Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition

Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, *Fellow, IEEE*,
and Huiguang He, *Senior Member, IEEE*

***Abstract*—Electroencephalogram (EEG) has been widely used in emotion recognition due to its high temporal resolution and reliability. Since the individual differences of EEG are large, the emotion recognition models could not be shared across persons, and we need to collect new labeled data to train personal models for new users. In some applications, we hope to acquire models for new persons as fast as possible, and reduce the demand for the labeled data amount. To achieve this goal, we propose a multisource transfer learning method, where existing persons are sources, and the new person is the target. The target data are divided into calibration sessions for training and subsequent sessions for test. The first stage of the method is source selection aimed at locating appropriate sources. The second is style transfer mapping, which reduces the EEG differences between the target and each source. We use few labeled data in the calibration sessions to conduct source selection and style transfer. Finally, we integrate the source models to recognize emotions in the subsequent sessions. The experimental results show that the three-category classification accuracy on benchmark SEED improves by 12.72% comparing with the nontransfer method. Our method facilitates the fast deployment of emotion recognition models by reducing the reliance on the labeled data amount, which has practical significance especially in fast-deployment scenarios.**

***Index Terms*—Brain–computer interface, emotion recognition, transfer learning (TL).**

I. Introduction

EMOTION plays an important role in the human–human interaction [1]. In the human–machine interaction (HMI), we also expect machines to communicate with us according to our emotions, where emotion recognition is the key problem [2]. Among the many emotion recognition methods, electroencephalogram (EEG) shows advantage in reliability [3] and accuracy [4]. In recent years, the affective brain–computer interface (aBCI) [5] has attracted great interests in the research field, which endows BCI system with the ability to detect, process, and respond to the affective states of humans using EEG signals [6]. Fig. 1 shows the fundamental structure of aBCI [5]. Emotion stimulus are used to evoke the desired emotions. Recent studies tend to use film stimulus, for films contain both scene and audio, which expose the users to real-life scenarios to elicit strong subjective and physiological changes [7]. EEG signals are recorded during the stimuli, which are used for recognizing emotions.

The individual differences of EEG make it difficult to acquire general models that are applicable across persons [8]–[10]. Therefore, the conventional method is to train new models for new persons [8], [11]. Researchers first collect data in some training sessions and then use these data to train models, which are then used in the subsequent sessions. In some application scenarios, we hope to reduce the time cost in the collection of the training data. However, existing researches have indicated that if the number of training data is small compared to the size of the feature vectors, the model will most probably give poor results [12]. Considering the feature dimension in the multielectrode EEG, acquiring models with few training data is challenging. In this paper, we present a transfer learning (TL) [13] method to explore and exploit information from existing subjects to make up for the insufficiency of the training data. The method is designed to make the target statistically similar to the sources so as to share source models. Unlike conventional methods where there are substantial unlabeled target data [14], we handle with the situation where a small amount of labeled data are available.

Manuscript received September 28, 2018; revised January 11, 2019; accepted February 27, 2019. Date of publication March 27, 2019; date of current version June 16, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 91520202 and Grant 81701785, in part by the Chinese Academy of Sciences (CAS) Scientific Equipment Development Project under Grant YJKYYQ20170050, in part by the Beijing Municipal Science and Technology Commission under Grant Z181100008918010, in part by the Youth Innovation Promotion Association CAS, and in part by the Strategic Priority Research Program of CAS. This paper was recommended by Associate Editor H. A. Abbass. *(Corresponding author: Huiguang He.)*

J. Li and S. Qiu are with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lijinpeng2015@ia.ac.cn; shuang.qiu@ia.ac.cn).

Y.-Y. Shen is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shenyuanyuan2015@ia.ac.cn).

C.-L. Liu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liucl@nlpr.ia.ac.cn).

H. He is with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: huiguang.he@ia.ac.cn).

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/TCYB.2019.2904052

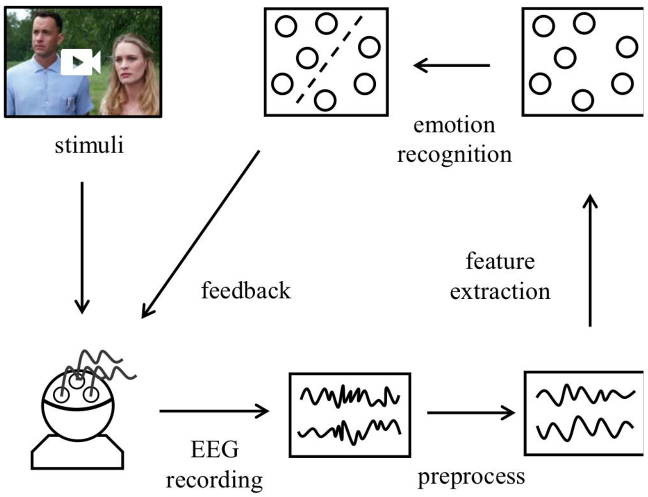


Fig. 1. Fundamental modules for affective brain–computer interface. Emotion recognition plays the key role in the system.

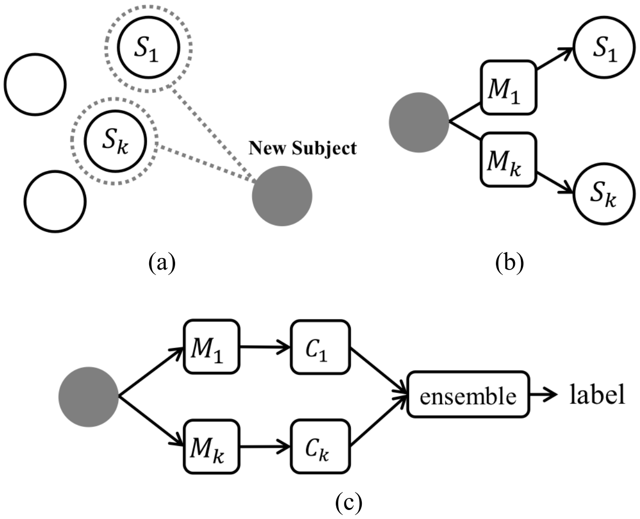


Fig. 2. Multisource TL framework for emotion recognition. *S*: source or subject; *M*: STM; and *C*: individual classifier. (a) Source selection. (b) Learn feature mapping. (c) Classifier ensemble.

Fig. 2 shows the multisource TL method. There are multiple existing subjects and their individual classifiers. For a new subject, we select some suitable sources (e.g., S_1 and S_k) with samples in some calibration sessions and learn style transfer mapping (STM) [15], [16] (M_1, M_k) to reduce the domain differences between the target and each selected source. To minimize the time cost of preliminary experiments, we explore the minimal samples needed in the calibration sessions. During the test, we transform each EEG sample in subsequent sessions with STMs and integrate classifiers (C_1, C_k) to deduce emotion labels. Considering the support vector machines (SVMs) are the most popular classifiers in the EEG-based emotion recognition [6], we use SVM in this paper. There are two key techniques in the framework.

1) *Source Selection:* We use a simple but effective method to locate suitable knowledge-exporting sources.
2) *Mapping Destination in STM Learning:* We explore two types of mapping destination settings.

The rest of this paper is organized as follows. Section II is a brief review of TL and its applications in BCIs. Section III is the multisource TL framework, including source selection and STM. In Section IV, we conduct experiments and present the results to show the superiority of our method. Section V discusses and analyzes the results. Section VI is the conclusion.

II. Related Work

There are two types of generalization problems in BCI.

1) *Session-to-Session Generalization for a Same Subject:* The EEG signals of different sessions show consistency, while there are still differences [7]. EEG has nonstationary characteristic, and we can view every session a slightly new task [8]. Most classical methods to tackle the session-to-session difference were based on the common spatial pattern (CSP) [17], which assumed the existence of a set of linear filters that could work across sessions. For example, Krauledat *et al.* [18] extracted prototypical spatial filters with good generalization properties by clustering in the calibration sessions, and used them to improve the accuracy in the last session. The method reduced the preliminary experiment time at a cost of a slightly drop on the accuracy. Their work was for a same person.
2) *Subject-to-Subject Generalization:* EEG signals are subject-specific and vary considerably between individuals [8]. As a consequence, researchers train individual models for new users. However, the collection of the training data appears to be time-consuming and expensive when there is a need to quickly build a reliable model to track a new user's affective states. The CSP methods have been used to tackle the subject-to-subject differences. For example, Fazli *et al.* [19] constructed an ensemble of classifiers derived from subject-specific temporal and spatial filters using a large database containing 45 subjects in a movement imagination task, and tested the model on a new subject. Their work was offline, which means the label information of the new subject were not taken into consideration. Inspired by the advances in machine learning, Morioka *et al.* [20] used sparse coding to perform the transfer, which learns a sparse representation on multiple subjects and transforms the EEG of a new subject to that space, or finds a stationary subspace with the data recorded across multiple subjects [21], [22]. These approaches excavate and fasten on the common structures to confront the variability, and usually require substantial data.

TL refers to a class of practical algorithms that are able to reduce the dependence on the labeled data when training models for new tasks by using prior knowledge concluded from relevant sources. According to Pan and Yang *et al.* [13], there are three most commonly used TL methods.

1) Instance transfer, which uses reweighted samples from the source to help the model training in the target [23], [24].

TABLE I
NOTATION SUMMARY

Notation	Meaning
$\boldsymbol{A}^S$	source data, $\boldsymbol{A}^S = \{\boldsymbol{A}^{Sp}, p = 1, \ldots, N\}$
$\boldsymbol{C}^S$	source classifiers, $\boldsymbol{C}^S = \{\boldsymbol{C}^{Sp}, p = 1, \ldots, N\}$
$\boldsymbol{A}_L^T$	labeled target data, $\{x_j, y_j\}_{j=1}^n$
$\boldsymbol{A}_U^T$	unlabeled target data, $\{x_j, y_j\}_{j=n+1}^u$
$\boldsymbol{A}^T$	target data, $\boldsymbol{A}_L^T \cup \boldsymbol{A}_U^T$

2) Feature transfer [25], [26], where the transferred knowledge is encoded into the feature representation [9].
3) Parameter transfer [27], [28], which assumes the target shares parameters with the source in different tasks.

In recent years, researchers have applied TL to aBCIs for subject-to-subject generalization.

Zheng and Lu [9] personalized emotion recognition models by adopting various TL techniques, including transductive component analysis (TCA) [25], transductive SVM (TSVM) [29], and transductive parameter transfer (TPT) [30]. After conducting transfer, the three-category classification accuracy improved significantly comparing with the generic classifier (trained on the combination of source samples) [9]. Their work showed the possibility of expanding the application scope of aBCIs across persons with TL. TCA was a kind of the feature transfer method, where the source and the target were projected to a new space to reduce the discrepancy. It involved the training of new classifier in the new space, and brought additional computational burden for large datasets. TSVM was a type of instance transfer, where the instance reweighting was conducted. TPT learned a regression model to predict classifier parameters based on the data attributes, which required a considerable amount of sources to learn the regression model. The main drawback of their methods is that they used all the data in the target [9], which means all the EEG recordings of the new subject were at hand before doing knowledge transfer. However, in some practical applications, we only have a small amount of data, and plan to recognize his (her) emotion as quickly as possible. The chief motivation of this paper is to adapt existing classifiers to new users with a small amount of labeled data collected in his (her) preliminary experiments.

III. METHODS

A. Source Selection

Existing studies have indicated that brute force leveraging of the sources poorly related to the target may decrease the performance, which is referred to as "negative transfer" [31]. To avoid negative transfer, we select appropriate sources before transfer. For a new subject, we first determine from which sources to borrow knowledge. To make the narrative clearer, we summarize the notations in Table I. $\boldsymbol{A}_L^T$ are labeled target data from some calibration sessions, and $\boldsymbol{A}_U^T$ are unlabeled data to be recognized in subsequent sessions. Since $\boldsymbol{A}_L^T$ has label information, source selection is intuitive. We enumerate the N classifiers in the sources to classify $\boldsymbol{A}_L^T$ and locate the top N_S classifiers with high accuracies. We regard their corresponding training EEG data as appropriate sources. The assumption behind the criteria is that the difference between $\boldsymbol{A}_L^T$ and $\boldsymbol{A}_U^T$ is not large, and we accept it as the data belong to a same subject in a same experiment. Feature mapping functions are learned between the target and each selected source, individually. In the classifier ensemble (test), the selected classifiers are combined according to weights determined by their accuracies on $\boldsymbol{A}_L^T$.

B. Style Transfer Mapping

In the standard TL term, we call the knowledge-exporting side the "source," and the importing side the "target." In STM, we map $\boldsymbol{A}^T$ to $\boldsymbol{A}^{Sp}$ to bridge the two distributions. In addition, we are not going to map $\boldsymbol{A}^T$ to $\boldsymbol{A}^{Sp}$ directly, but to find some representational patterns in $\boldsymbol{A}^{Sp}$ (prototypical clustering centers, class mean values) to map $\boldsymbol{A}^T$ to. For simplicity, we call the representational patterns in $\boldsymbol{A}^{Sp}$ the "destination," and samples in $\boldsymbol{A}^T$ the "origin."

STM is an effective TL method that achieved state-of-the-art performance in style transfer tasks [15], [16]. The objective of STM is to map data from origin to destination via an affine mapping. In this way, the classifier of the destination is more "familiar" with the samples in the origin, and thus yield better performance on it. The destination point set is noted as

$$D = \{d_i \in R^m \mid i = 1, \ldots, n\} \tag{1}$$

where n refers to the data amount in the point set and m is the feature dimensionality. D is composed of the representational patterns of $\boldsymbol{A}^{Sp}$. The mapping origin of STM is

$$O = \{o_i \in R^m \mid i = 1, \ldots, n\}. \tag{2}$$

The change from d_i to o_i is called "concept drift." Suppose d_i is transformed to o_i with confidence $f_i \in [0, 1]$, and we learn the inverse transformation function to transform o_i back to d_i with affine transformation $Ao_i + b$. The parameters $A \in R^{m\times m}$ and $b \in R^m$ are learned by minimizing the weighted squared error with regularization items to avoid overtransfer

$$\min_{A\in R^{m\times m}, b\in R^m} \sum_{i=1}^{n} f_i \|Ao_i + b - d_i\|_2^2 + \beta\|A - I\|_F^2 + \gamma\|b\|_2^2 \tag{3}$$

where $\|\cdot\|_F^2$ is the Frobenius norm of matrix and $\|\cdot\|_2$ is the L_2-norm of vector. The second item of (3) is a constraint on A to prevent it from being too far away from the identity matrix I, and the third item is introduced to make sure that b is not far from 0. In this way, β and γ control the tradeoff between nontransfer and overtransfer. If the values are large, then A is close to I, and b is close to 0, which means nontransfer. On the contrary, small values of β and γ will result in overtransfer. β and γ are determined via cross-validations. Considering the influence of data scaling, we suggest setting them as [15]

$$\beta = \tilde{\beta}\frac{1}{d}\mathrm{Tr}\left(f_i o_i o_i^T\right), \quad \gamma = \tilde{\gamma}\sum_{i=0}^{n} f_i \tag{4}$$

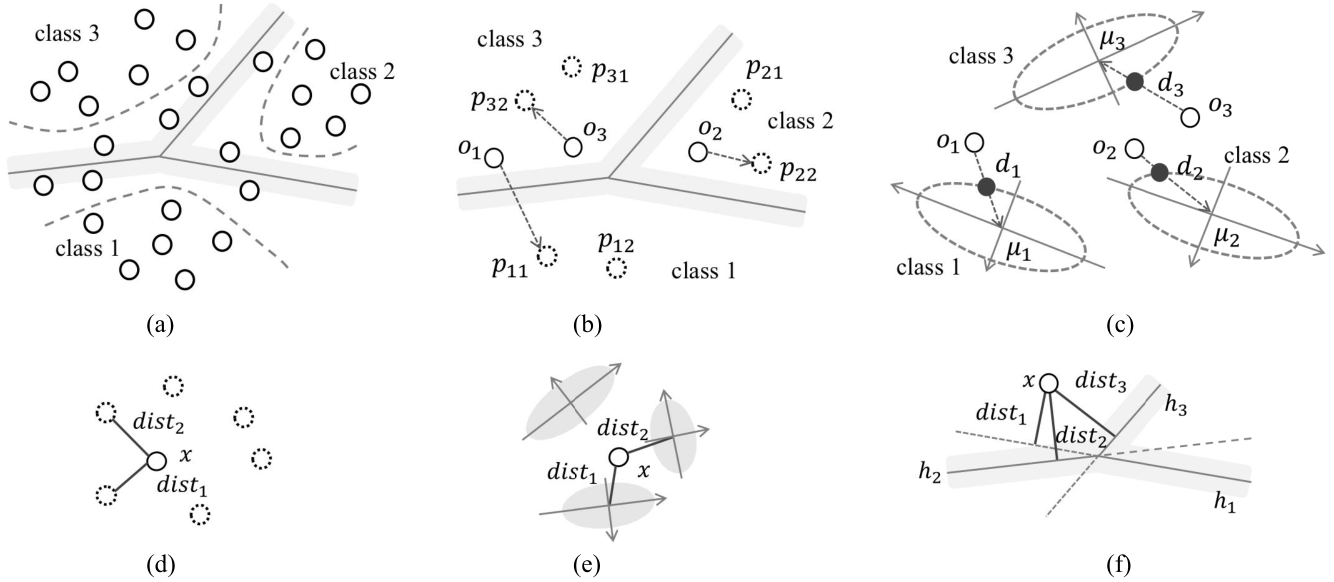


Fig. 3. Mapping destination and confidence in STM. (a) Illustration of decision boundaries. (b) Destination based on nearest prototype. (c) Destination based on the Gaussian model. (d) Confidence based on prototypes. (e) Confidence based on Gaussian centers. (f) Confidence based on multiclass SVM.

where $\mathrm{Tr}(\cdot)$ is the trace of a matrix, and $\widetilde{\beta}$ and $\widetilde{\gamma}$ could be selected efficiently between 1 and 3. Formulation (3) is a convex quadratic programming problem, which has a closed-form solution

$$A = QP^{-1},\ b = \frac{1}{\hat{f}}\left(\hat{d} - A\hat{o}\right) \tag{5}$$

where

$$Q = \sum_{i=1}^{n} f_i d_i o_i^T - \frac{1}{\hat{f}}\hat{d}\hat{o}^T + \beta I \tag{6}$$

$$P = \sum_{i=1}^{n} f_i o_i o_i^T - \frac{1}{\hat{f}}\hat{o}\hat{o}^T + \beta I \tag{7}$$

$$\hat{o} = \sum_{i=1}^{n} f_i o_i,\ \hat{d} = \sum_{i=1}^{n} f_i d_i \tag{8}$$

$$\hat{f} = \sum_{i=1}^{n} f_i + \gamma. \tag{9}$$

The optimizing process involves inverse matrix computing (P^{-1}). Since P is a symmetric matrix and in most cases a positive definite matrix, P^{-1} could be computed efficiently.

C. Mapping Origin and Destination

The mapping origin is $\boldsymbol{A}^T$. The key task is to define mapping destination in $\boldsymbol{A}^{Sp}$. Suppose there are M emotions. We train $M(M-1)/2$ binary SVMs between each pair of emotions, and use one-to-one voting strategy [32] for SVM combination. The hyper-planes of the binary SVMs are denoted as

$$h_i = \{w_i, b_i\},\ i = 1, \ldots, M(M-1)/2. \tag{10}$$

If the distance between a datum $x \in R^m$ and a hyper-plane h_i is $1/\|w_i\|$, the datum is one of the support vectors (SVs) of h_i.

In each source domain, SVs are the nearest samples from the decision boundaries, which means they are difficult to categorize. We hope the mapped origin points to be classified with high reliability. Therefore, the SVs in the source domain may have negative influence on the classification in the target domain. We remove the SVs after source classifier training, that is, the SVs are responsible to derive the decision boundaries, but do not participate in the mapping destination derivation. As Fig. 3(a) shows, in the source domain, only the non-SV data are enrolled in to derive the mapping destination (we draw three-class for illustration). This operation has potential danger of deleting some useful patterns near the boundaries, which should be examined by experiment. We explore two techniques to derive mapping destinations in the source domain ($\boldsymbol{A}^{Sp}$).

1) Nearest Prototype: There are numerous methods to derive prototypes [33], [34]. The simplest way is clustering. As Fig. 3(b) shows (two prototypes per class for illustration), we perform K-means clustering on the non-SV data in each class to obtain prototypes

$$p_{ij} \in R^m,\ j = i, \ldots,\ n_i, i = 1, \ldots, M \tag{11}$$

where n_i is the number of prototypes in class i. We define the nearest prototype of a sample $x \in R^m$ from class i in mathematical representation

$$N(x, i) = p_{ij},\ \text{where } j = \mathrm{argmin}_{j'=1}^{n_i} \left\|x - p_{ij'}\right\|_2^2. \tag{12}$$

The destination point of a sample $x \in R^m$ in $\boldsymbol{A}^T$ is defined as the nearest prototype from its genuine class (labeled data) or deduced class (unlabeled data)

$$D_{\text{proto}}(x, y) = N(x, y). \tag{13}$$

2) Gaussian Model: Gaussian models [35] assume the conditional density to be Gaussian-distributed, which is not always true. Fig. 3(c) shows the destination based on Gaussian models. The mean value of class i is μ_i, and Σ_i is the covariance matrix. We define the projection of a pattern $x \in R^m$ onto the contour surface of the Mahalanobis distance of class i as

$$P(x, i) = \mu_i + \min\left\{1, \frac{\rho}{d(x, i)}\right\} \tag{14}$$

where $d(x, i) = \sqrt{(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}$ is the Mahalanobis distance of class i. The contour is scalable under the control of ρ, which constrains the deviation of the projected point from the class mean. Large ρ may result in nontransfer, and small ρ results in overtransfer. The destination point of a sample $x \in R^m$ in $\boldsymbol{A}^T$ could be defined as the projection onto the genuine class (labeled data) or deduced class (unlabeled data)

$$D_{\text{Gauss}}(x, y) = P(x, y). \tag{15}$$

D. Confidence Setup

There are two possible strategies to compute the STM.

1) Supervised method only uses the labeled data $\boldsymbol{A}_L^T$ of the calibration sessions, which is an inductive transfer method [36].
2) Semi-supervised method uses both $\boldsymbol{A}_L^T$ and the test data $\boldsymbol{A}_U^T$ to learn STM, which is a transductive transfer method [37].

In the supervised learning of STM, there is no need to set confidence. From another perspective, we set the confidence as 1. In the semisupervised learning of STM, confidence is important. For an unlabeled datum in $\boldsymbol{A}^T$, we deduce its label, and find the mapping destination in the corresponding class. The label-deduction is not absolutely reliable. If the deduced label is wrong, STM will map the datum to a wrong class. Therefore, we tag each transformation with a confidence value, which emerges in the first item of (3). If the confidence of a sample is high, its influence on the STM computing is high, and vice-versa. Fig. 3(d)–(f) shows three confidence setups. The former two are offered in [15], and the last is a new method.

Fig. 3(d) shows the definition based on prototypes, where dist_1 is the distance between the datum and its nearest prototype (e.g., in class 1), and dist_2 is the distance between the datum and its nearest prototype in the rest of the classes (e.g., in class 2). The larger the value $\text{dist}_2 - \text{dist}_1$ is, the higher confidence we gain when deducing its label. The confidence is defined as

$$F(x) = \psi(\text{dist}_2 - \text{dist}_1) \tag{16}$$

where $\psi(\cdot) \in [0,\ 1]$ is a monotonically increasing function. We adopt sigmoidal function as follows:

$$\psi(c) = \frac{1}{1 + e^{\theta c + \tau}},\ \theta < 0. \tag{17}$$

The two parameters are determined by cross-validation. For simplicity, we designate θ as -1 and τ as 1.

Fig. 3(e) is the confidence based on the mean values of Gaussian models. The computation is (16) as well, but the prototypes are replaced by mean values.

Fig. 3(f) is the confidence defined in multiclass SVM. We measure the confidence as the weighted combination of the decision values of the binary SVMs. The distances between a datum $x \in R^m$ and each hyper-planes are

Algorithm 1 Supervised STM

```
Input:
    $A_L^T = \{x_i, y_i\}_{i=1}^n$
    $A_U^T = \{x_i, y_i\}_{i=n+1}^u$
    destination function $D(x, y)$
    hyper-parameters $\beta, \gamma$
for $i = 1$ to $n$ do
    origin $o_i = x_i$
    destination $d_i = D(x_i, y_i)$
    confidence $f_i = 1$
end for
learn STM $\{A_0, b_0\}$ with $\{o_i, d_i, f_i\}_{i=1}^n$ with (5)–(9)
transform by $\{A_0, b_0\} : A_0 x_i + b_0 \rightarrow x_i,\ \forall i = n+1, \ldots, u$
for $i = n+1$ to $u$ do
    prediction $y_i = C^{Sp}(A_0 x_i + b_0)$
end for
Output: predicted labels $l_i, i = n+1, \ldots, u$
```

$$\text{dist}_i = \frac{W_i x + b_i}{\|W_i\|_2}, i = 1, \ldots, M(M-1)/2. \tag{18}$$

In each binary SVM, large distance means high confidence in classification. We write the weighted sum of each binary classifier's confidence as

$$c = \sum_{i=1}^{M} w_i \text{dist}_i \tag{19}$$

where w_i is evaluated by "how many false classification occurs in the ith binary SVM." If the datum is classified with high confidence by an easy-to-deceive SVM, STM pays more attention to it. The final confidence configuration is

$$F(x) = \psi(c). \tag{20}$$

E. Algorithm Implementation

STM is learned between $\boldsymbol{A}^T$ and each $\boldsymbol{A}^{Sp}$. After that, each $\boldsymbol{C}^{Sp}$ makes a prediction on $\boldsymbol{A}_U^T$. We ensemble the predictions with weighted voting. The method has two versions, that is, multisource supervised STM (MS-S-STM, see Algorithm 1) and multisource semisupervised STM (MS-Semi-STM, see Algorithm 2). The difference is whether to use $\boldsymbol{A}_U^T$ in STM learning. For simplicity, we only show the pseudocode of STM, which is learned on each source, individually. The supervised STM $\{A_0, b_0\}$ is learned according to (5)–(9) with $\boldsymbol{A}^{Sp}$ and $\boldsymbol{A}_L^T$. We transform $\boldsymbol{A}_U^T$ by $\{A_0, b_0\}$ and make predictions.

In Algorithm 1, $\boldsymbol{A}_U^T$ is only used for test. They might be useful in STM learning. In some offline applications, the test data could also be used in STM learning. The semisupervised STM is thus summarized in Algorithm 2. After the supervised STM, we have $\{A_0, b_0\}$ at hand. We transform $\boldsymbol{A}^T$ with $\{A_0, b_0\}$ and enter into the self-training framework, which generates more precise estimations for the label along with the iteration. Since $\boldsymbol{A}_L^T$ has already been used when computing $\{A_0, b_0\}$, we diminish their influences by replacing the confidence as $0 \leq \alpha \leq 1$ in the iteration. In this way, the influence

Algorithm 2 Semisupervised STM

```
Input:
    $\boldsymbol{A}_L^T = \{x_i, y_i\}_{i=1}^n$
    $\boldsymbol{A}_U^T = \{x_i, y_i\}_{i=n+1}^u$
    destination function $D(x, y)$
    confidence function $F(x)$
    hyper-parameters $\alpha, \beta, \gamma, iterNum$
learn a STM $\{A_0, b_0\}$ with $\{x_i, y_i\}_{i=1}^n$ according to (5)–(9)
transform by $\{A_0, b_0\} : A_0 x_i + b_0 \rightarrow x_i, \forall i = 1, \ldots, u$
for $i = 1$ to $n$ do
    origin $o_i = x_i$,
    destination $d_i = D(x_i, y_i)$
    confidence $f_i = \alpha$
end for
do self-training: initial $A = I$, $b = 0$
for $iter = 1$ to $iterNum$ do
    for $i = n + 1$ to $u$ do
        origin $o_i = x_i$
        prediction $y_i = \boldsymbol{C}^{Sp}(Ax_i + b)$
        destination $d_i = D(x_i, y_i)$
        confidence $f_i = F(Ax_i + b)$
    end for
    learn STM $\{A, b\}$ with $\{o_i, d_i, f_i\}_{i=1}^u$ with (5)–(9)
end for
Output: predicted labels $l_i, i = n + 1, \ldots, u$
```

scope of $\boldsymbol{A}_L^T$ is mainly located outside the iteration. Finally, the semisupervised STM is expressed as $\{AA_0, Ab_0 + b\}$, which takes advantage of both $\boldsymbol{A}_L^T$ and $\boldsymbol{A}_U^T$ to learn more reliable transformation [15].

IV. Experiments and Results

A. Data

Database "SEED" [7], [9], [38] was used in our experiments, where film clips were used to evoke emotions. In the preliminary study for SEED, a pool of emotional film clips were selected from famous Chinese films. Twenty participants were recruited to assess their emotions when watching the candidate film clips using arousal scores (1–5) and valence keywords (positive, neutral, and negative). The criteria for selecting the film clips were: 1) to avoid visual fatigue, the length should not be too long; 2) the videos should be understood without explanation; and 3) the videos should elicit a single desired emotion. Finally, 15 Chinese film clips (five positive, five neutral, and five negative) were chosen from the pool of materials, which received an arousal score of 3 or higher on the mean ratings from the 20 participants [35]. Each clip lasted about 4 min.

Fifteen subjects aged 23.27±2.37 years with normal or corrected-to-normal vision participated in the EEG experiments for database SEED. A consent form was obtained from each subject before the formal experiment.

The experiment of database SEED has been described in detail previously [7], [38]. Fig. 4 is the experiment procedure. Each subject has 15 sessions. Each session includes a 5 s hint of start, the clip stimuli, a 45 s self-assessment, and a 15 s rest. To make sure that the subjects were evoked with the expected emotion, the self-assessment had three questions: 1) what do you really feel in response to the clip; 2) have you watched this movie before; and 3) do you understand the clip. Our TL scheme assumes multiple existing subjects with their individual classifiers (trained on 15 sessions), and we select N_S subjects as sources (Section III). For a new subject, we take few samples from the first three sessions (one for each emotion) as calibration data, which are used to learn STMs and determine the ensemble weights. The 12 test sessions of the new subject are transformed by STM, and the predictions are made by the classifier ensemble model associated with voting weights.

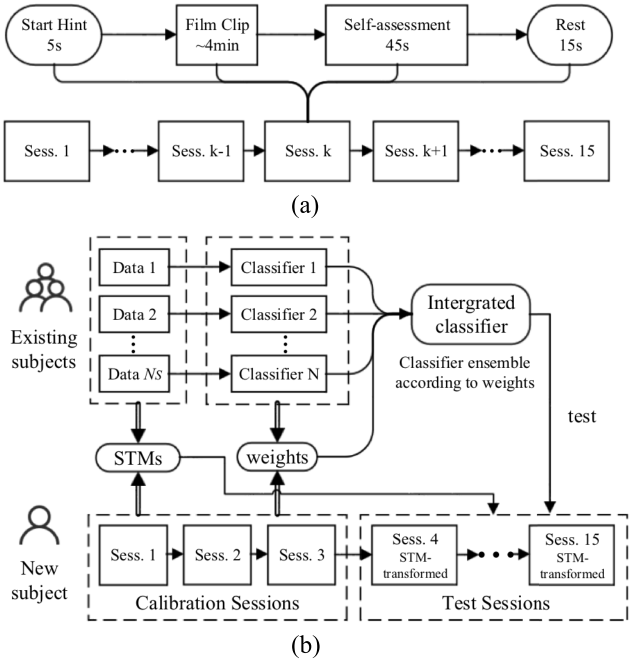


Fig. 4. Experiment procedure. (a) Experiment paradigm. Each subject has 15 sessions, and each session includes four steps. (b) TL strategy. There are multiple existing subjects (N_S selected subjects). Data means the EEG of each subject, and the individual classifiers are trained. We learn STM between the new subject and each existing subject based on the data in the new subject's three calibration sessions and the existing subject's data (15 sessions). The ensemble weights are determined by the classification accuracy of each existing classifier on the calibration data. The 12 test sessions of the new subject are transformed by STM, and each existing classifier makes a prediction. We integrate the results according to the ensemble weights to recognize the new subject's emotions in the test sessions.

During the experiment, EEG data were recorded using a Neuroscan Amplifier System from 62 active AgCl electrodes. The layout followed the international 10–20 system (see the Appendix) [39]. The sampling rate was 1000 Hz. A notch filter with 50 Hz was used during data acquisition.

B. Preprocessing

The raw EEG data were downsampled to 200 Hz sampling rate. EOG was recorded in the experiment, and later used to identify blink artifacts in the EEG data. To further filter the noise and remove the artifacts, the signals were then processed with a bandpass filter between 0.3 and 70 Hz.

After preprocessing, the EEG of each channel was divided into 1 s segments without overlapping [7]. The total number of the segment is about 3400. Features are extracted on each EEG segment. Zheng *et al.* have proved that differential entropy (DE) [40] was the most accurate and stable feature for emotion recognition than traditional features, including power spectrum density (PSD), autoregressive parameters, fractal dimension, and sample entropy [7], [40], [41]. Therefore, we use DE to characterize the EEG segments, which is defined as

$$h(x) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2} \log\frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2}$$
$$= \frac{1}{2}\log 2\pi e\sigma^2 \tag{21}$$

where x is a time series and σ is the variance of x. DE is a simple evaluation of the complexity of a time series. For each segment we compute DE on five critical frequency bands [45]: 1) Delta (1–3 Hz); 2) Theta (4–7 Hz); 3) Alpha (8–13 Hz); 4) Beta (14–30 Hz); and 5) Gamma (31–50 Hz). Since the EEG had 62 channels, the feature dimension of each segment was 62 × 5 = 310.

In view of the high correlation among samples in a same session, we select the labeled samples in a session-dependent way, which is of practical significance. We randomly take out few samples in the first three sessions to learn the STMs, and test the TL algorithm in the subsequent 12 sessions. The first three sessions include a positive session, a neutral session, and a negative session, and each one provides a same amount of data. Therefore, the calibration data contain an equal amounts of samples of the three emotion states.

C. Baseline Method

Classifier ensemble is an effective way to improve the generalization ability [42], [43]. We use weighted voting strategy to integrate the classifiers from 14 subjects to classify the emotions for the new subject in his (her) last 12 sessions. The voting weights are determined by the classification accuracies on the new subject's calibration data, as has been shown in Section III. The mean accuracy achieved across the 15 subjects is 76.2%. We use the integrated classifiers as our baseline method.

D. Transfer Learning With Multisource STM

1) Parameter Settings: The mapping origin is $\boldsymbol{A}^T$. We define the destination by either (13) or (15) in $\boldsymbol{A}^{Sp}$. For (13), we use K-means to find 15 clustering centers for each emotion. K-means is sensitive to initial values. To reduce the impact of initial values, we use K-means++ [44] and repeat the clustering operation ten times to find reliable cluster centers. We regard them as prototypes.

For Gaussian model, we appoint ρ as 50 for all subjects. This value keeps the balance between transfer and nontransfer.

The supervised STM has two parameters: the regularization parameters β and γ. We use leave-one-out cross-validation to search for the best match of $\widetilde{\beta}$ and $\widetilde{\gamma}$ between 0 and 3. The searching spaces are 0.2. The semisupervised STM has an additional parameter α. We set α as 0.8, which means the confidence of the labeled data will be replaced by 0.8 after initializing the supervised STM $\{A_0, b_0\}$. We set the iteration number as 5, that is, for each semisupervised STM learning, the self-training loop repeats five times.

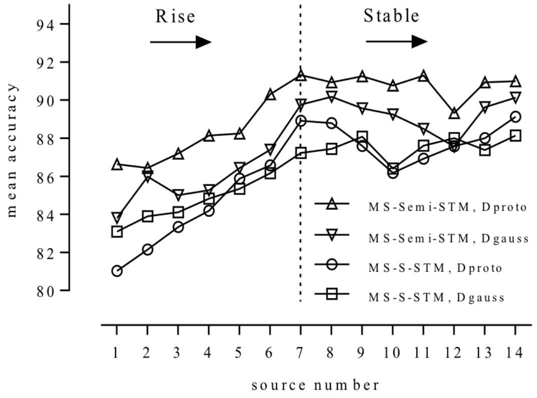


Fig. 5. Mean accuracy achieved with different source numbers. The accuracies rise significantly before seven sources ($p < 0.01$) and then remain stable. D_{proto} and D_{Gauss} specifies the way we define the mapping destination.

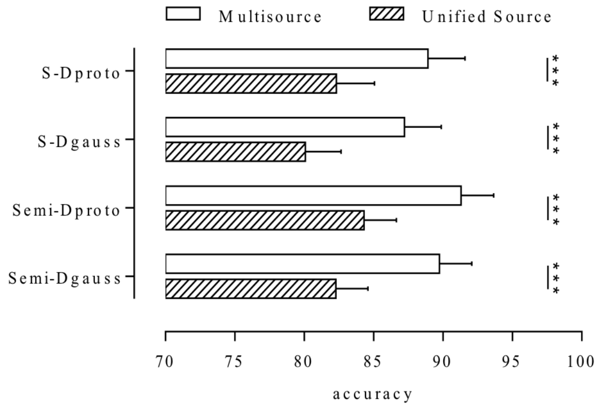


Fig. 6. Accuracies of the multisource transfer and the unified-source transfer. The error bars are standard error of mean (SEM). The multisource transfer shows better performance (***: $p < 0.0001$).

2) Source Number Evaluation: In the multisource TL, the source number N_S is an important factor. More sources means we integrate more classifiers to predict the target emotions. In view of the weak correlation between some subjects, blind increasing of the source number may not improve the accuracy, and brings computational burden [31]. To evaluate appropriate source numbers, we compare the accuracies achieved with MS-S-STM and MS-Semi-STM under different source numbers in Fig. 5. The number of calibration samples are 60 (20 samples for each emotion).

As the source number grows from 1 to 7, the accuracy rises sharply ($p < 0.01$ for the majority), which means more sources encourages better performance. Whereas after the number reaches 7, the accuracy generally stays stable (not significant in t-test), and there are even some fluctuations on the accuracy curve. To reduce the computational burden, we use seven sources in the following experiments.

3) Transfer With Unified Source: STM is associated with the multisource framework, where each subject is seen as an

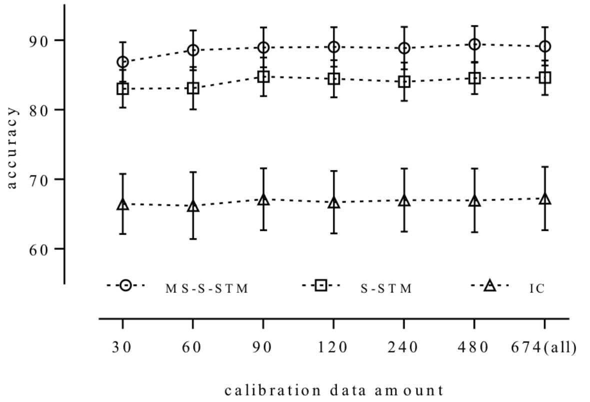


Fig. 7. Performance comparison of three methods with different calibration data amounts. The error bars are SEM. S-STM is the unified-source version of our method. S-STM and MS-S-STM are with D_{proto}. IC: individual classifier trained on the calibration data, and test on the unlabeled data.

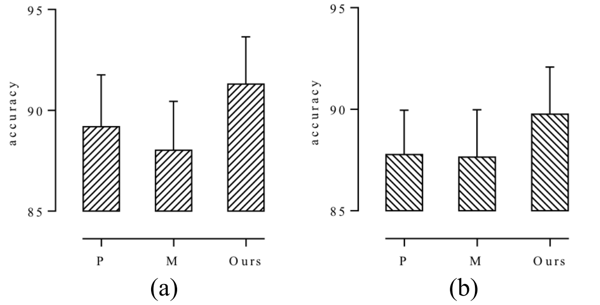


Fig. 8. Results of MS-Semi-STM across 15 subjects with different confidence settings. *P*: confidence with nearest prototype. *M*: confidence with class mean value. *P* and *M* are existing methods. Ours: confidence with multiclass SVM. (a) D_{proto} destination. (b) D_{Gauss} destination. Our confidence setting method shows advantage over *P* and *M*. Error bar: SEM.

independent source. As another choice, we could combine the data of all available subjects as a unified source [9]. We compare the accuracies achieved with multisource methods and unified-source methods in Fig. 6. In the unified-source transfer, to reduce the computational burden, we randomly pick up 5000 samples from 14 subjects as the knowledge-exporting source, and the left subject is the target. The multisource transfer shows significant advantage over the unified source transfer. Therefore, using multiple sources is more favorable than using a unified source.

4) Calibration Data Amount: The target domain has few labeled calibration data, and they play an important role in both source selection and STM computing. They are randomly picked out with balanced number for each class in the first three sessions, which means the preliminary experiments for the new subject should cover all the emotions with basically balanced samples per class in practice. Fig. 7 shows the accuracies with different numbers of calibration data. The repeated-measure analysis of variance (RM-ANOVA) results are: $F = 1246$, $p << 0.001$. The posterior analysis of MS-S-STM shows that as the calibration data grow (30–60–90), the accuracies improves significantly ($p < 0.001$), and remain stable afterward. For S-STM, the calibration data amount has little influence on the performance (not significant). In addition, MS-S-STM is superior to S-STM by introducing in more sources ($p < 0.05$ for all the considered calibration data amount). S-STM is superior to IC ($p < 0.01$).

Besides conducting transfer, the calibration data can be used to train individual classifiers directly. Posterior analysis shows that no matter how many data we use to train the individual classifiers (even when all the data in the calibration sessions are used), their performance on the 12 subsequent sessions is inferior to S-STM ($p < 0.01$ for all the considered calibration data amount) and MS-S-STM ($p < 0.001$ for all the considered calibration data amount). The transfer methods show advantage over the individual classifiers when the calibration samples are few. This is the foothold for us to do TL.

5) Confidence Evaluation: Confidence is important in MS-Semi-STM. Fig. 8 compares our confidence-setting method (based on multiclass SVM) with the nearest prototype (*P*) and class mean value (*M*). Our method shows superiority to *P* and *M* on the majority of the subjects. The RM-ANOVA results under D_{proto} are: $F = 1375$ and $p << 0.001$. Posterior analysis shows the superiority of our method to *P* ($p < 0.01$) and *M* ($p < 0.01$). The RM-ANOVA results under D_{Gauss} are: $F = 775$ and $p << 0.001$. Posterior analysis shows the superiority of our method to *P* ($p < 0.01$) and *M* ($p < 0.01$).

6) Individual Performance: To show the results more intuitively, Table II summarizes the results on the 15 subjects. The source number is 7. The calibration data amount is 60. For most subjects, the optimal $\widetilde{\beta}$ ranges from 0.2 to 0.8, and $\widetilde{\gamma}$ is from 0 to 3. The baseline is the classifier ensemble with a mean accuracy of 76.2%.

In MS-S-STM, STMs are learned with the calibration data alone. "No SV" means the SVs are removed. "SV" means they are preserved. Under all the conditions, whether to remove the SVs has little influence on the result (no statistical significance). Since the mean accuracies after SV removal is higher, we suggest removing the SVs. Under D_{proto}, the mean accuracy rises by 12.72% comparing with the baseline. Under D_{Gauss}, the accuracy rises by 11.02%. The mapping destinations defined by D_{proto} and D_{Gauss} have basically performance (no statistical significance). The mean accuracies for D_{proto} is higher. MS-Semi-STM improves the performance of MS-S-STM by absorbing the additional information of the test data. The sharpest rise is 15.11%. We compute the *p*-values of the paired-sample *t*-test between the baseline and our methods. All the *p*-values are below 0.01, indicating the advantage of our method.

Both MS-S-STM and MS-Semi-STM bring impressive improvement to the accuracy. However, the improvement scales on individuals differ greatly. For example, the improvement is uncertain on subject 5, whereas on subject 11, it is more than 20%. The effect of STM varies from person to person.

Our method alleviates the preliminary experiments of new subjects, and the evidences are shown in Fig. 9. We denote $p{-}{>}\ q$ as training model on the first *p* sessions, and test it on the following *q* sessions. For a same subject, more training sessions does not always mean better classification results. For

TABLE II
PERFORMANCE OF MULTISOURCE TL

Subject index	Classifier ensemble	MS-S-STM				MS-Semi-STM			
		No SV		SV		No SV		SV	
		D_{proto}	D_{Gauss}	D_{proto}	D_{Gauss}	D_{proto}	D_{Gauss}	D_{proto}	D_{Gauss}
1	78.68	95.27	84.98	89.06	85.97	97.63	87.93	95.57	88.82
2	86.69	93.80	93.88	91.71	93.91	94.47	95.43	95.10	94.65
3	55.44	65.79	71.74	69.76	71.89	68.80	83.95	70.97	85.42
4	87.87	100.00	98.14	99.06	92.22	100.00	93.43	100.00	93.43
5	80.88	81.60	71.01	79.73	71.67	93.07	66.38	95.46	71.39
6	77.83	95.57	98.65	95.86	98.91	95.57	100.00	95.42	98.47
7	80.96	99.94	98.84	97.48	98.99	98.71	98.76	98.14	99.94
8	77.61	91.71	78.36	85.61	79.54	98.81	83.99	86.63	83.25
9	71.99	88.62	88.99	85.16	87.55	89.72	90.86	88.18	90.27
10	65.63	76.04	75.46	74.57	76.63	79.50	82.26	78.21	82.33
11	49.67	73.40	80.24	71.60	83.32	81.85	86.41	76.49	88.07
12	74.52	91.67	94.68	90.86	88.62	94.57	99.61	94.68	91.49
13	82.39	86.87	80.82	88.10	80.35	83.25	86.04	88.99	81.68
14	92.76	99.12	99.94	99.65	99.35	100.00	100.00	100.00	100.00
15	80.22	94.33	92.57	91.89	90.88	93.66	91.28	93.25	91.88
Mean	76.20	**88.92**	87.22	87.34	86.65	**91.31**	89.76	90.47	89.41
Improve	-	12.72	11.02	11.14	10.45	15.11	13.56	14.27	13.21
SD	11.64	10.35	10.32	9.65	9.31	9.07	9.00	8.91	7.79
Significance	-	***	**	***	*	***	**	***	**

The accuracies are shown in percentage (%). Mean: mean value. SD: standard deviation (*: P<0.01; **: P<0.001; ***: P<0.0001).

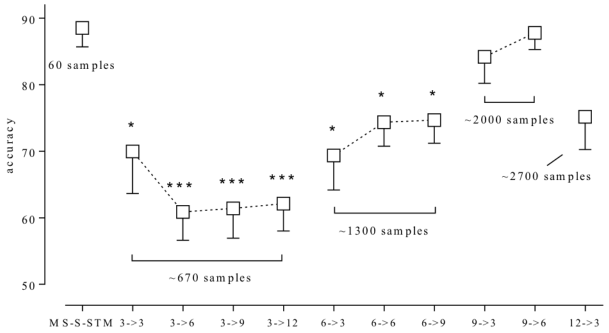


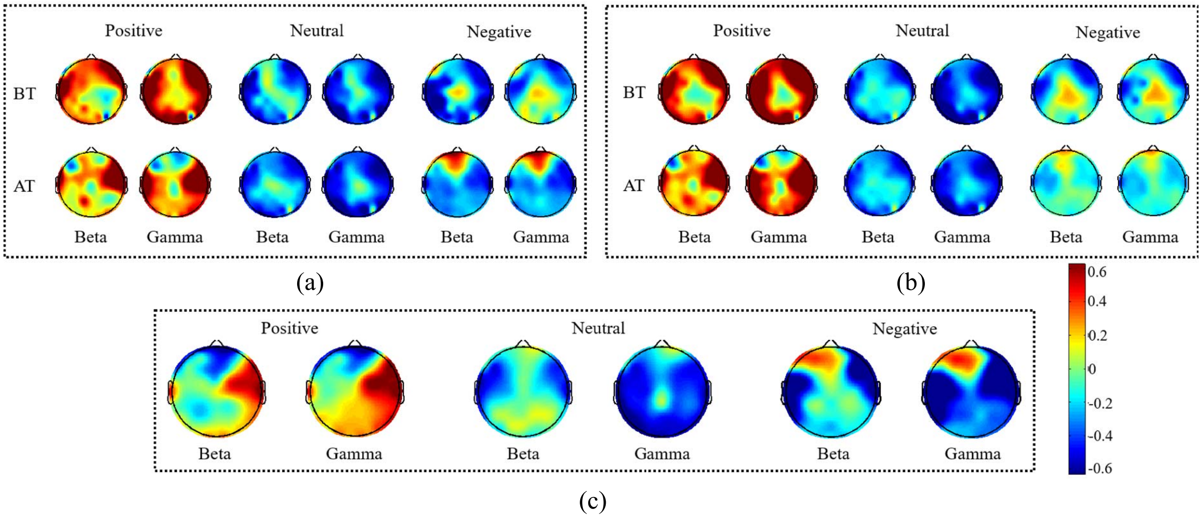
Fig. 9. Method uses fewer training data. *p*-> *q*: train models on the first *p* sessions, and test on the following *q* sessions. *: $p < 0.01$; **: $p < 0.001$; and ***: $p < 0.0001$. The baseline is MS-S-STM. We mark the sample amounts on the graph. Error bar: SEM.

example, 12->3 does not perform well. This may be caused by the nonstationary characteristics of EEG, where the correlation between far-apart sessions is weak. The performance of MS-S-STM is basically the same as that of 9->6 and 9->9. To reach an equivalent accuracy, our method uses much fewer training sessions (3 versus 9) and much fewer training samples (60 versus ~2000). In addition, MS-S-STM uses information in the first three sessions to predict the subsequent 12 sessions, which is similar to 3->12. We compare them, and find that MS-S-STM shows reliable classification on the 12 sessions. MS-S-STM shows tolerance to the nonstationary EEG signal.

V. DISCUSSION

Noticing the individual differences across subjects, Zheng and Lu [9] explored some TL methods to confront the distribution change. However, they used all the unlabeled data in the target domain, which is not realistic in practice. In this paper, we use as few as possible samples (e.g., 60) from the first few sessions to quickly adapt emotion classifiers to new persons with TL approach. In our experiment setting, MS-S-STM shows an accuracy of 88.92% for three-category emotion recognition, which is of practical significance in fast-deployment applications. MS-Semi-STM achieves 91.31% by using the test data in a transductive way.

Fig. 10 is an illustration of the effect of STM. DE features on five frequency bands are used in this paper. Among these five bands, Beta and Gamma are the most emotion-related [7]. Here, we draw the scalp maps on these two bands for illustration. We take the first subject as target, and the last as source. The first row shows the average DE values of the target for both the 60 labeled samples in the three calibration sessions



(a)

(b)

(c)

Fig. 10. STM reduces the domain discrepancy. (a) Labeled target. (b) Unlabeled target. (c) Source. BT: before transformation. The huge activation differences make the source classifier inapplicable to the target. AT: after transformation (with STM). For the positive emotion, STM makes the activations of the target closer to temporal electrodes (like the source). For the neutral emotion, there are no visible differences. For the negative emotion, STM moves the activation patterns to the frontal electrodes (like the source). In general, STM makes the target more statistically similar with the source while retaining its own characteristics. The color bars correspond to DE values (best viewed in electronic version).

and the 3334 unlabeled samples in the 12 test sessions. The EEG patterns of the labeled and unlabeled data are similar, indicating the consistency of EEG of one person in a same experiment. We use the labeled samples to compute supervised STM, and the STM-transformed target is shown in the second row. The third row shows the average DE values of the source. Existing researches have shown that there are two main EEG areas correlated with emotion activity: 1) the frontal portion of the temporal lobe and 2) the prefrontal part of the brain [6]. For the positive emotion of the target, the DE values (both Beta and Gamma) in the temporal and the prefrontal regions are consistently higher than the neutral and negative emotion. As for the source, the values in the temporal region (especially the right temporal) are large. The source and the target patterns are in consistent with [45]–[47]. In the second row, STM moves the DE patterns of the target to the right temporal lobe, making the DE patterns similar to the source while preserving their own characteristics in the left temporal and frontal regions. For the negative emotion of the target, the main Beta and Gamma activations emerge in the parietal region [6], however, the main activations of the source are in the left frontal. In the second row, STM moves the target activations to the left frontal, making the DE topology closer to the source. For the neutral emotion, we find no visual difference. The effect of STM is also reflected in the classification accuracy. For the target, the accuracy of the classifier trained on the 60 samples is 51.80%; the accuracy of the source classifier is 57.62%. They both show poor generalization on the 12 test sessions. After transfer, the source classifier achieves 72.38%, which is close to the 9->6 scheme (see Section IV) on the target (74.57%). Therefore, STM encourages the similarity between the two distributions so as to improve the classification accuracy.

Selecting appropriate sources is important, for the individual differences between subjects are huge. Blind increasing of the source number would not help improve the performance. In addition, we suggest taking each subject as an individual source, rather than combining multiple sources as a unified source.

To adapt STM to SVM, we investigate the role of the close-to-boundary samples (SVs). SVs might mislead the mapping to the direction where classification is more difficult to perform. On the other hand, the removal of SVs might delete useful patterns. We conduct experiments to evaluate the influence of SVs. The results show that the SV removal enhances the mean accuracies. In practice, we suggest removing the SVs.

We validate the effectiveness of the confidence setup based on multiclass SVM in MS-Semi-STM. As the difficulty of distinguishing each pair of emotions is not the same, we tag weights to the binary classifiers. The results prove that the proposed confidence setting is effective.

In this paper, the classifiers are trained on the source, and then used in the target. Considering the labeled samples in the target could introduce extra supervision for the classifiers, each classifier could be retrained based on the combination of the source and the target's calibration data [48]. We preserve the experiment settings and repeat the experiments, and find this strategy improves the mean accuracies by around 2% (for both MS-S-STM and MS-Semi-STM). The cost for a 2% increase in accuracy is to retrain all the classifiers, which is a tradeoff between accuracy and computational complexity.

The results are sensitive to β, but not to γ. The regularization on b is less important than the regularization on A.

MS-S-STM and MS-Semi-STM show superiority, where a few labeled data are available. However, the unsupervised STM does not perform well. We think the reason is as follows: to compute STM parameters, each target sample should be assigned with a mapping destination. If the target samples are linked with destinations in the wrong classes, the

TABLE III
ACCURACIES ACHIEVED WITH DIFFERENT METHODS

Method	Mean	SD
CE	76.20	11.64
MS-KPCA	80.97	15.10
MS-TSVM	82.51	13.08
MS-TCA	82.14	15.94
MS-TPT	83.67	11.39
MS-S-STM, Dproto	87.22	11.02
MS-S-STM, Dgauss	**88.92**	12.72

CE: classifier ensemble.

STM learning will collapse. The limitation of our method is twofold. First, we focus on the machine learning methods and their applications in aBCI, and the explanations from the perspective of neurophysiology is few. Second, the number of categories in this paper is few (three emotions). For further work, we will evaluate the method with more emotion categories on more datasets, and exploit zero-training strategies for cross-subject transfer.

VI. Conclusion

We have proposed a multisource TL method to generalize existing emotion recognition models to new persons. The method reduces the demand for the calibration data amount effectively, and integrates models for new persons with few calibration samples. It facilitates the fast acquisition of emotion recognition models for new users, which is of great significance especially in fast-deployment scenarios. The results show the superiority of our method.

Appendix

To evaluate whether existing methods work well with few calibration data (as in our case), we summarize the performance of different methods in Table III. We repeat the experiments in [7] with our multisource framework using weight voting strategy. The source number is 7, and the calibration data amount is 60. All the experimental conditions are the same. MS-S-STM with D_{proto} has the highest accuracy. The reasons are as follows. STM is a linear (affine) mapping. Compared with the nonlinear methods, linear methods are expected to reduce overfitting when the training samples are few. In addition, STM uses the emotion labels (semantic information) of the target when learning the affine transformation, while in the others, the semantic information is largely ignored.

The EEG placement followed the international 10–20 system (see Fig. 11).

We show an example of the transfer matrix A in Fig. 12(a) and the bias b in Fig. 12(b). A is a diagonally dominant matrix, where the diagonal elements are close to 1, and the left elements are close to 0. The b values wander around 0. They are learned with (3). Under the effect of the regularization, the algorithm does not lead the model to overtransfer, and the basic structure of the target EEG is preserved with the affine mapping.

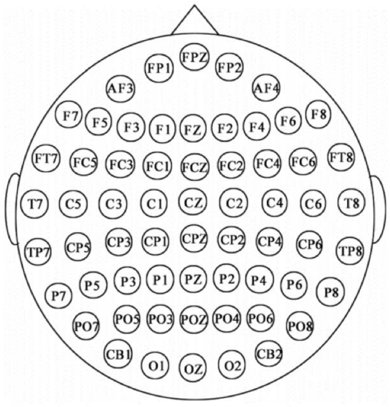
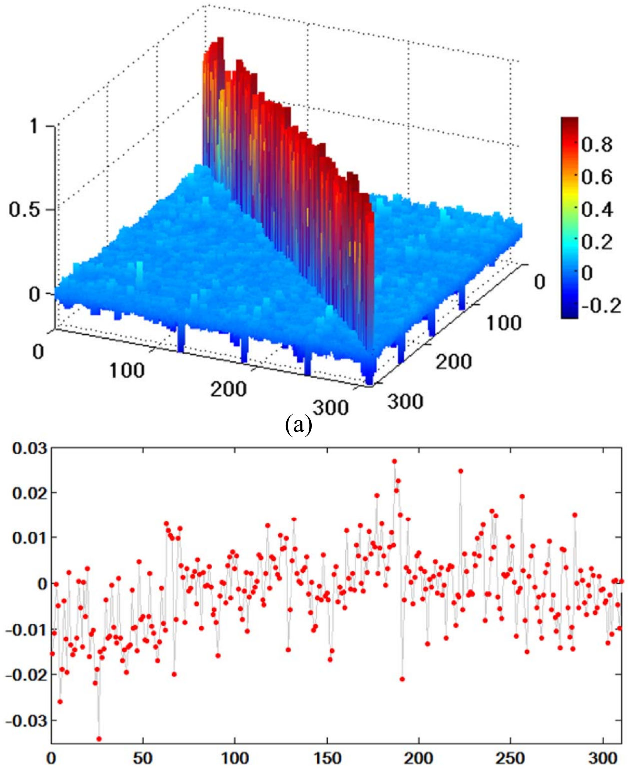


Fig. 11. Placement of the 62-channel EEG.



(a)

(b)

Fig. 12. Transfer matrix A and the bias b. A is a diagonally dominant matrix and b is close to zero vector. The transformation preserves the basic structure of the target EEG data.

Acknowledgment

The authors would like to thank Prof. B. Lu for providing the SEED dataset. They would also like to thank the anonymous reviewers for the excellent comments, which have helped tremendously to improve the quality of this paper.

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: Massachusetts Inst. Technol., 1995.

[2] R. Cowie *et al.*, “Emotion recognition in human–computer interaction,” *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[3] G. L. Ahern and G. E. Schwartz, “Differential lateralization for positive and negative emotion in the human brain: EEG spectral analysis,” *Neuropsychologia*, vol. 23, no. 6, pp. 745–755, 1985.

[4] X. Huang *et al.*, “Multi-modal emotion analysis from facial expressions and electroencephalogram,” *Comput. Vis. Image Understanding*, vol. 147, pp. 114–124, Jul. 2016.

[5] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, “A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges,” *Brain–Comput. Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[6] S. M. Alarcao and M. J. Fonseca, “Emotions recognition using EEG signals: A survey,” *IEEE Trans. Affective Comput.*, to be published.

[7] W.-L. Zheng and B.-L. Lu, “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[8] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, “Transfer learning in brain–computer interfaces,” *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 20–31, Feb. 2016.

[9] W.-L. Zheng and B.-L. Lu, “Personalizing EEG-based affective models with transfer learning,” in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, New York, NY, USA, 2016, pp. 2732–2738.

[10] J. Li, Z. Zhang, and H. He, “Hierarchical convolutional neural networks for EEG-based emotion recognition,” *Cogn. Comput.*, vol. 10, no. 2, pp. 1–13, 2017.

[11] Y.-P. Lin *et al.*, “EEG-based emotion recognition in music listening,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[12] F. Lotte *et al.*, “A review of classification algorithms for EEG-based brain–computer interfaces,” *J. Neural Eng.*, vol. 15, no. 3, 2007, Art. no. 031005.

[13] S. J. Pan and Q. A. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[14] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, “Visual domain adaptation: A survey of recent advances,” *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[15] X.-Y. Zhang and C.-L. Liu, “Writer adaptation with style transfer mapping,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1773–1787, Jul. 2013.

[16] X.-Y. Zhang and C.-L. Liu, “Style transfer matrix learning for writer adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 393–400.

[17] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, “Optimal spatial filtering of single trial EEG during imagined hand movement,” *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[18] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, “Towards zero training for brain–computer interfacing,” *PLoS ONE*, vol. 3, no. 8, 2008, Art. no. e2967.

[19] S. Fazli *et al.*, “Subject-independent mental state classification in single trials,” *Neural Netw.*, vol. 22, no. 9, pp. 1305–1312, 2009.

[20] H. Morioka *et al.*, “Learning a common dictionary for subject-transfer decoding with resting calibration,” *NeuroImage*, vol. 111, pp. 167–178, May 2015.

[21] P. Von Bünau, F. C. Meinecke, F. C. Király, and K. R. Müller, “Finding stationary subspaces in multivariate time series,” *Phys. Rev. Lett.*, vol. 103, no. 21, 2009, Art. no. 214101.

[22] W. Samek, C. Vidaurre, K. R. Müller, and M. Kawanabe, “Stationary common spatial patterns for brain–computer interfacing,” *J. Neural Eng.*, vol. 9, no. 2, 2012, Art. no. 026013.

[23] W. S. Chu, F. De la Torre, and J. F. Cohn, “Selective transfer machine for personalized facial action unit detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 3515–3522.

[24] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, “Correcting sample selection bias by unlabeled data,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, pp. 601–608.

[25] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[26] A. Argyriou, T. Evgeniou, and M. Pontil, “Multitask feature learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, p. 41.

[27] A. Schwaighofer, V. Tresp, and K. Yu, “Learning Gaussian process kernels via hierarchical Bayes,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1209–1216.

[28] J. Gao, W. Fan, J. Jiang, and J. Han, “Knowledge transfer via multiple model local structure mapping,” in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2008, pp. 283–291.

[29] R. Collobert, F. Sinz, J. Weston, and L. Bottou, “Large scale transductive SVMs,” *J. Mach. Learn. Res.*, vol. 7, pp. 1687–1712, Aug. 2006.

[30] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, “We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer,” in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 357–366.

[31] Y. Yao and G. Doretto, “Boosting for transfer learning with multiple sources,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 1855–1862.

[32] J. Weston and C. Watkins, “Multiclass support vector machines,” Dept. Comput. Sci., Roy. Holloway, Univ. London, London, U.K., Rep. CSD-TR-98-04, May, 1998.

[33] J. C. Bezdek and L. I. Kuncheva, “Nearest prototype classifier designs: An experimental study,” *Int. J. Intell. Syst.*, vol. 16, no. 12, pp. 1445–1473, 2001.

[34] L. I. Kuncheva and J. C. Bezdek, “Nearest prototype classification: Clustering, genetic algorithms, or random search?” *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp. 160–164, Feb. 1998.

[35] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, “Modified quadratic discriminant functions and the application to Chinese character recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 1, pp. 149–153, Jan. 1987.

[36] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Glob., 2010, pp. 242–264.

[37] M. Rohrbach, S. Ebert, and B. Schiele, “Transfer learning in a transductive setting,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 46–54.

[38] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, “Identifying stable patterns over time for emotion recognition from EEG,” *IEEE Trans. Affective Comput.*, to be published.

[39] (2012). *TransCranialTechnologies, 10/20 System Positioning—Manual*. Accessed: Feb. 2016. [Online]. Available: http://www.transcranial.com/local/manuals/10 20 pos man v1 0 pdf.pdf

[40] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, “Differential entropy feature for EEG-based emotion classification,” in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, 2013, pp. 81–84.

[41] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, “Differential entropy feature for EEG-based vigilance estimation,” in *Proc. IEEE 35th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 6627–6630.

[42] X. Ceamanos *et al.*, “A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data,” *Int. J. Image Data Fusion*, vol. 1, no. 4, pp. 293–307, 2010.

[43] X. Zhu, P. Zhang, X. Lin, and Y. Shi, “Active learning from stream data using optimal weight classifier ensemble,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 6, pp. 1607–1621, Dec. 2010.

[44] D. Arthur and S. Vassilvitskii, “k-means++: The advantages of careful seeding,” in *Proc. 18th Annu. ACM-SIAM Symp. Discr. Algorithms*, 2007, pp. 1027–1035.

[45] N. Jatupaiboon, S. Pan-ngum, and P. Israsena, “Emotion classification using minimal EEG channels and frequency bands,” in *Proc. 10th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE)*, 2013, pp. 21–24.

[46] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, “Asymmetric spatial pattern for EEG-based emotion detection,” in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2012, pp. 1–7.

[47] P. Sarkheil, R. Goebel, F. Schneider, and K. Mathiak, “Emotion unfolded by motion: A role for parietal lobe in decoding dynamic facial expressions,” *Soc. Cogn. Affective Neurosci.*, vol. 8, no. 8, pp. 950–957, 2012.

[48] L. Duan, D. Xu, and I. W.-H. Tsang, “Domain adaptation from multiple sources: A domain-dependent regularization approach,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.

Jinpeng Li received the B.E. and M.E. degrees in automatic control from the University of Science and Technology Beijing, Beijing, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing.

He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing. His current research interests include pattern recognition, machine learning, deep learning, transfer learning algorithms, and their applications in brain–computer interfaces and medical image analysis.

Shuang Qiu received the B.S. and Ph.D. degrees in biomedical engineering from Tianjin University, Tianjin, China, in 2010 and 2017, respectively.

From 2014 to 2016, she was a visiting student with the Department of Physical Medicine and Rehabilitation, Harvard Medical School, Boston, MA, USA. She is currently an Assistant Professor with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing. Her current research interests include biosignal processing, machine learning, and their applications in rehabilitation technology, with emphasis on human–machine interface and neurofeedback treatment.

Yuan-Yuan Shen received the B.S. degree from Anhui University, Hefei, China, in 2010 and the M.E. degree in computer science and technology from Xiamen University, Xiamen, China, in 2015. She is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

She is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing. Her current research interests include machine learning, pattern recognition, and data stream classification.

Cheng-Lin Liu (F'15) received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, in 1989, the M.E. degree in electronic engineering from Beijing Polytechnic University, Beijing, China, in 1992, and the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences (CAS), Beijing, in 1995.

He was a Post-Doctoral Fellow with the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, and later with the Tokyo University of Agriculture and Technology, Tokyo, Japan, from 1996 to 1999. From 1999 to 2004, he was a Research Staff Member and later a Senior Researcher with the Central Research Laboratory, Hitachi, Ltd., Tokyo. Since 2005, he has been a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, CAS, where he is currently the Director of the Laboratory. He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, and the Center for Excellence in Brain Science and Intelligence Technology, CAS. He has published over 200 technical papers at prestigious international journals and conferences. His current research interests include pattern recognition, image processing, neural networks, machine learning, and especially the applications to character recognition and document analysis.

Huiguang He (M'04–SM'10) received the B.S. and M.S. degrees from Dalian Maritime University (DMU), Dalian, China, in 1994 and 1997, respectively, and the Ph.D. degree (Hons.) in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China.

From 1997 to 1999, he was an Associate Lecturer with DMU. From 2003 to 2004, he was a Post-Doctoral Researcher with the University of Rochester, Rochester, NY, USA. From 2014 to 2015, he was a Visiting Professor with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He is currently a Full Professor with the Research Center for Brain-Inspired Intelligence, National Laboratory of Pattern Recognition, CASIA. He is also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, and the Center for Excellence in Brain Science and Intelligence Technology, CAS. His research has been supported by several research grants from the National Science Foundation of China. He has authored or co-authored over 120 peer-reviewed papers. His current research interests include pattern recognition, medical image processing, and brain–computer interfaces.

Dr. He was a recipient of the Excellent Ph.D. dissertation of CAS in 2004, the National Science and Technology Award in 2003 and 2004, the Beijing Science and Technology Award in 2002 and 2003, the K.C. Wong Education Prizes in 2007 and 2009, and the Jia-Xi Lu Young Talent Prize in 2009. He is an Excellent Member of Youth Innovation Promotion Association, CAS in 2016.