31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Dynamic Gesture Recognition based on Multiple Sensors Fusion Technology

Wang Wenhui, Chen Xiang, Wang Kongqiao, Zhang Xu, Yang Jihai

Abstract—**This paper investigates the roles of a three-axis accelerometer, surface electromyography sensors and a webcam for dynamic gesture recognition. A decision-level multiple sensor fusion method based on action elements is proposed to distinguish a set of 20 kinds of dynamic hand gestures. Experiments are designed and conducted to collect three kinds of sensor data stream simultaneously during gesture implementation and compare the performance of different subsets in gesture recognition. Experimental results from three subjects show that the combination of three kinds of sensor achieves recognition accuracies at 87.5%-91.8%, which are higher largely than that of the single sensor conditions. This study is valuable to realize continuous and dynamic gesture recognition based on multiple sensor fusion technology for multi-model interaction.**

I. Introduction

Vision-based, accelerometer (ACC)-based, and surface electromyography (EMG)-based techniques are three research branches in the field of hand gesture pattern recognition [1], [3], [5]. Vision-based methods claim to allow the user to perform gestures freely without any instrumentation attached to the body. But they suffer from temporal and spatial segmentation [4] and environmental changes. ACC-based technique is well suited to distinguish noticeable, larger scale gestures with different hand trajectories of forearm movements. EMG sensors have advantages in measuring wrist and subtle fingers movements. However, due to some problems inherent in the EMG measurements, including the separability and reproducibility of measurement, measuring the level of muscle contraction makes the number of discriminable gestures limited to 4-8 classes [7].

As mentioned above, different techniques have their own advantages and shortcomings for hand gesture recognition. Different sensors can capture similar and complementary information when a hand gesture is performed. For instance, vision and acceleration signals can depict motion trajectories in space from two different views, and EMG can provide subtle finger movements information by detecting muscles contraction, which is difficult to obtain from vision and acceleration signals. Therefore, we believe that multiple sensor technology fusing the information from different kinds of sensor could improve the recognition accuracies and the number of discriminable hand gestures.

This paper aims at exploring the improvement effect of multiple sensor fusion technology for dynamic gesture recognition. To reach the research goal, approaches based on EMG, ACC and vision signals using different classification strategies are described firstly. Then twenty hand gestures are defined and data acquiring schemes are designed and carried out to collect the three kinds of sensor data synchronously. And classification results using single sensor data and the combinations of multiple sensors are analyzed and compared. Finally, conclusions are given together with suggestions for future work.

II. Related Work

Many techniques based on vision, ACC or EMG signals have been used for gesture or sign language recognition. For vision-based technology, Starner and Pentland [1] developed a real-time system recognizing sentence-level American Sign Language (ASL) generated by 40 words. From a desk mounted camera, word accuracies achieved 91.9% with a strongly grammar and 74.5% without grammar respectively. Brashear and his colleagues [2] built a lab-based sign language recognition system using vision data and ACC data. The system was composed of a head-mounted camera and wireless accelerometers mounted on the wrist in a colored bracelet. 5-gesture set of words plus a begin gesture and an exit gesture were recognized with 90.48% accuracy. A real-time system with multi-channels EMG developed by Hargrove and his colleagues classified 7 classes with high accuracies [3]. Our pilot work [8] utilized one accelerometer and four EMG sensors classified 18 gestures and an average accuracy about 91.7% was achieved in real application.

Because the spatio-temporal variability of vision-based gesture makes the segmentation of vision gesture very difficult, many techniques about temporal segmentation of video gesture have been reported. McGuire et al. [5] used buttons to segment data manually for capturing meaningful sign language gestures. The signer pressed one button to begin capturing and pressed another to stop data gathering. In the work by Brashear and Starner et al. [2], they used a begin gesture and an exit gesture to start or stop the gesture data

Manuscript received April 23, 2009. This work was supported in part by National Nature Science Foundation of China (NSFC) under Grant 60703069, National High Technology of Research and Development Program of China (863 Program) under Grant 2009AA01Z322.

W.H. Wang, X. Zhang are with the Institute of Biomedical Engineering at the University of Science and Technology of China, Hefei China, (e-mail: wwh9712@mail.ustc.edu.cn).

X. Chen and J.H. Yang are professors of the Institute of Biomedical Engineering at the University of Science and Technology of China, Hefei China, (e-mail: xch@ustc.edu.cn).

K.Q. Wang is with Nokia Research Center, NOKIA (CHINA) Investment CO., LTD., Beijing, China (e-mail: Kongqiao.Wang@nokia.com).

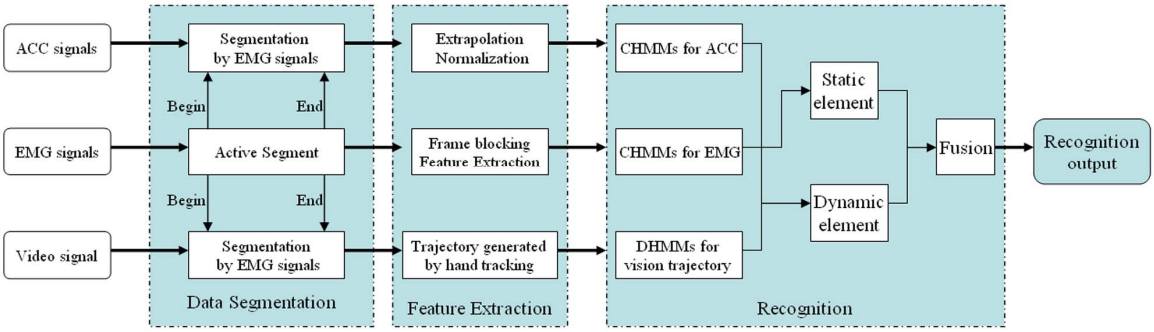

Fig. 1. The flow diagram of gesture recognition

collection. Our work in this paper significantly differs from others in the following respects:

1) The roles of three kinds of sensor (ACC, EMG and vision) and their combinations for dynamic gestures recognition are investigated.

2) A decision-level multiple sensor fusion method based on action elements are proposed for dynamic hand gestures recognition.

3) In order to acquire synchronous signals in the experiments automatically, the intensity of EMG signals was utilized to determine active segments of EMG, ACC and video signals.

III. METHODOLOGY

As Fig. 1 shows, the whole flow diagram of gesture recognition approach based on three kinds of sensor data consists of three main phases: the gesture action segmentation phase, the feature extraction phase, and the classification phase.

A. Gesture action segmentation

The start and end points of a meaningful gesture action called active segment must be specified firstly from continuous streams of input signals before feature extraction. The task of active segments extraction is very difficult due to gesture ambiguity [4].

Because EMG signals represent directly the level of muscle activities, the amplitudes of EMG are not sensitive to the withdrawal procedure of the hand gestures. To alleviate temporal segmentation difficulty and make gesture action segmentation easy, only EMG signals are used for active segments extraction. And the following 64-points moving average algorithm is applied.

First, the average absolute value of three-channel EMG signals at time t is computed according to Eq.1 (S is the number of EMG channels). And the average of the squared sums of the average absolute values is calculated with a moving window size of 64 sample points (see Eq.2).

$$EMG_{average}(t) = \frac{1}{S}\sum_{s=1}^{S}\left|EMG_s(t)\right| \quad (1)$$

$$EMG_{MA}(t) = \frac{1}{W}\sum_{i=t-W+1}^{t} EMG^2_{average}(i) \quad (2)$$

Then two thresholds (named as onset and offset) adjusted empirically are used respectively for determining the beginning and the ending of a gesture. When $EMG_{MA}(t)$ value reaches up to the onset threshold, it is thought that a gesture starts. When all sample points in 100ms have below the offset threshold, the gesture ends. A problem worthy to mention is that the offset threshold is lower than the onset threshold. The higher onset threshold can avoid mistaking trembling, while the lower offset threshold can prevent fragmentation during the gesture execution.

Finally, the vision and ACC signals are segmented synchronously with the start and end points determined by EMG signals. During the process of acquiring active segments, hand tracking is implemented. The skin-like regions in the image are detected from the hue-saturation-value (HSV) color space representation. CamShift algorithm combined with a Kalman filter is applied for hand tracking [6].

B. Feature Extraction

There are three types of features, i.e. features extracted from EMG, ACC and vision signals, used to differentiate the gesture classes.

For an active segment, ACC data is linearly re-sampled into 32 points, and then normalized and scaled as feature vector. So 3D ACC data of an active segment is represented by a 3x32 feature vector sequence. Before EMG feature extraction, overlapped windowing technique [8] is applied firstly to block EMG signal of an active segment into M frames of 256ms at every 128ms. Then the first three coefficients of forth-order Autoregressive (AR) model and Mean Absolute Value (MAV) are used to form the feature vector of each windowed frame. Hence, N-channel EMG signals of an active segment are represented by a 4 x N x M feature vector. With the purpose of simplification, we only calculate projected trajectory of vision signals in 2D plane. From hand tracking, hand centroids of motion video sequences are obtained. Then the angle of centroids between adjacent frames is calculated. The angles are quantified into 16 grades. Thus, the feature of visual trajectory of hand gesture is 1-dimensional vector represented by a set of discrete value ranging from 1 to 16.

C. Classification

Since the considered gestures are dynamic processes, HMM is an effective tool which can make use of temporal

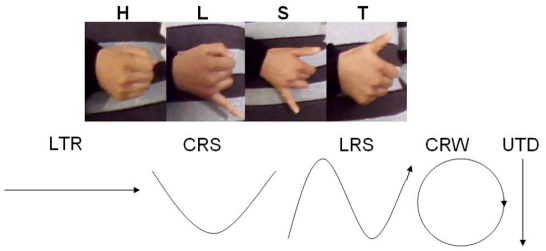

Fig. 2. Four static elements (H, L, S, T) and five dynamic elements (LTR, CRS, LRS, CRW, UTD)

characteristic of dynamic gestures. A five-state left-right Bakis HMM with self transitions and 1 skip state is used for representation of gestures. Three types of feature are represented by the same HMM topology. The normal Baum-Welch re-estimation method [9] is implemented to train HMM. Corresponding to the feature representation of each sensor data, continuous HMMs (CHMM) are used for modeling EMG and ACC, while discrete HMM (DHMM) for vision trajectory. As for CHMM, we employ 3 Gaussian Mixture Models as the emission probability of observation.

In this paper, a decision-level fusion method based on action elements is proposed for multiple sensor data classification.

According to the characteristics of dynamic gesture, we apply concept of action elements instead of whole gesture as the basic units for gesture recognition. A gesture action is decomposed into two types of action elements: static (finger configuration or certain pose) and dynamic (hand movement in the space) elements. EMG containing rich information about subtle finger movements is used to describe static element. ACC and visual trajectory, which depict the changing way in the position or orientation of the gesture movement, are used to represent dynamic element.

In classification phase, static element of the input gesture is classified firstly to a static element class which has the maximal likelihood corresponding to the EMG feature. Then the logarithmic likelihood of the input gesture belonging to the *c*th dynamic element class is calculated using multi-stream HMMs [8] according to Eq.3:

$$P(O_t \mid \lambda_c) = \delta_{ACC} P(O_t^{ACC} \mid \lambda_c^{ACC}) + \delta_{vision} P(O_t^{vision} \mid \lambda_c^{vision}) \quad (3)$$

Here t is the time and c is the class of dynamic elements. δ_{ACC} and δ_{vision} are ACC and visual trajectory weight factors respectively and satisfy the following restriction:

$$\delta_{ACC} + \delta_{vision} = 1,\ 0 \le \delta_{ACC}, \delta_{vision} \le 1 \quad (4)$$

The recognition result for dynamic element of the input gesture is the class which achieves the highest logarithmic likelihood.

Given static element set and dynamic element set are represented by P and Q respectively. There is an example of dynamic gesture classification: if posture is similar to hand grasp (condition p_1 of static element set P), the trajectory of hand gesture is like a circle (condition q_1 of dynamic element set Q), then the type of dynamic gesture is A, i.e. $p_1 \Theta q_1 \rightarrow A$. Here, the symbol Θ indicts the combination operator.

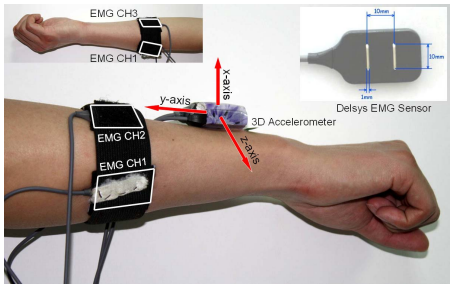

Fig. 3. The sensor placement for data acquisition system

IV. Experiment and Result Analysis

A. Experimental Setting

Data Acquisition: In order to record three kinds of sensor data synchronously, the acquisition system consisting of 3 surface EMG sensors, a 3-axis accelerometer and a webcam was built. The EMG and ACC signals are acquired by Delsys Myomonitor IV sensor system with inbuilt 20-1000Hz band-pass filters and 60dB-gain amplifiers. The ACC and EMG signals are digitized by a PCI Acquisition Card (PCI-6010 by NI) with 16-bit input resolution and 1 kHz sampling rate respectively. The camera is a webcam feeding video back to computer via USB. The video of gesture is captured with a spatial resolution of 640×480 pixels, at a rate of 30 frames per second. Each gesture-making takes about 1-3 seconds.

Gesture definition and sensor placement: In order to evaluate the improvement effect of multiple sensor fusion technology for gesture recognition, 20 dynamic gestures composed of 4 static elements and 5 dynamic elements were defined (shown in Fig. 2). To capture information of hand gesture exertion, following sensors placement scheme was adopted. As Fig. 3 shows, three EMG sensors were fixed around the circumference of the forearm approximately one-half of the distance from the elbow to the wrist. Two EMG sensors were fixed on the posterior aspect and one on the anterior aspect of the forearm. The 3D accelerometer was placed on the back of forearm near the wrist, and the camera was put beside the PC.

Subjects: Three healthy subjects (1 male and 2 females), with ages ranging from 20 to 25 years, were recruited for the data collection. In the experiments, the subjects, who all used their right hand to implement hand gesture, were requested to face to the camera with about 1m distance. Then each subject performed 20 defined gestures in a way felt natural. Each subject executed gestures five repetitions and each gesture was repeated about 10 times per repetition.

B. Experimental Results and Analysis

Forty gesture samples per each person were used to train and the rest were used for testing. The cross-validation was used to avoid the influence of the choice of training samples. In order to explore the roles of three kinds of sensor data and their combinations, the recognition results (given as mean

TABLE I
AVERAGE CLASSIFICATION RESULTS OF SINGLE SENSOR CONDITIONS

Conditions	Sub1		Sub2		Sub3	
	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)
EMG-only (CHMM)	34.70	19.26	47.17	24.36	71.40	16.01
ACC-only (CHMM)	41.90	18.86	52.67	27.52	55.00	20.77
vision-only (DHMM)	36.60	12.68	28.50	20.13	35.30	19.62

TABLE II
AVERAGE CLASSIFICATION RESULTS OF COMBINED SENSORS CONDITIONS

Conditions	Sub1		Sub2		Sub3	
	Mean (%)	Std (%)	Mean (%)	Std (%)	Mean (%)	Std (%)
EMG+ACC(A)	86.0	11.82	87.67	13.89	78.60	17.29
EMG+ACC(B)	89.8	9.33	89.33	10.46	80.50	16.39
EMG+vision(A)	64.5	16.66	67.50	24.35	65.60	20.36
EMG+vision(B)	85.6	10.85	56.33	30.98	74.10	17.96
Three-sensors(A)	87.5	11.01	81.83	17.88	82.20	12.99
Three-sensors(B)	91.8	8.48	90.67	9.77	87.50	11.27

and standard deviation over 20 defined hand gestures) of different experimental conditions for three subjects are given in Table I and II.

Table I shows the experimental results of hand gesture classification with data from single sensor. We can find that poor accuracies (from 28.5% to 71.4%) were obtained. Table II shows the recognition accuracies using multiple sensor data. To explore the capability of our proposed decision-level fusion classification method based on action elements, the classification results using multi-stream HMMs directly without action decomposition are also be given for comparison. In brief, classification method without action decomposition is represented with A, and our proposed method with B in Table II. We can find obviously from Table II that the accuracies are improved highly when classified with data from multiple sensors (overall from 56.33% to 91.80%), especially for the combination of EMG and ACC and the combination of three sensors conditions, the average recognition accuracies reach up to 78.60%-91.80%. These results demonstrate that multiple sensors can provide complementary classifiable information for defined dynamic gesture recognition.

Compared with the classification method without action decomposition (method A in Table II), the classification method based on action elements (method B in Table II) achieved higher rates in most of the tasks. Even the average classification rates of the combination of three kinds of sensor data using method A (81.83%-87.50%) are worse than that of the combination of EMG and ACC using method B (80.50%-89.80%). As we know, in the classification method based on action elements, the gesture is decomposed and the static element and dynamic element are used apart for recognition. These experimental results indicate that it is an effective way to improve the class separability by decomposing dynamic gestures into static and dynamic elements.

In experiments, we also found that there are also other advantages such as less training time cost and less recognition time delay in the implementation of the classification method based on action elements. We could easily understand this phenomenon because EMG is only used to classify 4 static elements classes in the proposed method, while 20 gesture classes in the method without action decomposition.

V. Conclusion and Future Work

The main work of this paper is to investigate multiple sensor fusion technology for dynamic gesture recognition. A classification method based on action elements is proposed for decision-level fusion of multiple sensor data. The experimental results show that the three kinds of sensor fusion condition achieve 87.5%-91.8% recognition accuracies, which is higher largely than that of single sensor conditions. These results demonstrate that multiple sensors can capture complementary classifiable information of the defined dynamic gestures, and multiple sensors fusion technology is a good choice for dynamic hand gesture pattern recognition.

Future work will be continued on the development of robust multiple sensor fusion algorithm for dynamic gesture recognition. And the multi-model interaction based on the multiple sensor fusion technique will also be investigated.

References

[1] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1371–1375, 1998.

[2] H. Brashear, T. Starner, P. Lukowicz and H. Junker, "Using Multiple Sensors for Mobile Sign Language Recognition", *Proceedings of the Seventh IEEE International Symposium on Wearable Computers*, pp. 45-52, 2003.

[3] L. Hargrove, Y. Losier, B. Lock, K. Englehart and B. Hudgins, "A Real-Time Pattern Recognition Based Myoelectric Control Usability Study Implemented in a Virtual Environment", *Processings of the 29th Annual International Conference of the Engineering in Medicine and Biology*, pp. 4842-4845, 2007.

[4] S. Mitra and T. Acharya, "Gesture Recognition: A Survey", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 37, pp. 311-324, 2007.

[5] R.M. McGuire, J. Hernandez-Rebollar, T. Starner, V. Henderson, H. Brashear and D.S. Ross, "Towards a One-Way American Sign Language Translator", *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 620-625, 2004.

[6] J. Zieren, N. Unger, and S. Akyol, "Hands Tracking from Frontal View for Vision-Based Gesture Recognition", *Lecture Notes in Computer Science*, pp. 531–539, 2002.

[7] M.A. Oskoei and H. Hu, "Myoelectric control systems—A survey", *Biomedical Signal Processing and Control*, Vol. 2(4), pp. 275-294, 2007.

[8] X. Zhang, X. Chen, W.H. Wang, J.H. Yang, V. Lantz, K.Q. Wang, "Hand Gesture Recognition and Virtual Game Control Based on 3D Accelerometer and EMG Sensors", *Proceedings of International Conference on intelligent User interfaces*, pp. 401-405, 2009.

[9] L. R. Rabiner. "A tutorial on Hidden Markov Models and selected applications in speech recognition". *Proceedings of the IEEE*, Vol. 77(2), pp. 257-286, 1989.