Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 49037, 10 pages
doi:10.1155/2007/49037

Research Article

Expectation-Maximization Method for EEG-Based Continuous Cursor Control

Xiaoyuan Zhu,[1] Cuntai Guan,[2] Jiankang Wu,[2] Yimin Cheng,[1] and Yixiao Wang[1]

[1] *Department of Electronic Science and Technology, University of Science and Technology of China, Anhui, Hefei 230027, China*
[2] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*

Received 21 October 2005; Revised 12 May 2006; Accepted 22 June 2006

Recommended by William Allan Sandham

To develop effective learning algorithms for continuous prediction of cursor movement using EEG signals is a challenging research issue in brain-computer interface (BCI). In this paper, we propose a novel statistical approach based on expectation-maximization (EM) method to learn the parameters of a classifier for EEG-based cursor control. To train a classifier for continuous prediction, trials in training data-set are first divided into segments. The difficulty is that the actual intention (label) at each time interval (segment) is unknown. To handle the uncertainty of the segment label, we treat the unknown labels as the hidden variables in the lower bound on the log posterior and maximize this lower bound via an EM-like algorithm. Experimental results have shown that the averaged accuracy of the proposed method is among the best.

1. INTRODUCTION

Brain-computer interface (BCI) is a communication system in which the information sent to the external world does not pass through the brain's normal output pathways. It provides a radically new communication option to people with neuromuscular impairments. In the past decade or so, researchers have made impressive progress in BCI [1]. In this paper our discussions focus on the Electroencephalogram (EEG) driven BCI. Different types of EEG signals have been used as the input of BCI system, such as slow cortical potentials (SCPs) [2], motor imagery signal [3], P300 [4, 5], and steady-state visual-evoked response (SSVER) [6]. In the recent years, EEG controlled cursor movement has attracted many research interests. In this kind of BCI, first, EEG is recorded from the scalp and digitalized in both temporal and spatial space by using acquisition system. Then the digitalized signals are subjected to one or more of feature extraction procedures, such as spectral analysis or spatial filtering. Afterwards, translation algorithm converts the EEG feature into command vector whose elements control different dimensions of cursor movement independently. Finally, the outputs of cursor control part are displayed on the screen. The subjects can learn from these feedbacks to improve their control performance. In Figure 1 we depict one-dimensional (1D) four-targets cursor-control system as an example. In the scenario of 1D four targets cursor control, there are four targets on the right side of the screen. Targets 1 to 4 are from top to bottom. The original position of the cursor is on the middle of the left side. During each trial, the cursor moves across the screen at a steady rate. The subjects' task is to move the cursor to the predecided target by performing vertical control at each time interval, usually every hundreds of milliseconds. Our aim is to continuously predict the cursor movement at each time interval as well as the final target.

Many groups have made great efforts to find effective translation algorithms to improve the performance of BCI system. The translation algorithms for cursor control BCI come under two categories: regression [7–9] and classification [10–12]. Each of them has its merits. Here, we adopt classification method to continuously predict cursor movement (up or down in 1D case as discussed here) using EEG signal. To train the classifier, we divide each trial into segments. Now the key issue is how to train a classifier, with given training data set where there is no knowledge of actual intended cursor movement at any time intervals during a trial. In other words, for each trial, although the final target label of the trial is known, the true label of each segment is unknown, which imposes great difficulties in classifier training. In this paper, we denote this issue as "unlabeled problem."

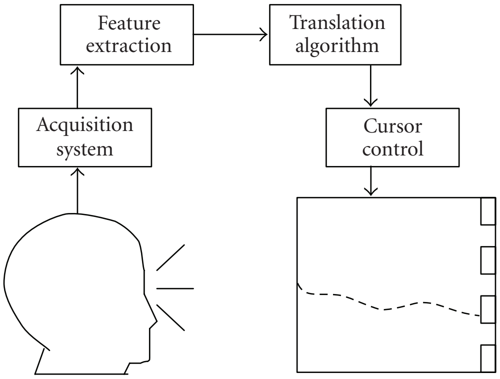


Figure 1: The diagram of EEG-based BCI system in one-dimensional four-targets cursor-control case.

In [10] Roberts and Penny extracted features using an AR model and classified them into cursor movement. Since their method is used in 1D two-targets cursor-control scenario, they can label the training data set, and use standard Bayesian learning method to train the classifier. In 1D four-targets cursor-control scenario, which we are discussing here, Cheng et al. [11] proposed a trialwise method to classify the cursor target position, and reported results on BCI Competition 2003 cursor-control data set $2a$ [13]. Blanchard and Blankertz [12] described both continuous method and trialwise method using common spatial pattern (CSP) for feature extraction, using Fisher linear discriminant (continuous method) and regularized linear discriminant (trialwise method) for classification. The results derived from their methods won the BCI Competition 2003 for cursor-control data set $2a$. As proposed in [12], a simple solution to solve the unlabeled problem for continuously predicting cursor movement is to only use trials of top and bottom targets for training the classifier. In this case, they can further label the training data set by assuming that in trials of top target the subject would try to make the cursor always go up. Then, they can use the classifier trained using partial training data set (top and bottom targets) to perform 4 targets cursor control. As we have seen from the above discussion, [12] simplified the unlabeled problem by reducing the number of labels from 2 (up and down) to 1 (either up or down depending on the target). We feel that the simplification in [12] is done at trial level, while the actual cursor control is carried out at finer time interval. For target on top, although the cursor has to go up to reach the final target, it is not necessarily true that the cursor always goes up at all time intervals.

In this paper, we propose a statistical learning method towards fully exploiting information contained in the BCI data. First we divide the training data set into segments whose labels are not known, and then represent the training data set by assigning a probability to the possible movement (the label of the segment) at each time interval. Then the unlabeled problem is solved by treating the uncertain labels as hidden variables in the lower bound on the log posterior, and maximizing this lower bound via an EM-like algorithm. The proposed algorithm can make full use of the incomplete data without the need for specifying a distribution for the unknown label. We tested our method on the BCI Competition 2003 cursor-control data set $2a$. The results show that the averaged classification accuracy of the proposed method is among the best.

The rest sections of this paper are organized as follows. The EM algorithm is reviewed in the lower bound point of view in Section 2. We derive the proposed algorithm in Section 3. In Section 4, we apply the proposed method in EEG-based 1D four-targets cursor-control scenario. The experimental results are analyzed in Section 5. Conclusions are drawn in Section 6.

2. THE LOWER BOUND INTERPRETATION OF EM ALGORITHM AND ITS EXTENSION

The expectation-maximization (EM) algorithm [14] is an iterative optimization algorithm specifically designed for the probabilistic models with hidden variables. In this section, we briefly review the lower bound form of the EM algorithm and its extension. Suppose that Z is the observed random variable, Y is the hidden (unobserved) variable, and θ is the model parameter we want to estimate. The maximum a posteriori (MAP) estimation concerns maximizing the posterior, or equally the logarithm of the posterior as follows:

$$L(\theta) = \ln P(Z,\theta) = \ln \sum_{Y} P(Z,Y,\theta). \tag{1}$$

Generally, the existence of the hidden variable Y will induce the dependencies between the parameters of the model. Moreover, when the number of hidden variable is large, the sum over Y is intractable. Thus it is difficult to maximize $L(\theta)$ directly.

To simplify the maximization of $L(\theta)$, we derive a lower bound on L by introducing an auxiliary distribution Q_Y over the hidden variable as follows:

$$\begin{aligned} L(\theta) &= \ln \sum_{Y} P(Z,Y,\theta) = \ln \sum_{Y} Q_Y(Y)\frac{P(Z,Y,\theta)}{Q_Y(Y)} \\ &\geq \sum_{Y} Q_Y(Y)\ln\frac{P(Z,Y,\theta)}{Q_Y(Y)} = F(Q_Y,\theta), \end{aligned} \tag{2}$$

where we have made use of Jensen's inequality. Then the maximization of $L(\theta)$ can be performed by the following two steps:

$$\begin{aligned} &\text{E-step:} \quad Q_Y^{(n+1)} \longleftarrow \arg\max_{Q_Y}\left[F\left(Q_Y,\theta^{(n)}\right)\right], \\ &\text{M-step:} \quad \theta^{(n+1)} \longleftarrow \arg\max_{\theta}\left[F\left(Q_Y^{(n+1)},\theta\right)\right]. \end{aligned} \tag{3}$$

This is the well-known lower bound derivation of the EM algorithm: $F(Q_Y,\theta)$ is the lower bound of $L(\theta)$ for any distribution Q_Y, attaining equality after each E-step. This can be proved by maximizing the lower bound $F(Q_Y,\theta)$ without putting any constraints on the distribution Q_Y:

$$P(Y \mid Z, \theta) = \arg\max_{Q_Y} [F(Q_Y, \theta)]. \tag{4}$$

Then the E-step can be rewritten as follows:

$$\text{E-step:} \quad Q_Y^{(n+1)} \longleftarrow P(Y \mid Z, \theta^{(n)}). \tag{5}$$

Furthermore, combining (2) and (5), we obtain

$$L(\theta^{(n)}) = F(Q_Y^{(n+1)}, \theta^{(n)}). \tag{6}$$

More detailed discussions on the lower bound interpretation of EM algorithm can be found in [15].

However, for many interesting models it is intractable to compute the full conditional distribution $P(Y \mid Z, \theta)$. In these cases we can put constraints on Q_Y (e.g., parameterizing Q_Y to be a tractable form) and still perform the above EM steps to estimate θ. But in general under these constraints of Q_Y, (4) is no longer held. This kind of algorithms which can be viewed as a computationally tractable approximation to the EM algorithm has been introduced in [16].

3. THE PROPOSED ALGORITHM FOR CONTINUOUS CURSOR PREDICTION

In this section, we propose a statistical framework to fully exploit information contained in the BCI data by solving the unlabeled problem based on the EM algorithm. First we formulate the learning problem as follows. Let $D = \{x_i, z_i\}_{i=1}^{N_D}$ stand for the learning data set of N_D independent and identically distributed (i.i.d.) items, where x_i denotes the ith trial and z_i denotes the target label of the ith trial. For continuous prediction, each trial is divided into certain number of segments. Let $x_i = \{x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ}\}$, where x_{ij} denotes the jth segment of the ith trial and J is the total number of the segments in a trial. Let $y_{ij} \in \Phi$ denote the label of x_{ij}, where Φ is the label set of segments. In this learning problem the segment label y_{ij} is hidden. Let θ denote the parameters of classifier which maps the input space of x_{ij} into label set Φ.

Based on the Bayesian theorem, parameter θ can be estimated under MAP criterion:

$$\begin{aligned} \arg\max_{\theta} P(\theta \mid D) &= \arg\max_{\theta} [P(D, \theta)] \\ &= \arg\max_{\theta} [P(D \mid \theta) P(\theta)]. \end{aligned} \tag{7}$$

Under the i.i.d. assumption of data set D, the likelihood $P(D \mid \theta)$ can be formulated as follows (strictly we only model the distribution of $\{z_i\}_{i=1}^{N_D}$ as suggested in [17], which falls into conditional Bayesian inference described in [18]):

$$P(D \mid \theta) = \prod_i P(z_i \mid x_i, \theta). \tag{8}$$

To estimate parameter θ, since the label of each segment is not known exactly, we sum the joint probability $P(z_i, y_{i,j=1:J} \mid x_i, \theta)$ on the hidden labels and model $P(z_i \mid x_i, \theta)$ as follows:

$$P(z_i \mid x_i, \theta) = \sum_{y_{i,j=1:J}} P(z_i, y_{i,j=1:J} \mid x_i, \theta), \tag{9}$$

$$\begin{aligned} P(z_i, y_{i,j=1:J} \mid x_i, \theta) &= P(z_i \mid y_{i,j=1:J}) P(y_{i,j=1:J} \mid x_{i,j=1:J}, \theta) \\ &= \frac{P(y_{i,j=1:J} \mid z_i) P(z_i)}{P(y_{i,j=1:J})} P(y_{i,j=1:J} \mid x_{i,j=1:J}, \theta) \\ &= \frac{1}{Z_D} \prod_{j=1}^{J} [P(y_{ij} \mid z_i) P(y_{ij} \mid x_{ij}, \theta)], \end{aligned} \tag{10}$$

where $y_{i,j=1:J}$ denotes variable set $\{y_{i1}, \ldots, y_{ij}, \ldots, y_{iJ}\}$, $x_{i,j=1:J}$ denotes variable set $\{x_{i1}, \ldots, x_{ij}, \ldots, x_{iJ}\}$, and in the last step of (10) the priors are set to be uniform distribution and the posteriors over hidden variables are fully factorized. From the above equations we obtain the logarithm of the posterior as follows:

$$\ln P(D, \theta) = \sum_{i,j} \ln \sum_{y_{ij} \in \Phi} [P(y_{ij} \mid z_i) P(y_{ij} \mid x_{ij}, \theta)] + \ln P(\theta), \tag{11}$$

where some constants are omitted.

To derive a lower bound on $\ln P(D, \theta)$, we introduce the auxiliary distribution $Q(y_{ij} \mid z_i)$. It should be noted that the function form of $Q(y_{ij} \mid z_i)$ is only determined by the value of z_i. Then according to (2), we obtain the lower bound as follows:

$$\sum_{i,j} \sum_{y_{ij} \in \Phi} Q(y_{ij} \mid z_i) \ln \frac{P(y_{ij} \mid z_i) P(y_{ij} \mid x_{ij}, \theta)}{Q(y_{ij} \mid z_i)} + \ln P(\theta). \tag{12}$$

And as suggested in [17], the prior $P(\theta)$ is modeled as the Gaussian distribution:

$$P(\theta) = N[0, \alpha^{-1} I], \tag{13}$$

where α is the precision parameter.

Therefore, by performing (3), the estimation of $Q(y_{ij} \mid z_i)$ and θ can be achieved via the following EM steps:

M-step:

$$\begin{aligned} \theta^{(n+1)} = \arg\max_{\theta} \Bigg\{ &\sum_{i,j} \sum_{y_{ij} \in \Phi} Q^{(n)}(y_{ij} \mid z_i) \\ &\times \ln \frac{P(y_{ij} \mid x_{ij}, \theta)}{Q}^{(n)} (y_{ij} \mid z_i) - \frac{\alpha}{2} \theta^T \theta \Bigg\}, \end{aligned} \tag{14}$$

E-step:

$$\begin{aligned} &Q^{(n+1)}(y_{ij} \mid z_i) \\ &= \frac{P(y_{ij} \mid z_i) \sqrt[N_{z_i}]{\prod_{m \in \{m \mid z_m = z_i\}} \prod_n P(y_{mn} = y_{ij} \mid x_{mn}, \theta^{(n+1)})}}{\sum_{y_{mn} \in \Phi} P(y_{mn} \mid z_i) \sqrt[N_{z_i}]{\prod_{m \in \{m \mid z_m = z_i\}} \prod_n P(y_{mn} \mid x_{mn}, \theta^{(n+1)})}}, \end{aligned} \tag{15}$$

where N_{z_i} is the total number of segments belonging to the trials having the same target label z_i.

To see further about the proposed algorithm, we rewrite (14) as follows:

$$\theta^{(n+1)} = \arg\min_{\theta} \left\{ \sum_{i,j} \sum_{y_{ij} \in \Phi} Q^{(n)}(y_{ij} \mid z_i) \times \ln \frac{Q^{(n)}(y_{ij} \mid z_i)}{P(y_{ij} \mid x_{ij}, \theta)} + \frac{\alpha}{2} \theta^T \theta \right\}, \tag{16}$$

where the precision parameter α here acts as a regularization constant. From (16) we can see that in the M-step $Q^{(n)}(y_{ij} \mid z_i)$ is used to supervize the optimization process by minimizing the Kullback-Leibler distance between $Q^{(n)}(y_{ij} \mid z_i)$ and $P(y_{ij} \mid x_{ij}, \theta)$ according to θ, which will let $P(y_{ij} \mid x_{ij}, \theta)$ close to $Q^{(n)}(y_{ij} \mid z_i)$. In the E-step, $Q^{(n+1)}(y_{ij} \mid z_i)$ is iteratively updated by considering both the prior knowledge $P(y_{ij} \mid z_i)$ and the information extracted from training data. Moreover, in binary classification case, let $\Phi = \{C_1, C_0\}$ be the segment label set, where C_1, C_0 stand for the two classes. If we assume the label of x_{ij} is known and set $Q^{(n)}(y_{ij} \mid z_i)$ to be delta function $\delta(y_{ij}, C_1)$, then the above algorithm will degenerate to

$$\theta_{\text{MAP}} = \arg\min_{\theta} \left\{ -\sum_{i,j} \left[\delta(y_{ij}, C_1) \ln P(C_1 \mid x_{ij}, \theta) + (1 - \delta(y_{ij}, C_1)) \times \ln (1 - P(C_1 \mid x_{ij}, \theta)) \right] + \frac{\alpha}{2} \theta^T \theta \right\}, \tag{17}$$

where θ_{MAP} is the MAP estimation of parameter θ. This criterion has been successfully used in [10]. For comparison, we take this method as baseline.

To estimate θ in the M-step, we have to model classifier $P(y_{ij} \mid x_{ij}, \theta)$ first. For simplicity let us model $P(y_{ij} \mid x_{ij}, \theta)$ in binary classification case as follows:

$$P(C_1 \mid x_{ij}, \theta) = \frac{1}{[1 + \exp(-\theta^T x_{ij})]} = g(\theta^T x_{ij}) = g(a)$$
$$P(C_0 \mid x_{ij}, \theta) = 1 - P(C_1 \mid x_{ij}, \theta), \tag{18}$$

where a denotes $\theta^T x_{ij}$. Based on this logistic model, in the M-step we use conjugate gradient algorithm to find the minimum of the target function in (16) and then update the regularization constant α as part of the Bayesian learning paradigm using a second level of Bayesian inference [19], as follows:

$$\alpha^{(n+1)} = \frac{(K - \alpha^{(n)} \operatorname{trace}(H^{-1}))}{\theta^T \theta}, \tag{19}$$

where K is the dimension of θ, H is Hessian matrix.

We summarize the proposed algorithm as follows.

(1) n =0, set the initial values $\theta^{(0)}$, $Q^{(0)}(y_{ij} \mid z_i)$, $\alpha^{(0)}$, and the prior $P(y_{ij} \mid z_i)$.

(2) Perform conjugate gradient algorithm on the target function in (16) to estimate parameter $\theta^{(n+1)}$ and update $\alpha^{(n+1)}$ using (19).

(3) Update $Q^{(n+1)}(y_{ij} \mid z_i)$ using (15).

(4) $n = n + 1$, go to (2) until

$$\left| Q^{(n+1)}(C_1 \mid z_i) - Q^{(n)}(C_1 \mid z_i) \right| < \text{threshold}_P,$$
$$\left| \frac{[\alpha^{(n+1)} - \alpha^{(n)}]}{\alpha^{(n)}} \right| < \text{threshold}_\alpha. \tag{20}$$

4. IMPLEMENTATION OF THE PROPOSED ALGORITHM

In this section we evaluated the proposed algorithm in cursor control problem, specifically for 1D four-targets cursor control based on mu/beta rhythms.

4.1. *Feature extraction*

Our aim in the feature extraction part is to increase the SNR ratio and extract the relevant features centralizing on the alpha and beta bands from EEG data. The EEG inputs were sampled at 160 Hz and enhanced using a band pass IIR filter with the pass band around 9–31 Hz. Then, the common spatial patterns (CSP) analysis was performed on the samples. In binary classification case, CSP analysis [20] can derive weights for linear combinations of the data collected from every channel to get several (usually four) most discriminative spatial components. In our algorithm, the data belonging to target 1 and 4 served as the two classes. Since not all the channels were relevant for predicting cursor movement, only a subset of channels was used to do CSP analysis for each subject. Moreover, in this paper we just transformed EEG signal into the subspace of the most discriminative CSP spatial components. After the above processing, we assume that the EEG signal of each trial is in a 4 × 368 matrix, where 4 is the number of CSP components and 368 is the length of each trial. Then the whole trials x_i were blocked into overlapping segments of 300 milliseconds (48 samples) in duration where the overlap was set to 100 milliseconds (16 samples). Therefore, the data matrix of each segment is 4 × 48. The relevant spectral power features were extracted from each segment after performing FFT on the roles of the segment matrix. Furthermore, in order to regard the bias weight of linear part $\theta^T x_{ij}$ in the classifier as an element of the parameter vector θ, a constant element "1" is added at the end of the feature vector. Finally these feature vectors were transmitted to the classifier for further processing.

4.2. *Classification algorithm*

The central part of our BCI system is the classification algorithm. We applied the proposed classifier here to translate feature vectors x_{ij} into commands to control cursor movement. However under the above model, the output of classifier $P(y_{ij} \mid x_{ij}, \theta)$ has closed relation with parameter θ. Thus the estimation error of θ will make the output of the classifier overconfident. To solve this problem, we adopt the Bayesian learning treatment as suggested in [21] to integrate out the parameter θ and obtain the modified classifier as follows:

$$P(C_1 \mid x_{ij}, D) \simeq g(k(\sigma^2(x_{ij}))a_{\text{MAP}}(x_{ij})), \tag{21}$$

where $a_{\text{MAP}}(x_{ij}) = \theta_{\text{MAP}}^T x_{ij}$, $k(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}$, and $\sigma^2 = x_{ij}^T H^{-1} x_{ij}$.

In the training period, the prior $P(y_{ij} \mid z_i)$ is set to be flat. The initial values and thresholds are set as follows: $\theta^{(0)} = 0$, $\alpha^{(0)} = 0.5$, $\text{threshold}_P = 0.05$, and $\text{threshold}_\alpha = 0.01$. The setting of initial value $Q^{(0)}(y_{ij} \mid z_i)$ will be further discussed in the experimental part.

4.3. Control and decision

Let d_{ij} denote the displacement of cursor movement at the jth time interval of the ith trial, then we obtain

$$d_{ij} = \frac{[P(C_1 \mid x_{ij}, D) - 0.5]}{J}, \tag{22}$$

where J is the total number of the segments of the ith trial. Then we formulated the vertical displacement D_{ij} between the middle line of the screen and the cursor at the jth-time interval of the ith trial as follows:

$$D_{ij} = \sum_{k=1}^{j} d_{ik}. \tag{23}$$

To evaluate the performance of our algorithm, three thresholds $t_3 < t_2 < t_1$ were chosen to classify the final distance D_{iJ} into four categories, such that trial x_i belongs to target 1 if $t_1 < D_{iJ}$, and trial x_i belongs to target 2 if $t_2 < D_{iJ} < t_1$, and so forth. Since t_i is scale variable and $D_{iJ} \in [-0.5, 0.5]$, we perform one-dimensional search for each t_i according to the classification accuracy between neighbor targets in training period, for example, t_1 is set to achieve the best accuracy between targets 1 and 2.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, we tested it on the BCI Competition 2003 data set 2*a*. This data set consists of ten 30-minutes sessions for each of three subjects (AA, BB, CC). In each session, there are 192 trials. The training set consists of all the trials of 1–6 sessions. The test set consists of 7–10 sessions. Both the proposed method and two state-of-the-art methods: Bayesian logistic regression (baseline) and Fisher linear discriminant (FLD) [17], were applied on this data set. In the proposed method, all the trials of the first six sections were used to train the model. The rest sections were used for testing. To set the initial value of $Q(y_{ij}|z_i)$, a six-fold cross-validation was performed on the training data set. In each fold we trained the classifier on five sections and tested on the section which was left out. This procedure was then repeated until all the sections had been tested. Since in the baseline and the FLD methods we had to assign label to each segment, as proposed in [12], we assumed the labels for the segments of target 1 belong to C_1 and the labels of target 4 belong to C_0, and used the trials belonging to the first and fourth targets of the first six sections to train the classifier. In all methods, first, the spatial and spectral features were extracted from the EEG data. Then in the training stage, the model parameter θ was estimated and the three thresholds $\{t_i\}_{i=1,2,3}$ were chosen. In the testing stage, we calculated $P(C_1 \mid x_{ij}, D)$ to control the cursor at each time interval. In the end, the final distance D_{iJ} was classified using the thresholds. The accuracy was measured by N_1/N_2, where N_1 is the number of times D_{iJ} falls into the correct interval and N_2 is the total number of tests.

Table 1: A comparison of classification accuracies and information transfer rates of different methods for different subjects.

	Accuracy (%)				Trans. rate (bits/trial)
	AA	BB	CC	Avg	
Proposed	71.1	67.6	71.2	70	0.643
Baseline	68.4	63.5	66.3	66.1	0.539
FLD	70.1	65.2	68.9	68.1	0.591

5.1. Results and comparisons

In order to benchmark the performance of the proposed algorithm, the averaged accuracies of each method are listed in Table 1, where “Avg” denotes the averaged accuracy over all the subjects. We also converted the overall classification accuracy into information transfer rate as proposed in [9] by using

$$B = \log_2 N + p \log_2 p + (1 - p) \log_2 \left[\frac{1 - p}{N - 1}\right], \tag{24}$$

where B is bits, N is the number of possible targets (four in this case), and p is the probability that the target will be hit (i.e., accuracy). From Table 1 we can see that the proposed method outperforms all the other methods on every subject. The improvement of the averaged accuracy over all subjects is up to 4%. Furthermore, the information transfer rate is increased from 0.539 to 0.643 bits/trial, the improvement is 19% which is considerable for the BCI communication system. The above results also show that the performance of the proposed method is comparable to the most recent methods, such as Tsinghua’s method (66.0%) [11], Blanchard’s continuous method (68.8%), and trial-wise method (71.8%) [12].

To further study the performance of the proposed method, we illustrate the accuracies of individual tasks in Figure 2 and compare them with the baseline method.

From Figure 2 we can see that for the middle targets (tasks 2 and 3) which are difficult to reach, the proposed method outperforms baseline method clearly for all the subjects. However for the top and the bottom targets (tasks 1 and 4), the performance improvements are not consistent. These results show that since we incorporate the unlabeled segments of tasks 2 and 3 in the training procedure based on EM algorithm, the information extracted from the training data set improves the control performance for the middle targets significantly. But on the other hand this may also hamper the performance improvement for the top and the

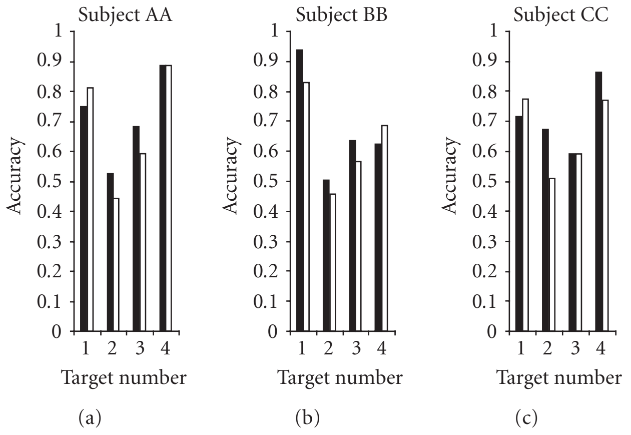


FIGURE 2: A comparison of the classification accuracy of individual task of the proposed method (black) and the baseline (white).

bottom targets according to the baseline. Due to the above reasons, the overall improvement of the proposed method is not always significant as compared with other methods, although the proposed method performs more steadily than baseline method on different targets.

Taking the above discussions as a whole, we can see that the performance of the BCI system is improved by handling the uncertainty of segment label properly. The classification accuracy of our method is among the highest.

5.2. A comparison of error rates with different initial probabilities

To study the effects of the initial probability, we performed a six-fold cross-validation on different settings of $Q^{(0)}(y_{ij} \mid z_i)$. In binary classification case, only one initial value, $Q^{(0)}(y_{ij} = C_1 \mid z_i)$, needs to be set for individual target z_i. Thus, we only need to set four initial probabilities with respect to different targets z_i in a six-fold cross-validation. For simplicity, in the rest of this section, we use $Q^{(0)}(z_i)$ instead of $Q^{(0)}(y_{ij} = C_1 \mid z_i)$. These initial probabilities are set as follows: first the initial value for the trials belonging to target one, $Q^{(0)}(z_i = 1)$, is set by using our prior knowledge. Then the rest initial values are set as follows:

$$\begin{aligned} Q^{(0)}(z_i = 4) &= 1 - Q^{(0)}(z_i = 1), \\ Q^{(0)}(z_i = 2) &= \frac{[2 \times Q^{(0)}(z_i = 1) + Q^{(0)}(z_i = 4)]}{3}, \\ Q^{(0)}(z_i = 3) &= \frac{[Q^{(0)}(z_i = 1) + 2 \times Q^{(0)}(z_i = 4)]}{3}. \end{aligned} \tag{25}$$

In the rest of this section, we take subject CC as an example to show the effects of initial probabilities. In our experiment, the initial probability $Q^{(0)}(z_i = 1)$ was increased 0.1 at each step from 0.6 to 1. At each step, a six-fold cross-validation was performed on the training data set. Therefore we got six convergence values of the initial probability for each target. Then, we calculated the mean and standard deviation of the convergence values for each target. The experimental results with different initial probabilities are depicted in Figure 3. The convergence property of initial probability is illustrated in Figure 4.

In the left of Figure 3, we compare the error rates (ERs) in three conditions: (i) the proposed algorithm (black), (ii) the proposed algorithm without updating initial probability (gray), and (iii) the baseline algorithm (white). Since there are no initial probabilities in the baseline method, the ERs of the baseline method are the same at each step. From the left of Figure 3 we can see that at every initial probability the performance of the BCI system is greatly improved by updating the initial probability iteratively and the ER reaches its minimum at $Q^{(0)}(z_i = 1) = 0.8$. Furthermore, without updating initial probability the ERs of our method are still lower than those of baseline method. Therefore the results in the left of Figure 3 confirm that it is effective to introduce $Q(y_{ij} \mid z_i)$ in the proposed algorithm to improve the performance of the classifier. In the right of Figure 3 we further compare the ERs in the first two conditions described above. The comparison is detailed to the error rates of different targets at different initial probabilities. In the right of Figure 3 ERs are significantly reduced on targets 2 and 3 by updating initial probability iteratively. For target 1 the improvement is slight, and for target 4 the performance is enhanced at $Q^{(0)}(z_i = 1) = 0.8$.

It is important to choose initial probability for the proposed algorithm. From Figure 4 we can see that when $Q^{(0)}(z_i = 1)$ is small (near 0.6), most of the initial probabilities are not changed after update. Thus in this case, the benefits of the update procedure are reduced, especially for targets 2 and 3 (right of Figure 3) and the averaged ER reaches its maximum (left of Figure 3, black bar). In the other case, when $Q^{(0)}(z_i = 1)$ is large (near 1), the standard deviations of the convergence probabilities in the six-fold cross-validation are increased, which means that the convergence values of the same initial probability are not consistent. This hampers the performance improvement of the classifier, which is confirmed by the fact that when $Q^{(0)}(z_i = 1)$ approaches 1, the ERs are increased (left of Figure 3, black bar). Therefore, the experimental results show that $Q^{(0)}(z_i = 1) = 0.8$ is the best initial value for this subject.

From the above discussions, we can see that although ERs vary with the initial values of $Q^{(0)}(z_i = 1)$, the proposed optimization algorithm clearly improves the performance of cursor control system at every initial probability, especially for the targets in the middle position. By choosing proper initial probability, the performance of the proposed algorithm can be improved.

5.3. A further study of the efficacy of the proposed algorithm

In this section, we demonstrate the efficacy of the proposed algorithm more in depth in two aspects. In the first aspect, we illustrate the control performance during a trial for

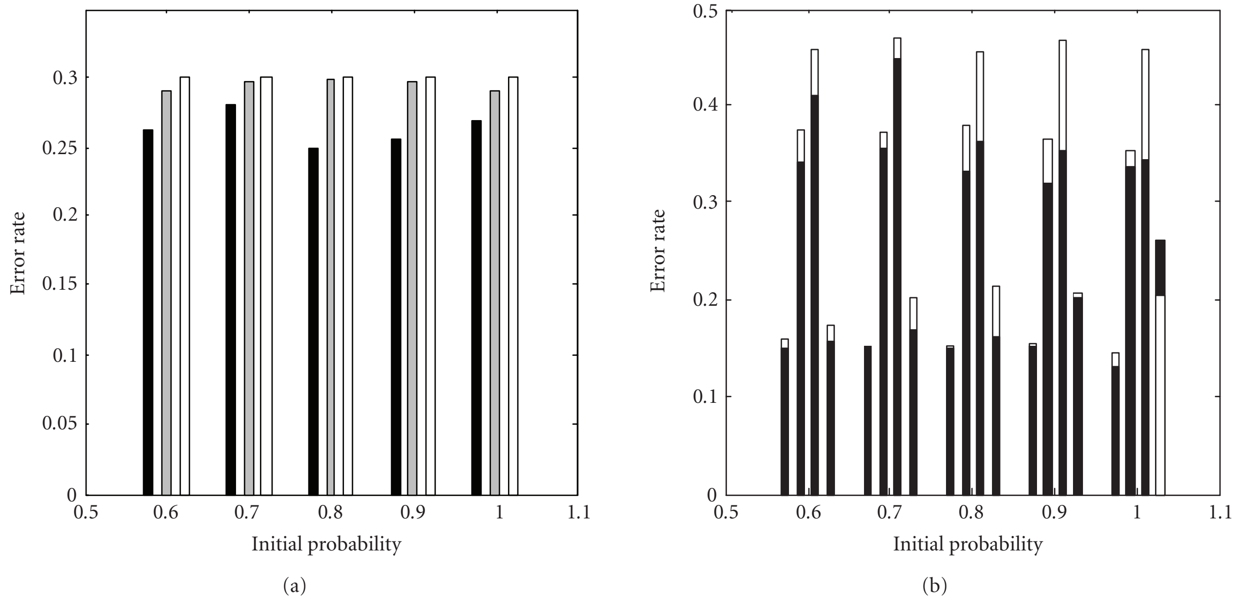


Figure 3: A comparison of the error rates (ERs) with different initial probabilities for subject CC. In the left, we compare the ERs averaged over targets with different initial probabilities in three conditions, the proposed method (black), the proposed method without updating initial probability (gray), and the baseline method (white). In the right, we compare ERs with and without updating initial probability in detail. There are five groups of error bars, and each group contains the error bars of the four targets (targets 1 to 4, from left to right). In each group, the bars with white top indicate that ER is reduced after update, and the length of the white part denotes that the amount of ER has been reduced. Similarly, the bars with black top indicate that ER is increased after update. The bars which are all black indicate that ER is unchanged after update.

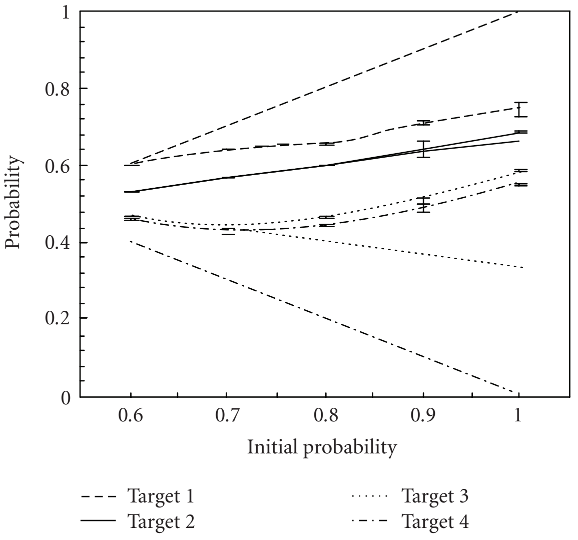
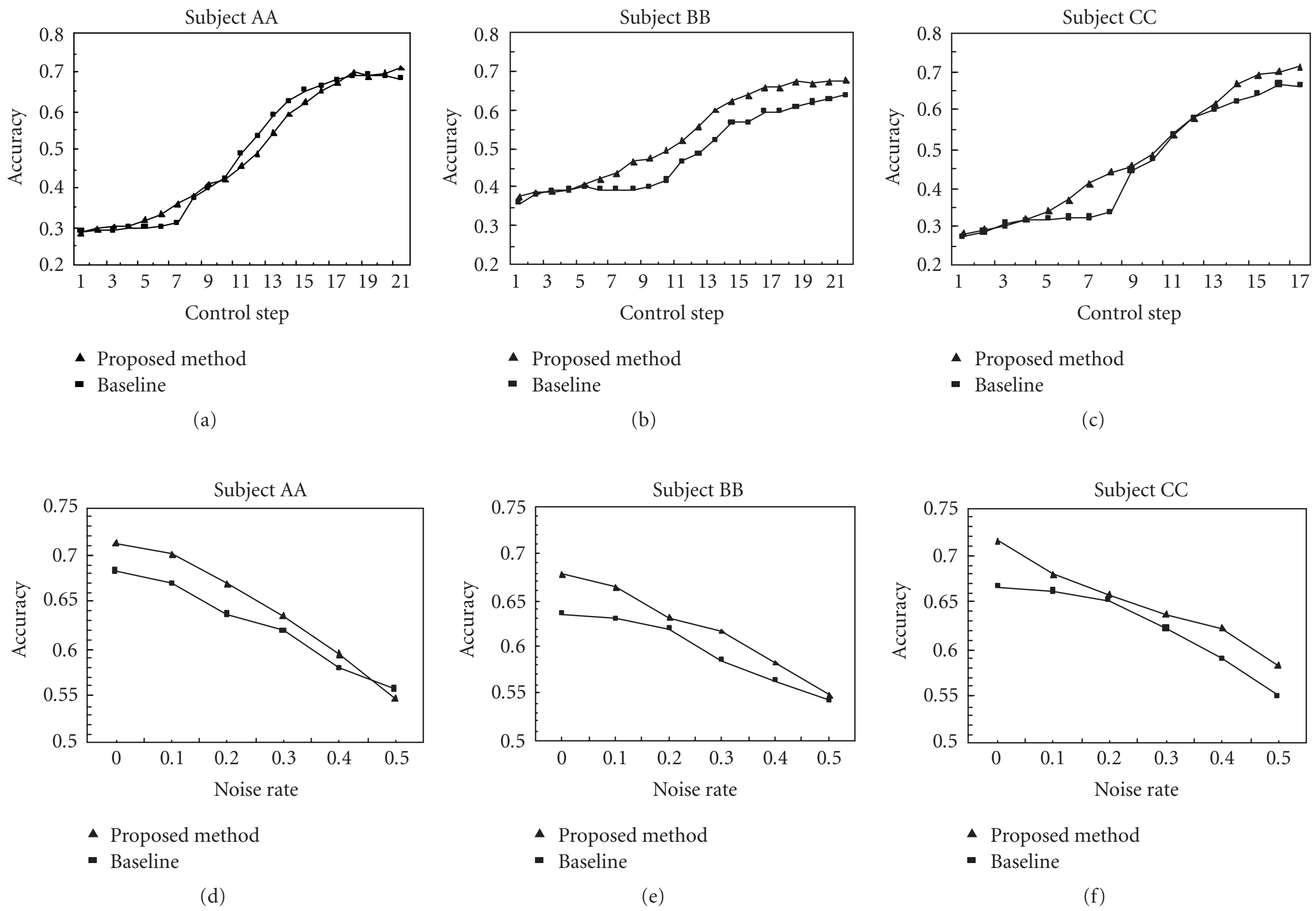


Figure 4: The convergence property of initial probability for subject CC. In this figure, we draw the mean values of the convergent probabilities and mark them with standard deviations at different initial values. For comparison, we also depict the curves of initial probability.

individual subject by categorizing D_{ij} into the four targets using the estimated thresholds at each control step. The results of the averaged cursor control accuracies are illustrated at the top of Figure 5. It shows that for all the subjects the accuracy increases sharply during the middle of the performance, which causes the form of the accuracy curves to be sigmoid. From the top of Figure 5, we can see that for subject AA, these two methods perform closely. While for the other two subjects, the proposed method (▲) performs clearly better than the baseline method (■) during the whole trial. Especially, for subject BB the classification accuracies are much higher than those of baseline almost at every control step from the beginning of the trial. These improvements indicate that by using the proposed method to translate EEG features into commands, one can achieve better performance consuming less time. This character is important for the EEG-based online cursor-control system.

In the second aspect, we manually corrupt the target labels of the training data with some fixed noise rate and show the effects of the noise rate with respect to the baseline method at the bottom of Figure 5. For each subject, the noise rate was increased 0.1 at each step from 0 to 0.5. The results show that although increasing the mislabel rate decreases the performance of both the two methods, the classification accuracies are much better than the random accuracy 25% (in four targets case), even the training data is half corrupted.

Furthermore, by comparing the two methods at the bottom of Figure 5, firstly, we can see that the proposed algorithm outperforms the baseline clearly almost at every noise rate on all the subjects by extracting information from the corrupted data effectively based on the EM algorithm.



(a) (b) (c)

(d) (e) (f)

Figure 5: We further study the efficacy of the proposed algorithm in two aspects. At the top, we compare the control performance between the proposed method (▲) and the baseline (■). At the bottom, the classification accuracies at different noise rates are compared.

Secondly, we also find that when the noise rate increases the advantage of the proposed algorithm is reduced, which is due to the fact that the proposed method has more parameters to be optimized than the baseline.

6. CONCLUSIONS

In this paper, we proposed a novel statistical learning method based on the EM algorithm to learn parameters of a classifier under MAP criterion. In most of the current methods the authors labeled the segments of the EEG data empirically. This will lead to the under-use of the training data. In the proposed method, we solved the "unlabeled problem" by treating the uncertain labels as the hidden variables in the lower bound on the log posterior. The parameters of the model were estimated by maximizing this lower bound using an EM-like algorithm. By solving the unlabeled problem, the proposed method can fully exploit information contained in the BCI data and improve the performance of the cursor control system. The experimental results have shown that the averaged classification accuracy of the proposed algorithm is higher than the results of other widely used methods up to 4% and the information transfer rate is improved up to 19%. Furthermore, the proposed method can achieve better performance consuming less time than the baseline, which is a desirable property for online application.

Moreover, our algorithm still has the potentials to be improved. From (10) we can see that the proposed criterion is based on the complete factorization of the likelihood $P(D \mid \theta)$. Thus in our method the dependence between neighbor segments has not been considered. While brain is a complex dynamic system, and EEG signal is a typical kind of nonstationary time series. Thus our proposed model is an approximation of the actual one. Therefore, one of the research directions is to add the dependence between segments (or predictions) into our model to model the nonstationary property of the EEG signal. As a final remark, although our method is derived to solve the cursor control problem of BCI system, the same formulation can also be used to handle the "unlabeled problem" in other pattern recognition systems.

ACKNOWLEDGMENTS

The authors are grateful to the reviewers for many helpful suggestions for improving this paper. The authors would like to thank Dr. Yuanqing Li, Dr. Manoj Thulasidas, and

Mr Wenjie Xu for their fruitful discussions, and to thank Wadsworth Center, NYS, Department of Health for providing the data set.

REFERENCES

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[2] N. Birbaumer, T. Hinterberger, A. Kübler, and N. Neumann, "The thought-translation device (TTD): neurobehavioral mechanisms and clinical outcome," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 120–123, 2003.

[3] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[4] L. A. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[5] P. Meinicke, M. Kaper, F. Hoppe, M. Heumann, and H. Ritter, "Improving transfer rates in brain computer interfacing: a case study," in *Advances in Neural Information Processing Systems*, pp. 1107–1114, MIT Press, Cambridge, Mass, USA, 2003.

[6] M. Middendorf, G. McMillan, G. Calhoun, and K. S. Jones, "Brain-computer interfaces based on the steady-state visual-evoked response," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 211–214, 2000.

[7] J. R. Wolpaw, D. J. McFarland, T. M. Vaughan, and G. Schalk, "The Wadsworth Center brain-computer interface (BCI) research and development program," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 204–207, 2003.

[8] J. R. Wolpaw and D. J. McFarland, "Multichannel EEG-based brain-computer communication," *Electroencephalography and Clinical Neurophysiology*, vol. 90, no. 6, pp. 444–449, 1994.

[9] D. J. McFarland and J. R. Wolpaw, "EEG-based communication and control: speed-accuracy relationships," *Applied Psychophysiology Biofeedback*, vol. 28, no. 3, pp. 217–231, 2003.

[10] S. J. Roberts and W. D. Penny, "Real-time brain-computer interfacing: a preliminary study using Bayesian learning," *Medical and Biological Engineering and Computing*, vol. 38, no. 1, pp. 56–61, 2000.

[11] M. Cheng, W. Jia, X. Gao, S. Gao, and F. Yang, "Mu rhythm-based cursor control: an offline analysis," *Clinical Neurophysiology*, vol. 115, no. 4, pp. 745–751, 2004.

[12] G. Blanchard and B. Blankertz, "BCI competition 2003-data set IIa: spatial patterns of self-controlled brain rhythm modulations," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1062–1066, 2004.

[13] B. Blankertz, K.-R. Müller, G. Curio, et al., "The BCI competition 2003: progress and perspectives in detection and discrimination of EEG single trials," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1044–1051, 2004.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood for incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, pp. 1–38, 1977.

[15] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed., pp. 355–368, Kluwer Academic, Dordrecht, The Netherlands, 1998.

[16] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, Ed., MIT Press, Cambridge, Mass, USA, 1999.

[17] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[18] T. Jebara, *Machine Learning: Discriminative and Generative*, Kluwer Academic, Dordrecht, The Netherlands, 2004.

[19] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[20] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[21] D. J. C. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, vol. 4, no. 5, pp. 698–714, 1992.

Xiaoyuan Zhu was born in Liaoning China, in 1979. He is currently a Ph.D. student in the University of Science and Technology of China (USTC). His research interest focuses on machine learning, Bayesian method, brain (EEG) signal recognition.

Cuntai Guan received his Ph.D. degree in electrical and electronic engineering in 1993. He worked in Southeast University, from 1993–1996, on speech vocoder, speech recognition, and text-to-speech. He was a Visiting Scientist in 1995 at CRIN/CNRS-INRIA, Lorraine, France, working on key word spotting. From September 1996 to September 1997, he was with City University of Hong Kong developing robust speech recognition under noisy environment. From 1997 to 1999, he was with Kent Ridge Digital Labs of Singapore, working on multilingual large vocabulary continuous speech recognition. He spent five years in industries, as a Research Manager and R&D Director, focusing on the development of spoken dialogue technologies. Since 2003, he is a Lead Scientist at the Institute for Infocomm Research, Singapore, heading Neural Signal Processing Lab and Pervasive Signal Processing Department. His current research focuses on investigation and development of effective framework and statistical learning algorithms for the analysis and classification of brain signals. His interests include machine learning, pattern classification, statistical signal processing, brain-computer interface, neural engineering, EEG, and speech processing. He is a Senior Member of the IEEE. He has published more than 50 technical papers.

Jiankang Wu received the B.S. degree from the University of Science and Technology of China, Hefei, and the Ph.D. degree from Tokyo University, Tokyo, Japan. Prior to joining the Institute for Infocomm Research, Singapore, in 1992, he was a Full Professor at the University of Science and Technology of China. He also worked in universities in the USA, UK, Germany, France, and Japan. He is the author of 18 patents, 60 journal publications, and five books. He has received nine distinguished awards from China and the Chinese Academy of Science.

Yimin Cheng was born in Xi'an, China, in 1945, graduated from the University of Science and Technology of China (USTC), Anhui, Hefei, China, in 1969. Currently, he is a Professor at the USTC. His research interests include digital signal processing, medicine image analysis, and computer vesion.

Yixiao Wang was born in 1945. Currently, he is an Associate Professor at USTC. His research interests focus on information hiding, video signal transfer and communication technique, computer vesion, and deep image analysis.