TN units. In the case of the ANN, the 354 units of the input layer are fully interconnected to a layer of 3 TN units. Finally, the fusion layer consists of 58 frames and 4 units per frame.

4 Experimental result

The experimental data consists of 19 utterances recorded by a male. The subject is directed to perform each activity at a time accompanied by corresponding speech in a quiet office environment. For training, 50 sets of activity and speech data are used, and the other 50 sets are used as test patterns. Speech is recorded by a SHURE microphone and the accelerometer-based activity is recorded by two MTx on his wrists. To train gesture and speech, the data set is provided to the system at the learning rate of 0.1.

Table 2 compares the performance of the proposed fusion system to the system that uses accelerometer alone. In table 2, when the signal-to-noise ratio(SNR) decreases, the Fusion method does not degrade as much as Accelerometers alone case.

Table 2. Comparison of recognition rate at various SNR(using white noise)

SNR	Accelerometers alone	Speech alone	Fusion
-5dB		97.15	99.26
-10dB	92.3	89.47	99.05
-15dB		53.68	96.21

As shown in Table 2, our system show a performance improvement of 6.96% at -5dB, 6.75% at -10dB and 3.91% at -15dB when compared to the use of accelerometer alone.

4 Conclusion

An accelerometer is one of the most useful wearable sensors for activity recognition. The accuracy of the previous work only using accelerometer was around 85% ~90%, which may be good enough for some applications. In this paper, we present a multisensory-based fusion recognition system having more enhanced performance than accelerometer-based activity recognition. In our work, we designed a hand activity recognizer that can classify acceleration data and utterance into nineteen activities : 10 natural activities and 9 emotional expressions by hand activity.

To improve performance of hand activity recognition system, we propose the modified TDNN architecture with a dedicated fusion layer and time normalization layer. Using the proposed fusion layer, we solved the synchronization problem in feature-level fusion. Our experiment shows performance improvement of 6.96% when compared to an activity system using only accelerometer.

References

1. Jeonghwa Yang, B. N. Schilit, and D.W. McDonald, "Activity Recogni- tion for the Digital Home," *in Computer*. vol. 41, pp. 102-104, April 2008.
2. L. Bao, and S. Intille, "Activity Recognition from user-Annotated Acceleration Data," In *Proc. Pervasive*, pp. 1-17, April 2004.
3. N. B. Bharatula, M. Stager, P. Lukowicz, and G. Troster, "Power and Size Optimized Multisensor Context Recognition Platform,"In *ISWC 2005*, pp. 194-195, Oct. 2005.
4. K. Kiani, C. J. Snijders and E. S. Gelsema, "Computerized analysis of daily life motor activity for ambulatory monitoring," Technol. Health Care vol. 5, pp. 307-318 Oct. 1997.
5. S. Lafon, Y. Keller, and R. R. Coifman, "Data Fusion and Multicue Data Matching by Diffusion Maps," *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol 28, pp. 1784-1797, Nov. 2006.
6. D. Ruta, and B. Garbrys, "An Overview of Classifier Fusion Method," *Computing and Information Systems*, vol. 7, pp. 1-10 February 2000.
7. J. W. Zhang, L. P. Sun, and J. Cao, "SVM for Sensor Fusion-a Comparison with Multilayer Perceptron Networks," In Proc. Machine Leaqrning and Cybernetics, pp. 2979-2984, Aug. 2006.
8. J. Young-Giu, H. Mun-Sung, and L. San Jo, "Development of an Optimized Feature Extraction Algorithm for Throat Signal Analysis," *ETRI Journal*, vol. 29, pp. 292-299, June 2007
9. H. Hermansky, and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech Audio Processing*, vol 2, pp. 578-589, Oct. 1994
10. N. Mache, M. Reczko and A. Hatzigeorgiou, "Multistate Time-Delay Neural Networks for the Recognition of Pol II Promoter Sequences," In *Proc. 10th Conf. Intelligent Systems for Molecular Biology,* St. Louis, 1996