A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors

Xu Zhang, Xiang Chen, *Associate Member, IEEE*, Yun Li, Vuokko Lantz, Kongqiao Wang, and Jihai Yang

***Abstract*—This paper presents a framework for hand gesture recognition based on the information fusion of a three-axis accelerometer (ACC) and multichannel electromyography (EMG) sensors. In our framework, the start and end points of meaningful gesture segments are detected automatically by the intensity of the EMG signals. A decision tree and multistream hidden Markov models are utilized as decision-level fusion to get the final results. For sign language recognition (SLR), experimental results on the classification of 72 Chinese Sign Language (CSL) words demonstrate the complementary functionality of the ACC and EMG sensors and the effectiveness of our framework. Additionally, the recognition of 40 CSL sentences is implemented to evaluate our framework for continuous SLR. For gesture-based control, a real-time interactive system is built as a virtual Rubik's cube game using 18 kinds of hand gestures as control commands. While ten subjects play the game, the performance is also examined in user-specific and user-independent classification. Our proposed framework facilitates intelligent and natural control in gesture-based interaction.**

***Index Terms*—Acceleration, electromyography, hand gesture recognition, hidden Markov models (HMMs).**

I. Introduction

HAND gesture recognition provides an intelligent, natural, and convenient way of human–computer interaction (HCI). Sign language recognition (SLR) and gesture-based control are two major applications for hand gesture recognition technologies [1]. SLR aims to interpret sign languages automatically by a computer in order to help the deaf communicate with hearing society conveniently. Since sign language is a kind of highly structured and largely symbolic human gesture set, SLR also serves as a good basic for the development of general gesture-based HCI. In particular, most efforts [7]–[10] on SLR are based on hidden Markov models (HMMs) which are employed as effective tools for the recognition of signals changing over time. On the other hand, gesture-based control translates gestures performed by human subjects into controlling commands as the input of terminal devices, which complete the interaction approaches by providing acoustic, visual, or other feedback to human subjects. Many previous researchers [2]–[4], [11], [12] investigated various systems which could be controlled by hand gestures, such as media players, remote controllers, robots, and virtual objects or environments.

According to the sensing technologies used to capture gestures, conventional researches on hand gesture recognition can be categorized into two main groups: data glove-based and computer vision-based techniques [1], [2]. In the first case, data gloves equipped with bending sensors and accelerometers are used to capture the rotation and movement of the hand and fingers. Fang *et al.* [9] reported a system using two data gloves and three position trackers as input devices and a fuzzy decision tree as a classifier to recognize Chinese Sign Language (CSL) gestures. The average classification rate of 91.6% was achieved over a very impressive 5113-sign vocabulary in CSL. However, glove-based gesture recognition requires the user to wear a cumbersome data glove to capture hand and finger movement. This hinders the convenience and naturalness of HCI [1]. In the later case, computer vision-based approaches can track and recognize gestures effectively with no interference on the user [7], [8], [10]. Starner *et al.* [7] developed an impressive real-time system recognizing sentence-level American Sign Language generated by 40 words using HMMs. From a desk-mounted camera, word accuracies achieved 91.9% with a strong grammar and 74.5% without grammar, respectively. Shanableh *et al.* [8] employed a spatiotemporal feature extraction scheme for the vision-based recognition of Arabic Sign Language (ArSL) gestures with bare hands. Accuracies ranging from 97% to 100% can be achieved in the recognition of 23 ArSL-gestured words. Nevertheless, the performance of this technology is sensitive to the use environment such as background texture, color, and lighting [1], [2]. In order to enhance the robust performance of vision-based approaches, some previous studies utilized colored gloves [7] or multiple cameras [33] for accurate hand gesture tracking, segmentation, and recognition. The use conditions limit their extensive applications, particularly in mobile environment.

Unlike the approaches mentioned earlier, the accelerometer (ACC) and electromyography (EMG) sensor provide two potential technologies for gesture sensing. Accelerometers can measure both dynamic accelerations like vibrations and static accelerations like gravity. The ACC-based techniques have been successfully implemented in many consumer electronics models for simple and supplementary control application [2],

Manuscript received January 18, 2010; revised August 24, 2010; accepted October 18, 2010. Date of publication March 22, 2011; date of current version October 19, 2011. This work was supported in part by the National Nature Science Foundation of China under Grant 60703069 and in part by the National High-Tech Research and Development Program of China (863 Program) under Grant 2009AA01Z322. This paper was recommended by Associate Editor T. Tsuji.

X. Zhang was with the Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China. He is now with the Sensory Motor Performance Program, Rehabilitation Institute of Chicago, Chicago, IL 60611 USA.

X. Chen, Y. Li, and J. Yang are with the Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China (e-mail: xch@ustc.edu.cn).

V. Lantz is with the Multimodal Interaction, Nokia Research Center, 33720 Tampere, Finland.

K. Wang is with the Nokia Research Center, NOKIA (CHINA) Investment CO., LTD., Beijing 100013, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2011.2116004

[3], [17], [37]. For instance, the hand gesture recognition system of Mäntyjärvi *et al.* [2] was studied as an interesting mobile interaction for media player control based on a three-axis accelerometer. The EMG, which measures the electrical potentials generated by muscle cells, can be recorded using differential pairs of surface electrodes in a nonintrusive fashion, with each pair of electrodes constituting a channel of EMG [4], [5]. Multichannel EMG signals which are measured by EMG sensors placed on the surface skin of a human arm contain rich information about the hand gestures of various size scales. EMG-based techniques, which provide us with the significant opportunity to realize natural HCI by directly sensing and decoding human muscular activity [12], are capable of distinguishing subtle finger configurations, hand shapes, and wrist movements. For over three decades, EMG has been used as means for amputees to use residual muscles to control upper limbed prostheses [5], [18], [19]. Recently, the EMG-based hand gesture interaction for common users in daily life has attracted more and more attentions of most researchers. Costanza *et al.* [4] investigated EMG-based intimate interfaces for mobile and wearable devices. Their study demonstrated the feasibility of using isometric muscular activities as inputs to discreetly interact with devices in an unobtrusive manner. Wheeler *et al.* [11] described gesture-based control using EMG taken from a forearm to recognize joystick movement for virtual devices. Saponas *et al.* [12] used ten sensors worn in a narrow band around the upper forearm to differentiate the position and pressure of finger presses. Although previous studies on EMG-based HCI have attained relatively good results, it has a significant distance from commercial applications for fine control due to some problems, including the separability and reproducibility of EMG measurements [5].

Since each sensing technique has its own advances and capabilities, the multiple sensor fusion techniques can widen the spread of potential applications. Many previous studies indicated that the combined sensing approach could improve the performance of hand gesture recognition significantly [13], [14]. Sherrill *et al.* [6] have compared the performance of ACC-based and EMG-based techniques in the detection of functional motor activities for rehabilitation and provided evidence that the system based on the combination of EMG and ACC signals can be built successfully. Our pilot study [15] demonstrated that ACC and EMG fusion achieved 5%-10% improvement in the recognition accuracies for various wrist and finger gestures. More recently, Kim *et al.* [32] examined the complementary functionality of both sensors in German Sign Language recognition for seven isolated words. Kosmidou and Hadjileontiadis [34] successfully applied the intrinsic mode entropy on ACC and EMG data acquired from the dominant hand to recognize isolated 60 Greek Sign Language signs. Aside from the information complementary characteristics, ACC and EMG sensors have some common advantages such as the low-cost manufacture and high portability for hand gesture capture. They can be easily worn on the forearm when used for HCI implementation. However, the ACC and EMG fusion technique for hand gesture recognition is still in the initial stage, and there is great potential for exploration.

As for intelligent interaction, it is important to automatically specify the start and end points of a gesture action [1]. However, most of the previous work has taken this for granted or accomplished it manually [2], [14], [17]. When performing gestures, the hand must move from the end point of the previous gesture to the start point of the next gesture. These intergesture transition periods are called movement epenthesis. The detection of movement epenthesis within a continuous sequence of gestures is often regarded as one of the main difficulties in continuous gesture recognition [35]. It is easy and natural to detect muscle activation with EMG sensors, which help to indicate meaningful gestures. In our method, the start and end points of gestures are detected automatically by the intensity of EMG signals, and then, both ACC and EMG segments are acquired for further processing.

The main contributions of this paper that significantly differ from others are as follows: 1) proposing a framework of hand gesture recognition using decision trees and multistream HMMs for the effective fusion of ACC and EMG sensors; 2) automatically determining the start and end points of meaningful gesture segments in the signal streams of multiple sensors based on the instantaneous energy of the average signal of the multiple EMG channels, without any human intervention, that can facilitate the relatively natural and continuous hand gesture recognition; and 3) conducting CSL recognition experiments with sentences formed by a 72-sign vocabulary and creating a prototype of an interactive system with gesture-based control to evaluate our proposed methods.

The remainder of this paper is organized as follows. Section II presents the framework for hand gesture recognition. Section III provides the experimental study on the recognition of CSL words and sentences to examine the proposed framework in continuous SLR. In Section IV, experiments on a virtual Rubik's cube game for gesture-based control are presented. The conclusions and future work are given in Section V.

II. Methodology

Fig. 1 shows the block diagram of our hand gesture recognition method using both multichannel EMG and 3-D ACC signals. The processing of the two signal streams is carried out in the following steps.

A. Data Segmentation

The multichannel signals recorded in the process of the hand gesture actions which represent meaningful hand gestures are called active segments. The intelligent processing of hand gesture recognition needs to automatically determine the start and end points of active segments from continuous streams of input signals. The gesture data segmentation procedure is difficult due to movement epenthesis [35]. The EMG signal level represents directly the level of muscle activity. As the hand movement switches from one gesture to another, the corresponding muscles relax for a while, and the amplitude of the EMG signal is momentarily very low during movement epenthesis. Thus, the use of EMG signal intensity helps to implement data segmentation in a multisensor system. In our method, only the multichannel EMG signals are used for determining the start and end points of active segments. The segmentation is based on a moving average algorithm and thresholding. The ACC signal stream is segmented synchronously with the EMG signal stream. Thus, the use of EMG would help the SLR system to

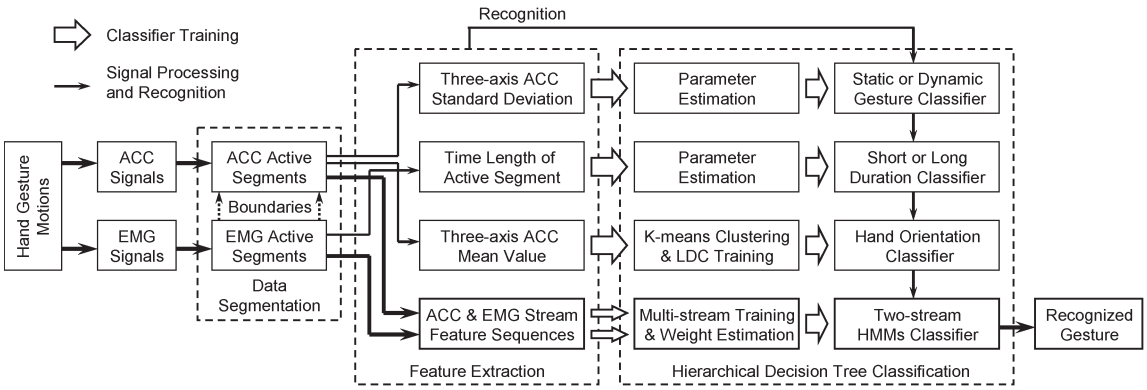


Fig. 1. Block diagram of proposed framework for hand gesture recognition.

automatically distinguish between valid gesture segments and movement epenthesis from continuous streams of input signals.

The detection of active segments consists of four steps based on the instantaneous energy of the average signal of the multiple EMG channels.

1) Computing the average value of the multichannel EMG signal at time t according to (1), where c is the index of the channel and N_c is the number of channels

$$EMG_{\text{avg}}(t) = \sum_{c=1}^{N_c} EMG_c(t). \tag{1}$$

2) Applying the moving average algorithm with a window size of $W = 60$ samples on the squared average EMG data to calculate the moving averaged energy stream $E_{MA}(t)$ according to

$$E_{MA}(t) = \frac{1}{W} \sum_{i=t-W+1}^{t} EMG^2_{\text{avg}}(i). \tag{2}$$

3) Detecting active segments using two thresholds, the onset and offset thresholds. Typically, the offset threshold is lower than the onset threshold. The active segment begins when $E_{MA}(t)$ is above the onset threshold and continues until all samples in a 100-ms time period are below the offset threshold. The higher onset threshold helps to avoid false gesture detection, whereas the lower offset threshold is for preventing the fragmentation of the active segment as $E_{MA}(t)$ may vibrate near the onset threshold during the gesture execution.
4) As the final step, abandoning the segments whose lengths are less than a 100-ms time period as measurement noise.

Hence, active gesture segments for both EMG and ACC signals are determined by the same boundaries.

B. Feature Extraction

1) Feature for ACC: The 3-D accelerometer measures the rate of change of velocity along three axes (x, y, z) when hand gestures are performed. Since the acceleration signals changing with time can directly represent patterns of hand gesture trajectories, the 3-D ACC active segments are scaled and extrapolated as feature vector sequences. The amplitude of the 3-D data in an active segment is scaled using a linear min–max scaling method. Then, the scaled ACC active segment is linearly extrapolated to 32 points so that the temporal lengths of all the 3-D ACC data sequences are the same. These two steps normalize the variations in the gesture scale and speed and thus improve the recognition of the type of the gesture [2], [17]. Normalized ACC active data segments are regarded as 3-D feature vector sequences as such.

In addition to the time-domain feature vector sequences as calculated earlier for ACC signals, we further extracted some statistical features, such as the mean value and standard deviation (SD) of each ACC axis. These simple features will be used by the following classifiers in a decision tree.

2) Feature for EMG: Various kinds of features for the classification of the EMG have been considered in the literature [23], [24]. These features have included a variety of time-domain, frequency-domain, and time–frequency-domain features. It has been shown that some successful applications can be achieved by time-domain parameters [19], for example, zero-crossing rate and root mean square (rms). The autoregressive (AR) model coefficients [25] of the EMG signals with a typical order of 4–6 yield good performance for myoelectric control. Many time–frequency approaches, such as short-time Fourier transform, discrete wavelet transform, and wavelet packet transform, have been investigated for EMG feature extraction [18]. However, time–frequency-domain features require much more complicated processing than time-domain features. Considering our pilot study [26], the combination of mean absolute value (MAV) and fourth-order AR coefficients as a feature set is chosen to represent the patterns of myoelectric signals with high test–retest repeatability.

In active segments, the EMG stream is further blocked into frames with the length of 250 ms at every 125 ms utilizing an *overlapped windowing technique* [19]. Each frame in every EMG channel is filtered by a Hamming window in order to minimize the signal discontinuities at the frame edges. Then, each windowed frame is converted into a parametric vector consisting of fourth-order AR coefficients and MAV. Hence, each frame of an n-channel EMG signal is presented by a $4n$-dimensional feature vector, and the active EMG segments are represented by $4n$-dimensional vector sequences of varying length. Additionally, the duration of the active segment is also regarded as an important statistical feature, which will be used by the following classifiers in a decision tree.

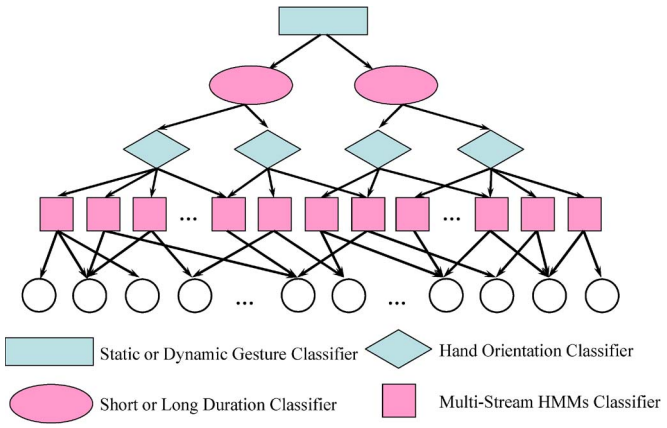


Fig. 2. Structure of decision tree for hand gesture recognition.

C. Tree-Structure Decision

A decision tree is a hierarchical tree structure consisting of a root node, internal nodes, and leaf nodes for classification based on a series of rules about the attributes of classes in nonleaf nodes, where each leaf node denotes a class [9]. The input sample data, including the value of different attributes, are initially put in the root node. By the rules in nonleaf nodes, the decision tree splits the values into different branches corresponding to different attributes. Finally, which class the input data belong to is assigned at the leaf nodes.

Decision trees are simple to understand and interpret. Their ability for diverse information fusion is well suited for pattern classification with multiple features. They also take advantage of their sequential structure of branches so that the searching range between classes for classification can be reduced rapidly. Decision trees are robust for good performance with large data in a short time [9], which is the significant advantage to realize real-time classification systems.

Fig. 2 shows the structure of the proposed four-level decision tree for hand gesture recognition, where each nonleaf node denotes a classifier associated with the corresponding gesture candidates and each branch at a node represents one class of this classifier. All the possible gesture classes form the gesture candidates of the root node, and then, the gesture candidates of a nonleaf node are split into the child nodes by the corresponding classifier of the parent node. For hand gesture recognition, unknown gesture data are first fed into a static or dynamic classifier, then into a short- or long-duration classifier, further into a hand orientation classifier, and, at last, into the multistream HMM classifier to get the final decision. The classifiers in each level of the decision tree are constructed as follows.

1) Static or Dynamic Gesture Classifier: The gestures can be static (a hand posture with a static finger configuration and an arm keeping a certain pose without hand movement) or dynamic (hand movement with a certain trajectory and finger motion). The three-axis SD of the ACC active segment can reflect the intensity of the hand or arm movements. Therefore, the rms value of the three-axis SD of the ACC active segment is compared with a certain threshold: If the value is lower than the threshold, it is considered a static hand gesture, and if higher, a dynamic gesture. The threshold is determined by the training samples of typical static gestures, such as the word "you," "good" in CSL, and hand grasping without arm movements. Usually, the threshold is assigned as the maximum of the rms value of the three-axis SD in these training samples.

After all the training samples are classified, the candidate gestures associated with static or dynamic gestures are generated, which will be used by the following short- or long-duration classifier.

2) Short- or Long-Duration Classifier: The time durations of gesture performance can be short (a simple posture) or long (a relatively complex posture or motion trajectory), which is a useful indicator to distinguish different gestures. A short- or long-duration classifier can be used as the supplementary classification of hand gestures with various attributes. Similar to the static or dynamic gesture classifier, the time-duration feature extracted from the EMG active segment is compared with a certain threshold: If the value is less than the threshold for the short, on the contrary, it is more than the threshold for the long. The threshold is determined by the training samples of typical short gestures, such as the word "good," "bull" in CSL, and hand grasping without arm movements. Usually, the threshold is assigned as the maximum of the time-duration value in these training samples of short gestures. However, those gestures that cannot be robustly determined will appear in both the candidate gestures of short gestures and those of long gestures.

3) Hand Orientation Classifier: The orientation of the hand can be described as the following two terms: 1) the direction toward which the hand and the arm are pointing and 2) the facing of the palm [9]. Since different hand orientations can cause the projection of gravity with different component values along three axes of the accelerometer, which is usually placed on the forearm near the wrist, the mean values of three-axis ACC active segments can effectively reflect the orientation of the hand for static hand gesture. Although the three-axis ACC mean features can be varied due to different movement patterns of dynamic hand gestures, these features for the same gesture are still consistent. Thus, in our method, the fuzzy K-means clustering and linear discriminant classifier (LDC) are proposed for the training and classification of the hand orientation classifier. The algorithms of the hand orientation classifier are described as follows.

Fuzzy K-means Clustering: In fuzzy clustering, each element has a degree of belonging to clusters, called as fuzzy membership degree, rather than completely belonging to just one cluster [27]. In statistical pattern recognition, fuzzy K-means clustering is a method of cluster analysis which aims to partition several finite elements into K clusters in which each element belongs to the cluster with the highest fuzzy membership degrees.

Given a set of elements $(\mathbf{g}_1, \mathbf{g}_2 \ldots \mathbf{g}_n)$, where each element is a three-axis ACC mean feature vector and n is the number of all the training samples, for each element $\mathbf{g}_j$, there is a fuzzy membership degree of being in the kth cluster $\hat{P}(\omega_k|\mathbf{g}_j)$

$$\hat{P}(\omega_k|\mathbf{g}_j) = \frac{(1/d_{kj})^{1/(b-1)}}{\sum_{i=1}^{K}(1/d_{ij})^{1/(b-1)}} \tag{3}$$

where ω_k denotes the kth cluster, d_{kj} denotes the Euclidean distance between the element $\mathbf{g}_j$ and the centroid of the kth cluster

$\boldsymbol{\mu}_{jk}$: $d_{kj} = \|\mathbf{g}_j - \boldsymbol{\mu}_k\|$, and the free parameter b is chosen to normalize and control the fuzzy degree of the algorithms. In our method, b is kept constant to the value of 1.5 for allowing each pattern to belong to multiple clusters. Then, the sum of those fuzzy membership degrees for any $\mathbf{g}_j$ is defined to be 1

$$\forall \mathbf{g}_j, \qquad \sum_{k=1}^{K} \hat{P}(\omega_k|\mathbf{g}_j) = 1. \tag{4}$$

The centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster

$$\boldsymbol{\mu}_k = \frac{\sum_{j=1}^{n} \hat{P}(\omega_k|\mathbf{g}_j)^b \mathbf{g}_j}{\sum_{j=1}^{n} \hat{P}(\omega_k|\mathbf{g}_j)^b}. \tag{5}$$

In the clustering approach, the initial centroids are randomly selected from the basic hand orientations determined by the experts from the hand gesture dictionary. Then, the fuzzy K-means clustering algorithm is employed to update the centroids in the training set according to (3) and (5). A set of new centroids is obtained after iterating the aforementioned process until the centroids of all clusters are not changed. Hence, the three-axis ACC mean features of all the training samples are assigned to the cluster whose fuzzy membership is the highest. Each resulting cluster denotes one pattern branch which indicates a kind of hand orientation, respectively. The candidate gesture set associated with each corresponding pattern is generated as the classes which the training samples in the cluster belong to. The candidate gesture set of each pattern branch will be used by following the multistream HMM classifier.

LDC Training: In order to determine the pattern branch of input data, the LDC is used in this low-dimensional space for hand orientation classification after the clustering process. The LDC is a probabilistic classifier based on applying Bayes's theorem with strong independence assumptions [28]. The probability model for the LDC is a conditional model in which an *a posteriori* probability function of ω_k given an input three-axis ACC mean feature $\mathbf{g}$ is defined as

$$P(\omega_k|\mathbf{g}) = \frac{P(\mathbf{g}|\omega_k)P(\omega_k)}{P(\mathbf{g})}. \tag{6}$$

The training of the LDC involves the estimation of the conditional probability density function for each class (or cluster). In our method, the within-class densities are modeled as normal distributions

$$P(\mathbf{g}|\omega_k) = (2\pi)^{-N/2}|\boldsymbol{\Sigma}_k|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{g}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{g}-\boldsymbol{\mu}_k)\right\} \tag{7}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of class ω_k, respectively. In our method, $\boldsymbol{\mu}_k$ is directly assigned as the centroid of the kth cluster, and $\boldsymbol{\Sigma}_k$ is calculated by the training samples belonging to the kth cluster after the fuzzy K-means clustering

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k - 1} \sum_{\mathbf{g}_j \in \omega_k} (\mathbf{g}_j - \boldsymbol{\mu}_k)(\mathbf{g}_j - \boldsymbol{\mu}_k)' \tag{8}$$

where n_k is the number of training samples belonging to the kth cluster.

LDC Classification: The maximum *a posteriori* decision rule for the LDC is

$$r = \arg\max_k \{P(\mathbf{g}|\omega_k)P(\omega_k)\} \qquad \mathbf{g} \in \omega_r. \tag{9}$$

The element $\mathbf{g}$ is classified to ω_r of whom an *a posteriori* probability given $\mathbf{g}$ is the largest among all the other classes. Hence, the unknown input gesture is classified into the corresponding pattern branch through the LDC.

4) Multistream HMM Classifier: Along the branch assigned by the hand orientation classifier, the unknown gesture sample is fed into the multistream HMM classifier with its ACC and EMG feature vector sequences. The final decision is determined among the candidates of this multistream HMM node.

Multistream Formalism: The multistream structure has the advantage that it can effectively combine several information sources, namely, feature streams, using cooperative Markov models. According to the multistream formalism [36], a hand gesture to be recognized is represented by an observation sequence $\mathbf{O}$, which is composed of K input streams $\mathbf{O}^{(k)}$. Moreover, each hypothesized model λ is composed of K models $\lambda^{(k)}$ attached to each of the K input streams. For the information fusion, the K stream models are forced to recombine using some proper recombination strategies.

Based on the Bayes theorem, the recognition problem can be directly formulated as the one of finding the gesture model λ^* that achieves the highest likelihood for the given observation sequence $\mathbf{O}$

$$\lambda^* = \arg\max_{\lambda \in \theta} P(\mathbf{O}|\lambda) \tag{10}$$

where θ is the set of all possible gesture hypotheses.

In order to determine the best gesture model λ^* that maximizes $P(\boldsymbol{O}|\lambda)$, three recombination strategies have been investigated in the literature [36].

1) Recombination at the HMM state level: Assuming that strict synchrony exists among the streams, it does not allow for asynchrony or different topologies of the stream models. In this case, the observation log-likelihood at each state is often calculated as the sum (or weighted sum) of the stream observation log-likelihoods [21], [22], [30], [31].
2) Recombination at the stream model level: Assuming that each stream is independent, it can allow for asynchrony or different topologies of the stream models. The streams are forced to be synchronous at the end of the gesture models [36]. It is really simple to perform a standard HMM algorithm to build each stream model separately based on single-stream observations.
3) Recombination by the composite HMM: It can be regarded as the integration of the aforementioned two strategies. Each state of the composite HMM is generated by merging a k-tuple of states from the K stream HMMs [36]. The topology of this composite model is defined so as to model multiple streams as a standard HMM. However, it requires an additional processing to build the composite HMM. When dealing with multiple streams,

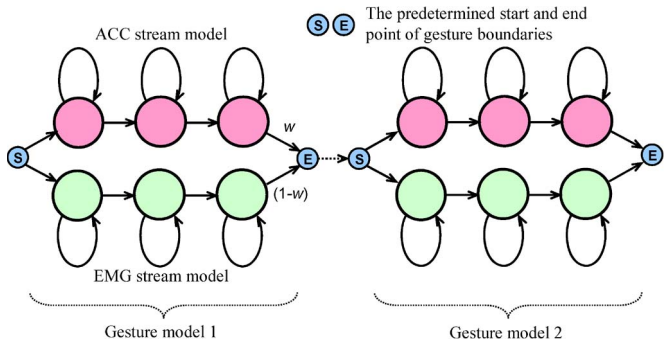


Fig. 3. Example of two gesture models with ACC and EMG streams.

the number of composite states increases significantly. That may cause much computational complexity [36].

In this paper, we choose to use the recombination strategy at the stream model level due to the assumption that the ACC and EMG streams representing different aspects (posture and trajectory) of the hand gesture are independent of each other. With this choice, each gesture model likelihood can be computed as depicted in

$$P(\mathbf{O}|\lambda) = \prod_{k=1}^{K} P^{w_k}\left(\mathbf{O}^{(k)}\left|\lambda^{(k)}\right.\right) \tag{11}$$

where w_k is the stream weight factor of the kth stream with the following restriction:

$$w_k \geq 0, \quad 1 \leq k \leq K, \quad \sum_{k=1}^{K} w_k = 1. \tag{12}$$

Most of the approaches use a linear weighted combination function of log-likelihood as follows:

$$\log P(\mathbf{O}|\lambda) = \sum_{k=1}^{K} w_k \log P\left(\mathbf{O}^{(k)}\left|\lambda^{(k)}\right.\right). \tag{13}$$

Multistream HMM Algorithm: The multistream HMM is implemented based on multiple single-stream HMMs, which independently model each stream between two synchronization points. The synchronization points are often the gesture boundaries to avoid the mistake of misalignment during recognition. In our method, a pair of synchronization points is predetermined as the start and end points of the active segment corresponding to the gesture. Due to the data segmentation procedure, the continuous gesture recognition can be simplified as the concatenated recognition of every isolated gesture (see Fig. 3).

For the information fusion of both ACC and EMG, each gesture class (or control command) is represented by a multistream HMM consisting of ACC and EMG stream models, denoted as $\lambda^{(A)}$ and $\lambda^{(E)}$, respectively. Equation (13) can be rewritten as

$$\log P(\mathbf{O}|\lambda) = w \log P\left(\mathbf{O}^{(A)}\left|\lambda^{(A)}\right.\right) + (1-w)\log P\left(\mathbf{O}^{(E)}\left|\lambda^{(E)}\right.\right) \tag{14}$$

where $\mathbf{O}^{(A)}$ and $\mathbf{O}^{(E)}$ are observed feature sequences from both the ACC and EMG streams and w is the stream weight factor. The stream model likelihoods $P(\boldsymbol{O}^{(E)}|\lambda^{(E)})$ and $P(\boldsymbol{O}^{(A)}|\lambda^{(A)})$ can be calculated using the forward–backward algorithm [29]. Thus, the recognition result for an unknown gesture observation $\mathbf{O}$ can be determined according to (10).

Training the multistream HMM in this paper consists of two tasks. The first task is the training of its ACC and EMG stream HMMs. All the stream models are trained in parallel using the Baum–Welch algorithm applied on gesture samples. In our method, we utilize continuous density HMMs, where the observation data probability is modeled as a multivariate Gaussian distribution. Good results have been obtained in earlier studies [9], [17] by using left-to-right HMMs with five states and three mixture components. It has also been reported that these parameters do not have a significant effect on the gesture recognition results [17]. The same parameters for models are chosen here because of better recognition performance and less computational complexity. The second task is the estimation of appropriate stream weights, which is described hereinafter.

Stream Weight Estimation: The multistream HMM proposed earlier consists of ACC and EMG feature streams, and the final decision is generated from the summation of logarithmic likelihoods of ACC and EMG models weighted by stream weights. These stream weights should be determined properly in order to improve the classification performance. However, they cannot be estimated based on the HMM training approach. In recent years, a great interest has been devoted to the determination of stream weights for multimodal integration, including for the audiovisual automatic speech recognition (AV-ASR) system. Various criteria have been employed to optimize stream weights with limited training data sets, for example, the maximum entropy criterion and the minimum classification error criterion investigated by Gravier *et al.* [30] and the likelihood-ratio maximization criterion and the output likelihood normalization criterion proposed by Tamura *et al.* [31]. The visual information is often regarded as a supplement in most previous AV-ASR systems, particularly in low SNR environments. However, in our method, the ACC and EMG streams are of the same importance for hand gesture recognition, although there are differences between the two stream models in many fields, such as the input feature sequences extracted from two heterogeneous sensor data, the model topologies, and the parameters. The output log-likelihoods of two stream models may vary even in magnitude. If the output likelihood of one stream is significantly larger than that of the other stream, the contribution to the classification of the other stream will be ignored when the equal weights are used. Therefore, the stream weight estimation in our methods focuses on the balance of the two streams' contribution to the classification.

Our stream weight adaptation approach consists of evaluating the differential log-likelihoods for each stream and normalizing them as stream weights. The differential log-likelihoods of the gesture class c $(c = 1, 2, \ldots, C)$ for the ACC and EMG streams are defined, respectively

$$Diff_c^{(A)} = C \sum_{\mathbf{O}\in\lambda_c} \log P\left(\mathbf{O}^{(A)}\left|\lambda_c^{(A)}\right.\right) - \sum_{\mathbf{O}} \log P\left(\mathbf{O}^{(A)}\left|\lambda_c^{(A)}\right.\right) \tag{15}$$

TABLE I
LIST OF 72 SELECTED CSL WORDS

No.	Word	Meaning	No.	Word	Meaning	No.	Word	Meaning	No.	Word	Meaning
1	你	you	19	好	good	37	死亡	dead	55	推	push
2	我	I, me	20	饱	full	38	等号	equal sign	56	为	for
3	们	everybody	21	很	very	39	多	many	57	谁	who
4	先生	sir	22	不	no, not	40	翻	turn over	58	无锡	Wuxi city
5	韩	Han (surname)	23	还	or	41	范围	scope, range	59	相同	same
6	鱼	fish	24	哪里	where	42	故意	intentionally	60	下	down
7	虾	shrimp	25	中国	China	43	厚	thick	61	现在	now
8	猪	pig, pork	26	也	also	44	坏	bad	62	循环	circulate
9	牛	bull, beef	27	谢谢	thanks	45	加	plus	63	要	take, need
10	菜	vegetable, food	28	再见	bye	46	伎俩	trick	64	一定	definite, must
11	汤	soup	29	想	think, want	47	小孩	child	65	意义	meaning
12	酒	wine	30	这里	here	48	拉	pull	66	油	oil
13	吃	eat	31	报复	avenge	49	墙壁	wall	67	指挥	instruct
14	喝	drink	32	玻璃	glass	50	轻	soft	68	自己	self
15	去	go	33	也许	perhaps	51	舒服	comfort	69	棕	brown
16	有	have, exist	34	擦	wipe	52	水	water	70	昨天	yesterday
17	没有	not have	35	打	hit	53	上午	morning	71	山	mountain
18	是	yes, be	36	党	political party	54	提	lift	72	镇	town

$$Diff_c^{(E)} = C \sum_{\mathbf{O}\in\lambda_c} \log P\left(\mathbf{O}^{(E)} \middle| \lambda_c^{(E)}\right) - \sum_{\mathbf{O}} \log P\left(\mathbf{O}^{(E)} \middle| \lambda_c^{(E)}\right). \quad (16)$$

Their values denote the degree of distinguishing the class c from the other classes for each stream. Moreover, the stream weight w in (14) can be calculated as

$$w = \frac{\sum_c Diff_c^{(E)}}{\sum_c Diff_c^{(A)} + \sum_c Diff_c^{(E)}}. \quad (17)$$

Thus, with the stream weight inversely proportional to the differential logarithmic likelihoods, the ACC and EMG streams can play the same important role in hand gesture recognition.

III. SLR

A. Data Collection

In order to evaluate the performance of the hand gesture recognition method based on the information fusion of the ACC and EMG, the experiments on CSL recognition were conducted. Seventy-two CSL single-hand words were selected as pattern classes to form the gesture dictionary, as shown in Table I. Fig. 4 also specifies the actual movements corresponding to five typical words by example. Forty kinds of sentences were constituted by the aforementioned 72 CSL words. The practicality of the gesture segmentation and recognition method was tested by these sentences with continuous word streams.

The ACC and EMG signal measurements were made with our self-made sensor system. The three-axis accelerometer built by MMA7361 (Freescale Semiconductor, Inc., Austin, TX) was placed on the back of the forearm near the wrist to capture the information about hand orientations and trajectories (see Fig. 4). In each EMG sensor, there are two silver bar-shaped electrodes with a 10 mm × 1 mm contact dimension and a 10-mm electrode-to-electrode spacing. The differential

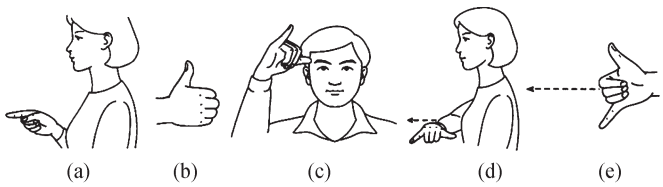


Fig. 4. Five examples of CSL words. (a) "You." (b) "Good." (c) "Bull." (d) "Also." (e) "Go."

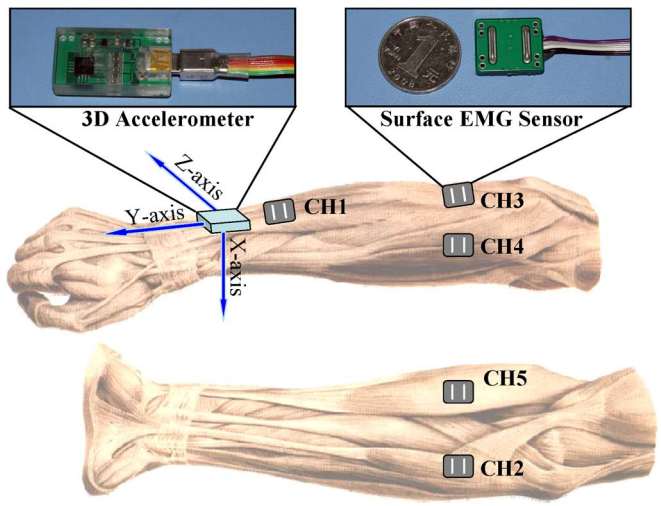


Fig. 5. Sensor placement of three-axis ACC and five-channel EMG. The anatomical pictures of the forearm muscles are adapted from [38].

EMG signals in each channel pass through a two-stage amplifier, which is formed by AD8220 and AD8698 (Analog Devices, Inc., Norwood, MA) with a total gain of 60 dB and a bandpass filtering of 20 to 1000 Hz bandwidth. Five-channel surface EMG sensors were located over five sites on the surface of the forearm muscles: *extensor digiti minimi*, *palmaris longus*, *extensor carpi ulnaris*, *extensor carpi radialis*, and *brachioradialis*, respectively, as shown in Fig. 5. The sampling rate for data collection was 1 kHz.

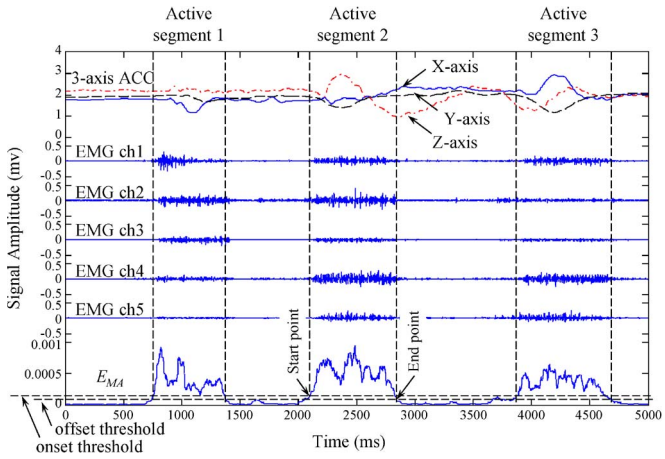


Fig. 6. Illustration of data segmentation.

Two right-handed subjects, male (age 27) and female (age 25), participated in the data collecting experiments. They were both healthy, had no history of neuromuscular or joint diseases, and were informed of the associated risks and benefits specific to the study. Each subject was required to participate in the experiments for more than 5 times (5 days with 1 experimental session per day). In each session, the subjects performed the selected 72 CSL words in a sequence with 12 repetitions per motion, and then, they further performed the defined 40 sentences with 2 repetitions per sentence. Both of the ACC and EMG signals were recorded as data samples for CSL recognition. The data set for experimental analysis consisted of 8640 CSL word samples and 800 sentence samples in total.

B. Experimental Results and Analysis

1) Data Segmentation Results: Fig. 6 illustrates the double-threshold principles of the data segmentation method. The three-axis ACC and five-channel EMG signals recorded when a subject was continuously performing the three CSL words "我-喝-汤" (which means "I drink soup" in English) are shown with the moving averaged energy stream $E_{MA}(t)$ below them in Fig. 6. The stream $E_{MA}(t)$ rising above the onset threshold denotes the start point, and $E_{MA}(t)$ falling down the offset threshold denotes the end point of the active segment. The three active segments corresponding to the three CSL words are successfully marked on the figure. For effective data segmentation, many factors should be considered to choose the values of the onset and offset thresholds, such as the strength that the user exerts when performing hand gestures and the environmental noises. We think that the noise is the dominant factor. If the noise level increases, the corresponding thresholds should also be adjusted higher to avoid the false detection caused by noises. Fortunately, the data collecting environment in our experiments is favorable so that we choose the onset threshold as 2% of the $E_{MA}(t)$ recorded by the experts when the user performs the hand grasping at maximum volume contraction (MVC), and the offset threshold is usually set as 75% of the onset threshold.

2) CSL Word Classification: The data collection experiments for each subject in five different sessions can generate five groups of data sets, respectively. The user-specific classification of 30 CSL words was carried out using the fivefold cross-validation approach. Four group data samples from four sessions for each subject were used as the training samples, and the other remaining group data were referred to as the testing samples.

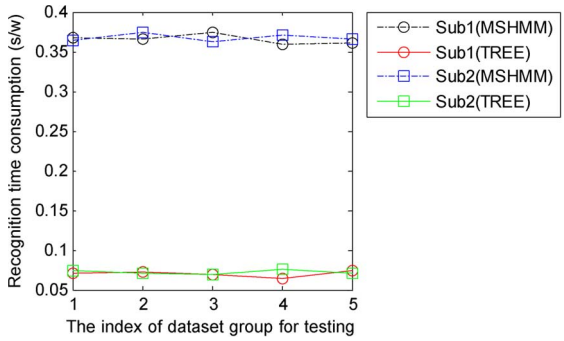


Fig. 7. Average recognition time consumption of the methods.

According to the classification method presented in Section III, the multistream HMM (represented by MSHMM) classifiers are the special part of the proposed decision tree (represented by TREE). The performances on 72-CSL-word classification with MSHMM and TREE are tested, respectively, which means that only multistream HMMs are used to recognize CSL words in the MSHMM condition for comparison (the signal processing and gesture recognition procedure marked with the thick dash in Fig. 1).Table II shows the test results of the MSHMM and TREE.

In Table II, the TREE approach achieved the average recognition accuracies of 95.3% for Sub1 and 96.3% for Sub2, and the MSHMM approach obtained the average recognition accuracies of 92.5% and 94.0%, respectively. On the basis of the MSHMM, the decision tree increased the overall recognition accuracy by 2.56% $(p = 1.068\mathrm{E}-6 < 0.001)$ for the two subjects. This may be attributed to two factors. One is the different additional features utilized through different classifiers in the nonleaf nodes of the decision tree which provided more information that enhanced the seperability. The other is that the decision tree reduced the searching range between word classes level by level, and some easily confused words that could cause recognition error might be excluded from the set of candidate words.

The recognition time consumptions of the MSHMM and TREE were also investigated for the approach of the cross-validation test. All the tests were realized on a PC (Intel E5300 at a 2.6-GHz CPU with a 2-GB RAM) using Matlab R2007a (The Mathworks, Inc., Natick, MA). As shown in Fig. 7, the average time consumption of the MSHMM was 0.366 second per word (s/w) for Sub1 and 0.368 s/w for Sub2. In contrast, the TREE approach obtained the average time consumption of 0.0704 and 0.0726 s/w, respectively. Experimental results indicated that the TREE approach could reduce the recognition time consumption significantly. The classifiers in the top of the TREE with effective classification rules but low computational complexity were applied prior to the MSHMM to exclude the most impossible word classes. Consequently, the searching range of the MSHMM, as well as the recognition time consumption, can be reduced effectively.

For a further investigation on the information complementarity of the EMG and ACC, five words (see Fig. 4) are selected from the 72 CSL words for classification in three conditions:

TABLE II
LIST OF 72 SELECTED CSL WORDS

Conditions	1st Test (%)		2nd Test		3rd Test (%)		4th Test (%)		5th Test (%)		Overall (%)	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Sub1(MSHMM)	90.5	19.8	91.1	17.1	92.9	17.7	93.2	17.3	94.9	11.7	92.5	14.7
Sub1(Tree)	93.4	15.5	94.4	9.89	96.1	9.17	95.7	10.5	97.0	7.70	95.3	9.44
Sub2(MSHMM)	91.2	16.7	94.4	17.1	93.3	13.2	93.5	17.2	97.3	10.5	94.0	11.6
Sub2(Tree)	93.8	12.4	97.1	9.19	95.6	10.1	96.6	13.5	98.2	8.18	96.3	8.25

TABLE III
CONFUSION MATRIX FOR CLASSIFICATION IN ACC-ONLY CONDITIONS

	you	good	bull	also	go	Accuracy
you	65	55	0	0	0	54.2%
good	25	95	0	0	0	79.2%
bull	0	0	120	0	0	100.0%
also	0	0	0	120	0	100.0%
go	0	0	0	0	120	100.0%

TABLE IV
CONFUSION MATRIX FOR CLASSIFICATION IN EMG-ONLY CONDITIONS

	you	good	bull	also	go	Accuracy
you	103	6	4	0	7	85.8%
good	0	119	1	0	0	99.2%
bull	0	2	79	36	3	65.8%
also	0	0	1	109	10	90.8%
go	0	0	9	43	68	56.7%

ACC-only, EMG-only, and fusion of ACC and EMG. In the ACC-only or EMG-only condition, only one stream HMM and features of the ACC or EMG are used in the decision tree for classification. Tables III–V show the composite confusion matrices of the fivefold cross-validation classification of five words for both Sub1 and Sub2 in three conditions, respectively. The total number of each word is 120. The words "you" and "good" are both static gestures with the same hand orientation (arm toward the front and palm toward the left) and different hand postures, index finger extension for "you" and thumb extension for "good," so these two words cannot be distinguished effectively using only ACC signals, whereas the EMG can overcome this. Since the word "bull" is also a static gesture but with a different hand orientation (arm toward up) from that of the word "you" or "good," the ACC can provide a relatively high confidence in its recognition. Contrarily, the words "bull," "also," and "go" are performed with the same hand posture (thumb and little finger extension), and different trajectories cannot be distinguished in EMG-only conditions without the supplement of ACC features. All the words can be classified successfully with high accuracies in the condition of ACC and EMG fusion. In addition to these five words, the complementary effect could be observed for all the words in our experiments. The aforementioned five words were selected as typical and intuitive examples. The complementary functionality of both ACC and EMG signals has been also examined by Kim *et al.* [32]. This paper expanded it for CSL recognition with a relatively larger vocabulary based on our own fusion strategy.

3) CSL Sentence Recognition: This experiment is to test the recognition performance on the CSL sentences using the proposed continuous hand gesture recognition approaches. For the user-specific classification, all the five groups of data samples

TABLE V
CONFUSION MATRIX FOR CLASSIFICATION IN FUSION CONDITIONS

	you	good	bull	also	go	Accuracy
you	115	5	0	0	0	95.8%
good	0	120	0	0	0	100.0%
bull	0	0	120	0	0	100.0%
also	0	0	0	120	0	100.0%
go	0	0	0	0	120	100.0%

TABLE VI
RECOGNITION RESULTS OF CSL SENTENCES FOR TWO SUBJECTS

Dataset	N	D	I	S	Ps	Pw	P
Sub1	1930	18	8	101	98.7%	93.4%	74.0%
Sub2	1930	17	9	113	98.7%	92.8%	71.0%
Overall	3860	35	17	214	98.7%	93.1%	72.5%

for each subject were used to train classifiers, and the collected sentence samples from the same subject were tested one by one. The output of the well-trained classifiers was the recognized CSL words in a sequence of detected active segments in a signal stream of each sentence. The sentence recognition results for Sub1 and Sub2, respectively, are listed in Table VI, where the word segment detection rate P_s and the word recognition rate P_w are computed through the following equations:

$$P_s = 1 - \frac{D + I}{N} \tag{18}$$

$$P_w = 1 - \frac{D + S + I}{N} \tag{19}$$

where D is the number of deletions, S is the number of substitutions, I is the number of insertions, and N is the total number of words which constitute all the sentences in the test set. The sentence recognition rates P are then calculated as the percentage of the correctly recognized sentences to the total sentence number, which is defined the same as in [10].

The sentence samples collected from each subject were constituted by 1930 CSL words. The word segment detection rate was 98.7% for both the two subjects. That means that the performed CSL word segments within a continuous sequence of gestures can be mostly detected with few deletions and insertions through the proposed data segmentation step. The word recognition rate was 93.4% for Sub1 and 92.8% for Sub2. The accuracy in the sentence recognition was lower than that of the word classification due to the signal variation of the words in the sentences. For collecting CSL word samples, each word was repeated 12 times one by one, but for sentence collection, the subjects were required to continuously perform a sequence of various words. The overall recognition rate of the total 800 sentences was 72.5% because of the stringent statistical criteria that the correctness of a sentence entails the correct recognition of all the words constituting the sentence

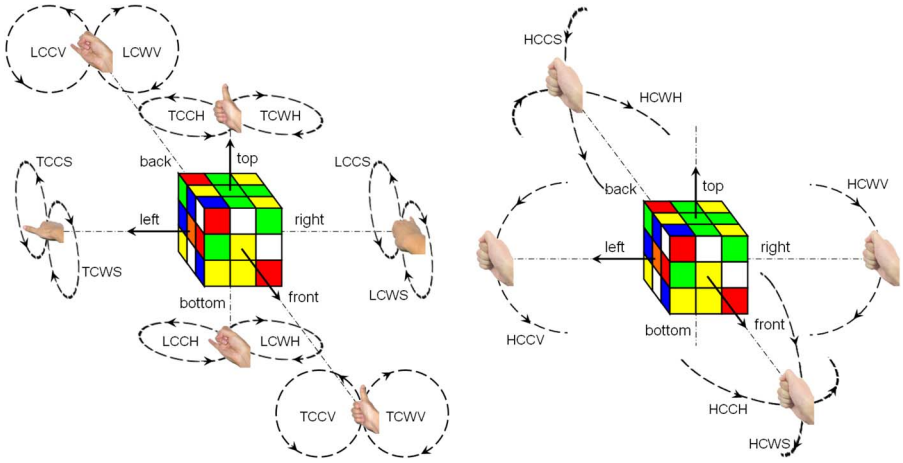


Fig. 8. (Left) Twelve circular gestures to turn the six planar faces of the cube. (Right) Six circular gestures to rotate the entire cube.

without any insertions, substitutions, or deletions. The errors of word segment detection and recognition were scattered in many sentences. That was the main factor that caused the low sentence recognition rates. We did not do any optimization in the CSL sentence recognition according to the grammar and syntax. If the factors mentioned earlier were considered, the performance of CSL sentence recognition could be improved. Exploring this remains future work.

IV. Gesture-Based Control of Virtual Rubik's Cube

In this section, an interactive system was established to evaluate our framework for hand gesture recognition with application to gesture-based control. In contrast with SLR in Section III, the system processes both ACC and EMG signals in real time, and the recognized gestures are translated into control commands. A virtual Rubik's cube was built in our interactive system to demonstrate the advantages of EMG and ACC fusion by providing multiple degrees of freedom in control. The experimental setups, including the principles of the Rubik's cube game and gesture control schemes, are introduced hereinafter.

A. Experimental Setups

1) Virtual Rubik's Cube Interface: Rubik's cube is a mechanical puzzle. In a standard $3 \times 3 \times 3$ cube, each of the six faces is covered by nine stickers, and they are colored with different solid colors (traditionally being white, yellow, orange, red, blue, and green). Each face is able to be turned independently, thus mixing up the color stickers inside the faces. For the puzzle to be solved, each face must be made of one solid color.

2) Selected Hand Gestures: Utilizing the complementary sensing characteristics of EMG and ACC signals, the selected hand gestures include three basic hand postures and six circular hand movements. Since any arbitrary transformation of the cube can be achieved by a series of steps of rotating the six external faces of the cube, we defined 12 circular gestures to rotate the six cube faces by 90° clockwise or counterclockwise, as illustrated in the left subgraph of Fig. 8. When these gestures are being performed, either the thumb or little finger needs to be extended for determining which side is to be rotated: the top or bottom, front or back, and left or right; moreover, the direction of the hand circles determines in which direction the side is turned. Since the interface screen can only show three faces (e.g., the top, front, and left as in Fig. 8) of the cube at the time, six gestures with hand grasping (as shown in the right subgraph of Fig. 8) are used for rotating the entire cube by 90° clockwise or counterclockwise around three axes so that all six faces of the virtual cube can be brought into the front view.

TABLE VII
Name Abbreviation of Gestures Used to Control the Cube

	Posture		Direction		Plane
H	Hand grasp	CW	Clockwise	H	In left-front plane
T	Thumb	CC	Counter-Clockwise	V	In top-left plane
L	Little finger			S	In front-top plane

Each gesture defined is named by a four-letter abbreviation. These names indicate gesture meanings which are described in Table VII. It is intuitive to comprehend the gesture controls of the virtual Rubik's cube. For example, the gesture TCWH means thumb extension and hand circles drawn clockwise in the left-front plane. This gesture makes the topmost face of the virtual Rubik's cube turn clockwise.

3) Sensor Placement: The sensor placement in this experiment was similar to that of the SLR in Fig. 5. A three-axis accelerometer and only three-channel EMG sensors (CH3–CH5 in Fig. 5) were utilized in game control. The three EMG sensors were attached to the inner side of a stretch belt for convenient sensor installation.

4) Testing Schemes: Ten users, five males and five females, aged from 21 to 27, participated into the gesture-based control experiments. In contrast with the aforementioned SLR experiment only conducted in user-specific classification, which means that the classifiers were trained and tested independently on data from each user, the gesture-based control experiments consisted of two testing schemes: user-specific and user-independent classification.

In the user-specific classification, each of the ten subjects participated into the experiments for three times (three days with one experimental session per day). In each session, the subjects performed the defined 18 kinds of hand gestures in a sequence with ten repetitions per motion and recorded training data samples firstly. Then, the system loaded the data recorded in the current session to train the classifiers, and the subjects

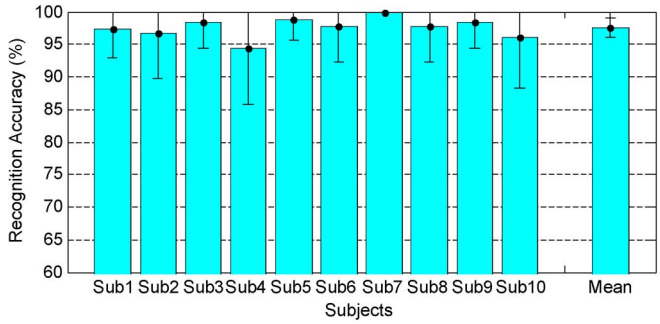


Fig. 9. User-specific classification accuracies of 18 kinds of hand gestures.

further performed ten testing samples per gesture for classification to evaluate the system performance.

The experiment in the user-independent classification was proposed to evaluate the generality of our interactive system based on the fusion of ACC and EMG. For this testing scheme, we used the leave-one-out method: The training data from nine users in all three sessions were mixed and loaded together to train the classifiers, which then were applied to recognize the gestures performed by the remaining user to test the results. Additionally, all users participated into the experiments in the way that they were playing an entertaining Rubik's cube game: The cube was initialized with each face randomly disordered as a puzzle; then, the subjects were required to sort the faces of the cube into a solid color as fast as possible to solve the puzzle using the defined gesture commands.

B. Experimental Results and Analysis

1) User-Specific Experiments: Fig. 9 shows the average classification accuracies of the 18 kinds of hand gestures in three sessions for each subject separately. The mean classification accuracy for all the subjects reached a 97.6% (SD: 1.45%) average accuracy. According to these satisfying results, the experiments can also be regarded as a practical example to demonstrate the feasibility of building a gesture-based control system using ACC and EMG signals.

2) User-Independent Experiments: To extend the user-specific classification results, we explored the system performance in user-independent classification. Every subject found it fun to play the virtual Rubik's cube game for puzzle solving. All the gesture commands were defined in pairs. If an occasional recognition error occurred, it seldom influenced the game: Users could easily perform the gesture controlling the counteraction of the error command and continue to play. Table VIII shows the statistical results for ten subjects to solve the Rubik's cube puzzle in user-independent classification.

The recognition results achieved with our system were satisfactory as the overall accuracy was 90.2%. It is not unexpected that the recognition rates in the user-independent classification are lower than that in the user-specific classification. Due to the individual differences of the biosignal-like EMG, there exist great challenges to establish user-independent EMG-based recognition and interaction systems. Although previous researchers have realized various outstanding prototypes with EMG-based interfaces, few studies on user-independent classification have been reported, and the limited testing results are not satisfying [12], [16], [32]. Our experimental results also indicate another advantage of the fusion of ACC and EMG sensors that the ACC and EMG information fusion technique not only enhances the performance of a gesture-based control system with high accuracies but also reduces the burden of a single sensor. To some extent, the main task of the EMG in our system was to distinguish three hand postures in 18 kinds of hand gestures so that it is easy to achieve relatively robust user-independent classification in our system. The average input rate for gesture commands was about 16/min. These figures indicate that the proposed gesture-based control method is efficient.

TABLE VIII
STATISTICAL RESULTS FOR TEN SUBJECTS TO SOLVE RUBIK'S CUBE PUZZLE IN USER-INDEPENDENT CLASSIFICATION

Subject	Gesture Performed	Correctly Recognized	Accuracy (%)	Time Consumption	Commands per Minute
Sub 1	32	29	90.6	1'35"	20.2
Sub 2	27	24	88.9	1'47"	15.1
Sub 3	27	26	96.3	1'26"	18.8
Sub 4	33	28	84.8	2'08"	15.5
Sub 5	24	20	83.3	1'23"	17.3
Sub 6	26	23	88.5	1'33"	16.8
Sub 7	29	28	96.5	1'32"	18.9
Sub 8	21	18	85.7	1'22"	15.4
Sub 9	19	18	94.7	1'09"	16.5
Sub 10	27	25	92.6	1'44"	15.6
Overall	265	239	90.2	15'39"	16.9

For the realization of natural gesture-based HCI, the subjects recruited in the experiments were asked to perform gestures in a way that felt natural to them. Consequently, how hard they performed the tasks could not be accurately quantified. Generally, each subject performed every hand gesture at 10%–20% of the MVC. The strength of performing hand gestures varied in subject due to the different personal habits, which were also attributed to individual differences and could affect the EMG amplitudes. In this paper, the performance of the user-independent classification suffered from the strength variation, whereas the user-specific classification was relatively insensitive to this factor because of the consistency of the strength exerted by the same subject. From the experiments on real-time gesture recognition, it was also observed that some subjects could adjust their strength to perform gestures in order to achieve higher classification rates in user-independent classification with the instantaneous visual feedback. We call this phenomenon as "user self-learning," which could partly support our idea that the strength for different subjects is a major factor of the individual difference that could influence the performance of hand gesture recognition in user-independent classification.

V. CONCLUSION AND FUTURE WORK

This paper has developed a framework for hand gesture recognition which can be utilized in both SLR and gesture-based control. The presented framework combines information from a three-axis accelerometer and multichannel EMG sensors to achieve hand gesture recognition. Experimental results on the classification of 72 CSL words show that our framework is effective to merge ACC and EMG information with the average accuracies of 95.3% and 96.3% for two subjects. On the basis of multistream HMM classifiers, the decision tree increases the overall recognition accuracy by 2.5% and significantly reduces

the recognition time consumption. The ability of continuous SLR by our framework is also demonstrated by the recognition results of 40 kinds of CSL sentences with an overall word accuracy of 93.1% and a sentence accuracy of 72.5%. The real-time interactive system using our framework achieves the recognition of 18 kinds of hand gestures with average rates of 97.6% and 90.2% in the user-specific and user-independent classification, respectively. We have shown by example of game control that our framework can be generalized to other gesture-based interaction.

There are further potential advantages of the combination of EMG and ACC signals. With the supplementary ACC data, the recognition system may effectively overcome some problems typical to EMG measurements, such as individual physiological differences and fatigue effects. Furthermore, EMG is capable of sensing muscular activity that is related to no obvious movement [4]. Such gestures are useful in mobile use contexts where the discretion of the interaction is an important issue. On all accounts, the combination of EMG and ACC measurements can enhance the functionality and reliability of gesture-based interaction.

Although we have researched into an effective fusion scheme for the combination of ACC and EMG sensors with successful applications, there are still some problems to be further studied.

1) The utilization of two hands and other useful parameters in sign language. The CSL recognition experiment in this paper only utilized some single-hand words to evaluate our proposed framework. Investigating the two-hand information fusion and other useful parameters in sign language, including gaze, facial expression, motion of head, neck, and shoulder, and body posture, is a further direction.
2) The effortless and fast customization of robust gesture-based interaction. In our experiments, the training data samples were collected by many subjects who were required to perform each predefined hand gesture with abundant repetitions in multiple sessions. This approach was the important factor to achieve relatively satisfactory results in this paper. Since hand gestures should be customizable, easy, and quick to train to meet the requirement of most common users, our future work will focus on enhancing the robustness of the system to enable effortless customization and extending our methods to other types of applications, for example, to gesture-based mobile interfaces. In addition, the design of tiny, wireless, and flexible sensors that are better suited for common users in real applications is another goal of our research.

ACKNOWLEDGMENT

The authors are grateful to all the volunteers for their participation in this study. We would like to express our special appreciation to Dr. Z. Zhao, W. Wang, and C. Wang for their assistance in the experiments.

REFERENCES

[1] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 3, pp. 311–324, May 2007.
[2] J. Mäntyjärvi, J. Kela, P. Korpipää, and S. Kallio, "Enabling fast and effortless customisation in accelerometer based gesture interaction," in *Proc. 3rd Int. Conf. Mobile Ubiquitous Multimedia*, New York, 2004, pp. 25–31.
[3] T. Pylvänäinen, "Accelerometer based gesture recognition using continuous HMMs," in *Proc. Pattern Recog. Image Anal.*, *LNCS 3522*, 2005, pp. 639–646.
[4] E. Costanza, S. A. Inverso, and R. Allen, "Toward subtle intimate interfaces for mobile devices using an EMG controller," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.* , Portland, OR, Apr. 2–7, 2005, pp. 481–489.
[5] M. Asghari Oskoei and H. Hu, "Myoelectric control systems—A survey," *Biomed. Signal Process. Control*, vol. 2, no. 4, pp. 275–294, Oct. 2007.
[6] D. M. Sherrill, P. Bonato, and C. J. De Luca, "A neural network approach to monitor motor activities," in *Proc. 2nd Joint EMBS/BMES Conf.*, Houston, TX, 2002, vol. 1, pp. 52–53.
[7] T. Starner, J. Weaver, and A. Pentland, "Real-time American Sign Language recognition using desk and wearable computer based video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1371–1375, Dec. 1998.
[8] T. Shanableh, K. Assaleh, and M. Al-Rousan, "Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic Sign Language," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 3, pp. 641–650, Jun. 2007.
[9] G. Fang, W. Gao, and D. Zhao, "Large vocabulary sign language recognition based on fuzzy decision trees," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 34, no. 3, pp. 305–314, May 2004.
[10] K. Assaleh, T. Shanableh, M. Fanaswala, H. Bajaj, and F. Amin, "Vision-based system for continuous Arabic Sign Language recognition in user dependent mode," in *Proc. 5th Int. Symp. Mechatron. Appl.*, Amman, Jordan, 2008, pp. 1–5.
[11] K. R. Wheeler, M. H. Chang, and K. H. Knuth, "Gesture-based control and EMG decomposition," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 36, no. 4, pp. 503–514, Jul. 2006.
[12] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, "Demonstrating the feasibility of using forearm electromyography for muscle–computer interfaces," in *Proc. 26th SIGCHI Conf. Human Factors Comput. Syst.*, Florence, Italy, 2008, pp. 515–524.
[13] A. Wilson and S. Shafer, "Between u and i: XWand: UI for intelligent spaces," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Ft. Lauderdale, FL, Apr. 2003, pp. 545–552.
[14] H. Brashear, T. Starner, P. Lukowicz, and H. Junker, "Using multiple sensors for mobile sign language recognition," in *Proc. 7th IEEE ISWC*, 2003, pp. 45–52.
[15] X. Chen, X. Zhang, Z. Y. Zhao, J. H. Yang, V. Lantz, and K. Q. Wang, "Hand gesture recognition research based on surface EMG sensors and 2D-accelerometers," in *Proc. 11th IEEE ISWC*, 2007, pp. 11–14.
[16] J. Kim, S. Mastnik, and E. André, "EMG-based hand gesture recognition for realtime biosignal interfacing," in *Proc. 13th Int. Conf. Intell. User Interfaces*, Gran Canaria, Spain, 2008, pp. 30–39.
[17] J. Kela, P. Korpipää, J. Mäntyjärvi, S. Kallio, G. Savino, L. Jozzo, and S. D. Marca, "Accelerometer-based gesture control for a design environment," *Pers. Ubiquitous Comput.*, vol. 10, no. 5, pp. 285–299, Jul. 2006.
[18] K. Englehart, B. Hudgins, and P. A. Parker, "A wavelet-based continuous classification scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 3, pp. 302–310, Mar. 2001.
[19] Y. Huang, K. Englehart, B. Hudgins, and A. D. C. Chan, "A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 11, pp. 1801–1811, Nov. 2005.
[20] A. V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Orlando, FL, 2002, pp. 2013–2016.
[21] H. Manabe and Z. Zhang, "Multi-stream HMM for EMG-based speech recognition," in *Proc. 26th Annu. Int. Conf. IEEE EMBS*, San Francisco, CA, 2004, pp. 4389–4392.
[22] M. Gurban, J. P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition," in *Proc. 10th Int. Conf. Multimodal Interfaces*, Chania, Greece, 2008, pp. 237–240.
[23] V. E. Kosmidou, L. J. Hadjileontiadis, and S. M. Panas, "Evaluation of surface EMG features for the recognition of American Sign Language gestures," in *Proc. IEEE 28th Annu. Int. Conf. EMBS*, New York, Aug. 2006, pp. 6197–6200.
[24] R. N. Khushaba and A. Al-Jumaily, "Channel and feature selection in multifunction myoelectric control," in *Proc. IEEE 29th Annu. Int. Conf. EMBS*, Lyon, France, Aug. 2007, pp. 5182–5185.

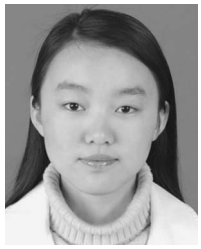
[25] X. Hu and V. Nenov, "Multivariate AR modeling of electromyography for the classification of upper arm movements," *Clinical Neurophysiol.*, vol. 115, no. 6, pp. 1276–1287, Jun. 2004.
[26] X. Chen, Q. Li, J. Yang, V. Lantz, and K. Wang, "Test–retest repeatability of surface electromyography measurement for hand gesture," in *Proc. 2nd ICBBE*, Shanghai, China, 2008, pp. 1923–1926.
[27] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Second ed. New York: Wiley, 2001, Section 10.4.4.
[28] C. Liu and H. Wechsler, "Robust coding schemes for indexing and retrieval from large face databases," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 132–137, Jan. 2000.
[29] L. Rabiner and B. Jaung, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[30] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," in *Proc. ICASSP*, 2002, p. 853.
[31] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, pp. 468–472.
[32] J. Kim, J. Wagner, M. Rehm, and E. André, "Bi-channel sensor fusion for automatic sign language recognition," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recog.*, Amsterdam, The Netherlands, 2008, pp. 1–6.
[33] C. Vogler and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis," in *Proc. 6th Int. Conf. Comput. Vis.*, Bombay, India, 1999, pp. 363–369.
[34] V. E. Kosmidou and L. J. Hadjileontiadis, "Sign language recognition using intrinsic mode sample entropy on sEMG and accelerometer data," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 12, pp. 2879–2890, Dec. 2009.
[35] D. Kelly, R. Delannoy, and J. Mc Donald, "A framework for continuous multimodal sign language recognition," in *Proc. ICMI-MLMI*, Cambridge, MA, 2009, pp. 351–358.
[36] Y. Kessentini, T. Paquet, and A. M. Ben Hamadou, "Off-line handwritten word recognition using multi-stream hidden Markov models," *Pattern Recognit. Lett.*, vol. 31, no. 1, pp. 60–70, Jan. 2010.
[37] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, pp. 1–5, Jan. 2011. DOI: 10.1109/TSMCA.2010.2093883.
[38] S. Gao and P. Yu, *Atlas of Human Anatomy (Revision)*. Shanghai, China: Shanghai Sci. & Tech. Publ., 1998, ch. 2, (in Chinese).

Xu Zhang received the B.S. degree in electronic information science and technology and the Ph.D. degree in biomedical engineering from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Postdoctoral Fellow with the Rehabilitation Institute of Chicago, Chicago, IL. His research interests include biomedical signal processing, pattern recognition for neurorehabilitation, and multimodal human–computer interaction.

Xiang Chen (A'11) received the M.E. and Ph.D. degrees in biomedical engineering from the University of Science and Technology of China, Hefei, China, in 2000 and 2004, respectively.

From 2001 to 2008, she was an Instructor with the Department of Electronic Science and Technology, University of Science and Technology of China, where she has been an Associate Professor since 2008. She is currently the Director of the Neural Muscular Control Laboratory, University of Science and Technology of China. Her research interests include biomedical signal processing, multimodal human–computer interaction, and mobile health care.



Yun Li received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2005, where she is currently working toward the Ph.D. degree.

Her research interests include biomedical signal processing, sign language recognition, and multimodal human–computer interaction.

Vuokko Lantz received the M.S. degree in system analysis and operation research and the Ph.D. degree in computer and information science from the Helsinki University of Technology, Helsinki, Finland, in 1999 and 2002, respectively.

Since 2003, she has been with the Nokia Research Center, Tampere, Finland. Currently, she leads the Multimodal Interaction team, Tampere Laboratory, Nokia Research Center. Her research interests include text entry, handwriting recognition, use-context analysis, mobile user testing, gaze tracking, gesture- and touch-based interaction, and dynamic audiotactile feedback.

Kongqiao Wang received the Ph.D. degree in signal and information processing from the University of Science and Technology of China, Hefei, China, in 1999.

He joined the Nokia Research Center Beijing laboratory in 1999. Currently, he is leading the research team focusing on multimodal and multimedia user interaction. Meanwhile, he is strongly pushing a global research program of Nokia, gestural user interface, as the program leader. The Nokia Plug and Touch delivered from the program has attracted strong media attentions through Nokia World 2011 in London. He is also one of the Nokia active inventors. His research interests include visual computing technologies, pattern recognition, and related user interactions.

Jihai Yang received the B.S. degree from Harbin Engineering University, Harbin, China, in 1969.

From 1992 to 2001, he was an Associate Professor with the Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, China, where he has been promoted to a Professor since 2002. His current research interests are biomedical signal processing, neuromuscular control, and modeling of a bioelectric process.