Multitask Feature Learning Meets Robust Tensor Decomposition for EEG Classification

Qingqing Zheng, *Member, IEEE*, Yi Wang, and Pheng Ann Heng, *Senior Member, IEEE*

***Abstract*—In this article, we study a tensor-based multitask learning (MTL) method for classification. Taking into account the fact that in many real-world applications, the given training samples are limited and can be inherently arranged into multidimensional arrays (tensors), we are motivated by the advantages of MTL, where the shared structural information among related tasks can be leveraged to produce better generalization performance. We propose a regularized tensor-based MTL method for joint feature selection and classification. For feature selection, we employ the Fisher discriminant criterion to both select discriminative features and control the within-class nonstationarity. For classification, we take both shared and task-specific structural information into consideration. We decompose the regression tensor for each task into a linear combination of a shared tensor and a task-specific tensor and propose a composite tensor norm. Specifically, we use the scaled latent trace norm for regularizing the shared tensor and the ℓ_1-norm for task-specific tensor. Further, we give a computationally efficient optimization algorithm based on the alternating direction method of multipliers (ADMMs) to tackle the joint learning of discriminative features and multitask classification. The experimental results on real electroencephalography (EEG) datasets demonstrate the superiority of our method over the state-of-the-art techniques.**

***Index Terms*—Electroencephalograph (EEG), Fisher discriminant criterion, multitask learning (MTL), tensor classification.**

Manuscript received April 27, 2019; revised August 2, 2019; accepted September 29, 2019. This work was supported in part by the Grant from 973 Program under Project 2015CB351706, in part by the National Natural Science Foundation of China under Grant 61701312, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515010847, in part by the Medical Science and Technology Foundation of Guangdong Province under Grant B2019046, and in part by the Natural Science Foundation of Shenzhen University under Grant 2018010. This article was recommended by Associate Editor D. Tao. *(Corresponding author: Yi Wang.)*

Q. Zheng is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, and also with Tencent Youtu Lab, Shenzhen 518057, China.

Y. Wang is with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: onewang@szu.edu.cn).

P. A. Heng is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, and also with the Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.

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

Digital Object Identifier 10.1109/TCYB.2019.2946914

I. Introduction

CLASSIFICATION techniques usually require great amounts of training samples to learn an accurate classifier. For example, the conventional supervised deep networks often need millions of labeled samples to train a classification model containing a large number of parameters [1]. However, such a requirement cannot be met in some applications, such as biological signal analysis since labeled samples in this area are hard to collect [2]. With limited labeled data, it is very challenging to learn an accurate classifier, even for shallow models. As a remedy, multitask learning (MTL) jointly learns multiple related tasks so that sample size can be effectively increased and knowledge obtained from each task can be shared in the learning process [3]–[7]. In this way, MTL is an effective learning paradigm to improve the generalization performance of multiple related tasks, with each of them having limited training samples [8].

Traditional MTL methods are designed for data with vector representation and do not explore the inherent structures embedded in the data, whereas many real-world applications produce data in the form of matrices or tensors, such as video sequences [9] and electroencephalography (EEG) signals [10]. When learning from such data of multiple indices, we have to naively vectorize them before applying traditional MTL methods [11]. In such a way, it would result in performance degradation due to the loss of structural information [12]. Moreover, such vectorization may suffer from the curse of dimensionality. For example, the popular EEG dataset IIa [13] contains 288 samples for each subject, and per-sample records the EEG signals with 750 time points at 22 electrodes. By preprocessing each sample with z (say $z = 6$) band-pass filters, each sample can be naturally represented as a tensor of size $750 \times 22 \times 6$. Vectorizing these tensors would produce samples of dimensionality as large as 99 000, which is much larger than the sample size 288.

With the development of high-order data analysis, some modern MTL methods are investigated to handle tensor-based features [14]. Romera-Paredes *et al.* [3] first proposed the overlapped trace norm to model shared tensor structures between related tasks. Tomioka and Suzuki [15] proposed an alternative method, also known as latent trace norm, to decompose a tensor into a mixture of latent tensors, where each is low rank in a specific order. Inspired by [15], Wimalawarne *et al.* [16] further studied the scaled latent trace norm and obtained better generalization when tensors have heterogeneous ranks. However, most of these methods focus only on tensor decomposition or tensor recovery. Directly extending these methods

for supervised tensor classification would fail to select discriminative features or control the nonstationarity[1] existing in the training samples. Again taking EEG data, for instance, a large variance may exist in the same category EEG signals, due to mental fatigue or distraction of subjects during the signal collection procedure [17]. Existing tensor-based MTL methods without feature selection [15], [16] would suffer performance degeneration when applied to EEG classification. Therefore, it is necessary for an efficient tensor-based MTL method to automatically detect discriminative and useful features for tensor classification.

To tackle these issues, we propose a regularized tensor-based MTL method for joint feature selection and classification. For feature selection, different from the existing methods [18], which learn discriminative features and classifiers in two separated steps by optimizing different objective functions, we select useful features and learn the classification models simultaneously in a unified framework. Specifically, we employ the Fisher discriminant criterion to minimize the within-class variance and meanwhile maximize the between-class distance. In such a way, the proposed method can enlarge the boundary between different classes and control the within-class variance. For classification, we take both the shared and task-specific structural information into consideration. We decompose the regression tensor for each task into a linear combination of a shared tensor and a task-specific tensor. To extract the shared patterns, we employ the scaled latent trace norm to regularize a mixture of latent tensors so that each is low rank for a specific mode over multiple related tasks. Meanwhile, the ℓ_1-norm is leveraged to constrain the sparsity of the task-specific tensor to detect individual patterns or outliers. To solve the resulting optimization problem, we newly develop an efficient algorithm based on the alternating direction method of multipliers (ADMMs) framework [19]. We investigate the performance of our method and the state-of-the-art techniques on three real EEG datasets. The experimental results demonstrate that the proposed method yields competitive performance in all datasets.

The novelties of this article are attributed to three aspects. First, to the best of our knowledge, our method is the first multitask tensor classification in the context of EEG classification. Our method learns multiple related tasks simultaneously and leverages the shared knowledge to improve the classification, especially, when the training data are limited for each task but related among multiple tasks. Second, compared with the existing models, our method addresses the nonstationarity issue and selects discriminative features by adding the Fisher discriminant measure to the cost function. In such a way, our method can jointly optimize feature selection and classification. Third, the sparse and low-rank decomposition are less explored in the tensor data. Our method considers both the shared low-rank structural information among multiple tasks and the sparse task-specific information for each task.

The remainder of this article is organized as follows. Section II introduces the related works on MTL and tensor-based classification. Section III gives the notations and preliminaries that run throughout this article. Section IV presents the details of our regularized multitask feature learning for tensor data classification. Section V shows the experimental setting and results. The discussion and the conclusion of this article are given in Sections VI and VII, respectively.

[1]The situation where EEG signals may change rapidly over time or over sessions.

II. Related Works

A. MTL

Most existing MTL algorithms assume that the input data and models are both N-dimensional vectors. Then, they stack T tasks into an $N \times T$ sized matrix $\mathbf{W}$. Despite different motivations and applications, many regularization-based MTL methods are proposed by adding constraints on $\mathbf{W}$. For example, Liu *et al.* [20] studied an $\ell_{2,1}$-norm on $\mathbf{W}$ to jointly select features from multiple tasks for regression. Ji and Ye [21] proposed a trace norm on $\mathbf{W}$ to learn the low-rank structure among multiple tasks. Similarly, an earlier work [22] first decomposed the linear model for each task t as a summation of a common vector $\mathbf{w}_0$ and a task-specific vector $\mathbf{v}^t$, and employed an ℓ_2-norm to regularize both $\mathbf{w}_0$ and $\mathbf{v}^t$. Based on [22], Li *et al.* [23] designed a composite norm to simultaneously learn shared regression parameters and shared features. Kumar and Daume, III [24] proposed a scheme for learning grouping and overlap structure in MTL, where parameters of each task group are assumed to lie in a low-dimensional subspace. Luo *et al.* [25] developed a heterogeneous MTL approach, which transfers the knowledge between different domains by finding a common subspace. In the induced subspace, the high-order divergences between all domains are exploited to learn more reliable metric. However, all of these MTL methods are built for vector-form data, which would result in the loss of structural information when the data are inherently modeled in tensor form.

B. Tensor-Based Classification

The objective is to investigate the classifiers that directly process tensor data without vectorization. Though for matrices, low-rank structure has been successfully applied to many applications, such as robust principal component analysis [26], missing data completion [27], and matrix-based classification [28], it is inherently more complex to find a proper low-rankness for tensors. Recently, several tensor norms, such as tensor trace norm [29], overlapped trace norm [3], latent trace norm [15], and scaled latent trace norm [16], have investigated high-order structures for tensor completion or decomposition. Based on [16], Wimalawarne *et al.* [30] further applied the scaled latent trace norm to single-task tensor classification. However, tensor structures in the supervised MTL framework have not yet been explored. Tao *et al.* [31] proposed a general tensor discriminant analysis (GTDA) as a tensor representation model for gait recognition. This article employed GTDA to extract features from tensors and further used conventional linear discriminant analysis (LDA) for recognition. In contrast, we propose a regularized tensor-based method for joint feature selection and classification. We learn multiple related tasks simultaneously and leverage the

shared knowledge to improve the tensor-form data classification. In addition, we employ the Fisher discriminant criterion to address the nonstationarity issue and select discriminative features to achieve better generalization performance.

III. Notations and Preliminaries

We first describe the notations and preliminaries that will be used in this article. Following the standard notations [32], we represent scalars by lowercase letters (e.g., x), vectors by bold lowercase letters (e.g., $\mathbf{x}$), matrices by bold uppercase letters (e.g., $\mathbf{X}$), and higher-order tensors (mode three or higher) by calligraphic uppercase letters (e.g., $\mathcal{X}$). We represent a K-mode tensor as $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ that contains $N = \prod_{k=1}^{K} n_k$ elements. A mode-k fiber of tensor $\mathcal{X}$ is an n_k-dimensional vector denoted by fixing all but the k_{th} index. A mode-k unfolding of $\mathcal{X}$, denoted by $\mathcal{X}_{(k)} \in \mathbb{R}^{n_k \times N/n_k}$, is obtained by arranging all of the N/n_k mode-k fibers along its column. $\mathcal{X}$ has multilinear ranks $(r_1, \ldots, r_K)$ and $r_k = \text{rank}(\mathcal{X}_{(k)})$. The inner product of two same-sized tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is denoted by $<\mathcal{X}, \mathcal{Y}> = \text{vec}(\mathcal{X})^T \text{vec}(\mathcal{Y})$. The Frobenius norm of tensor $\mathcal{X}$ is defined as $\|\mathcal{X}\|_F = \sqrt{<\mathcal{X}, \mathcal{X}>}$ and the ℓ_1-norm $\|\mathcal{X}\|_1$ is the sum of absolute values of all elements.

Proximal Operators: The proximal operator of a convex function $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$\text{prox}_{\lambda f}(\mathbf{X}) = \arg\min_{\mathbf{Z} \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2}\|\mathbf{Z} - \mathbf{X}\|_F^2 + \lambda f(\mathbf{Z}) \tag{1}$$

with any scalar $\lambda \geq 0$ [33]. Specifically, if $f = \|\mathbf{Z}\|_1$, the proximal operator for the ℓ_1-norm is the elementwise soft thresholding operator

$$\text{prox}_{\lambda \|\cdot\|_1}(\mathbf{X}) = \text{sgn}(\mathbf{Z}) \circ [|\mathbf{Z}| - \lambda]_+ \tag{2}$$

where sgn is the sign function and $[x]_+ = \max(x, 0)$. The operator $\circ$ in (2) denotes the elementwise product.

The proximal operator for nuclear norm $f = \|\mathbf{Z}\|_*$ is given by the shrinkage operation as follows:

$$\text{prox}_{\lambda \|\cdot\|_*}(\mathbf{X}) = \mathbf{U}\text{diag}\big(\text{prox}_{\lambda \|\cdot\|_1}(\mathbf{d})\big)\mathbf{V}^T \tag{3}$$

where $\mathbf{X} = \mathbf{U}\text{diag}(\mathbf{d})\mathbf{V}^T$ is the singular value decomposition of $\mathbf{X}$ with the singular values in $\mathbf{d}$.

IV. Method

In this section, we present our tensor-based MTL method for joint discriminative feature selection and classification. To solve the resulting optimization problem, we develop an efficient algorithm to obtain the global optimum based on the ADMM framework.

A. Problem Formulation

The focus of this article is the multitask classification problem for tensor data such as single-trial EEG signals. We set the categorization of EEG signals for each subject as a separate task; "multiple tasks" means to simultaneously learn classifiers for multiple subjects, and each of them undergoes the same mental tasks. Different from most existing EEG classification methods, we explore the MTL for EEG signals for two reasons.

1) It has been proved in the literature that due to the common ground, the principal feature characteristics are invariant across subjects [34]. And the MTL framework is to discover important shared characteristics of the related tasks.
2) In terms of EEG, it is very time consuming and inconvenient to collect enough annotated data [35]. The MTL framework can investigate data from multiple related tasks and thus overcome the data scarcity problem.

In this article, we consider T different but related EEG classification tasks. Each task consists of m_t sample pairs $\{(\mathcal{X}_i^t, y_i^t)\}_{1 \leq i \leq m_t}^{1 \leq t \leq T}$, where $\mathcal{X}_i^t \in \mathbb{R}^{n_1 \times \cdots \times n_K}$ is a covariate tensor drawn from a feature space shared by all tasks, and $y_i^t \in \{1, -1\}$ denotes the corresponding label. Based on MTL, the proposed method simultaneously learns T related models with the model parameters $\{\mathcal{W}^t, b^t\}_{t=1}^T$. Given the newly observed sample $\hat{\mathcal{X}}^t$, we can predict its label using the linear model with $\hat{y}^t = \text{sgn}(<\mathcal{W}^t, \hat{\mathcal{X}}^t> + b^t)$. Based on the regularized loss minimization framework, we formulate the proposed tensor-based MTL for classification as follows:

$$\min_{\{\mathcal{W}^t, b^t\}_{t=1}^T} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \ell\big(\mathcal{W}^t, b^t, \mathcal{X}_i^t, y_i^t\big) + R\big(\mathcal{W}^t\big) \tag{4}$$

where ℓ denotes an empirical loss function and R is a regularization term for simultaneously selecting discriminative features and leveraging the shared information between different tasks to improve the classification performance.

For the measurement of empirical loss, the hinge loss, as a relaxation of 0/1 loss, is desirable for its maximal margin principle. It has been widely used in the classification methods due to its robustness and sparsity such as SVM-like classifiers [36], [37]. Thus, we employ the hinge loss in (4).

The inherent nonstationarity in the EEG signals is so complex that it tends to deteriorate the classification performance. To tackle the nonstationarity of the training samples within the same class, we seek to identify the most discriminative features by penalizing the within-class variance and meanwhile maximizing the between-class distance based on the Fisher discriminant criterion. Specifically, the Fisher discriminant criterion is employed as a stationarity regularizer, which pushes the samples from two classes far away from the decision boundary and minimizes the within-class variations to alleviate the nonstationarity of the training samples. In this article, instead of selecting important elements within tensor data itself, we consider its dual-learning weights of the classifiers. The extracted tensor-form features, which are parameterized by the weights of classifiers, should be discriminative in the subspace. Thus, we employ the Fisher linear discriminant criterion on $\mathcal{W}^t$ to minimize the within-class variation and maximize the between-class boundary at the same time

$$\begin{aligned} S\big(\mathcal{W}^t\big) = &\sum_{c=1}^{2} \sum_{j=1}^{m_t^c} \Big(<\mathcal{W}^t, \mathcal{X}_j^t - \bar{\mathcal{X}}_c^t>\Big)^2 \\ &- \sum_{c=1}^{2} m_t^c \big(<\mathcal{W}^t, \bar{\mathcal{X}}_c^t - \bar{\mathcal{X}}^t>\big)^2 \end{aligned} \tag{5}$$

where $\bar{\mathcal{X}}_c^t$ denotes the mean of training samples of the cth class in the t-th task and $\bar{\mathcal{X}}^t$ represents the mean of training samples in the t-th task. m_t^c is the number of training samples of the cth class in the t-th task.

To make use of the fact that the related tasks learn from each other in the training process, we consider both the shared and task-specific patterns over multiple tasks. To leverage the shared knowledge from multiple tasks, we assume that the feature space is not independent among different subjects and only a subspace of feature space is useful for classification, namely, the EEG features should be of low rankness. Inspired by [12] and [38], we model the shared structural information by analyzing the singular value spectra of learning coefficients with low-rank regularization. On the other hand, considering that sparse representation modeling has been successfully employed for noise-robust EEG data classification [39]–[41], we employ the sparsity to model the divergence from the shared structural patterns for each individual task. Thus, we decompose the learning coefficients of each classifier to be the sum of a low-rank task-shared tensor and a sparse task-specific tensor

$$\mathcal{W}^t = \mathcal{P} + \mathcal{Q}^t, \ \text{s.t. rank}(\mathcal{P}) \le r, ||\mathcal{Q}^t||_0 \le s \tag{6}$$

where rank($\mathcal{P}$) denotes the multilinear ranks given in Section III and we constrain $r_k \le r$ for any k, and the ℓ_0-norm $||\cdot||_0$ for a tensor counts the number of its nonzero entries. In (6), the first component $\mathcal{P}$ captures the low-rank correlative structure shared across multiple related tasks. The second component $\mathcal{Q}^t$ identifies the sparse patterns for each individual task.

It is NP-hard to solve the tensor rank and ℓ_0 constraints. Similar to matrix analysis, we approximate the nonconvex ℓ_0-norm into a convex surrogate with the ℓ_1-norm. However, the tensor rank has not been well studied like in the matrix domain. The recent research in tensor decomposition led to a scaled latent trace norm, which has shown better generalization performance, especially for the case of heterogeneous multilinear ranks [42]. Inspired by this novel tensor norm, we are the first to investigate it to extract the shared structure information across multiple classification tasks. We decompose the target shared K-mode tensor $\mathcal{P}$ into a linear combination of K latent tensors as $\mathcal{P} = \sum_{k=1}^K \mathcal{P}^{(k)}$, and regularize the nuclear norm of the mode-k unfolding of the k_{th} latent tensor as

$$||\mathcal{P}||_{tr} = \inf_{\mathcal{P}^{(1)}+\cdots+\mathcal{P}^{(K)}=\mathcal{P}} \sum_{k=1}^K \frac{1}{\sqrt{n_k}} \left\| \mathcal{P}_{(k)}^{(k)} \right\|_*. \tag{7}$$

Different from the conventional methods, which learn the feature selection and classification in two separated steps, we jointly learn the discriminative features and multitask classification in a unified framework, leading to the following optimization problem:

$$\begin{aligned} \min_{\{\mathcal{W}^t, b^t, \mathcal{Q}^t\}_t^T, \mathcal{P}} \ & \sum_{t=1}^T \sum_{i=1}^{m_t} \left[1 - y_i^t \left(<\mathcal{W}^t, \mathcal{X}_i^t> + b^t\right)\right]_+ \\ & + \frac{\lambda}{2} \sum_{t=1}^T S\left(\mathcal{W}^t\right) + \tau \|\mathcal{P}\|_{tr} + \sum_{t=1}^T \beta \|\mathcal{Q}^t\|_1 \\ \text{s.t.} \quad & \mathcal{W}^t = \mathcal{P} + \mathcal{Q}^t, \qquad \forall t = 1, \ldots, T \end{aligned} \tag{8}$$

where λ, τ, and β are the positive regularization parameters. Intuitively, when $\lambda = 0$, the Fisher discriminant is ineffective and our method degrades to the MTL version of [30]; when $\tau = 0$, our method degrades to solve multiple single-task classification problems with the ℓ_1-norm; and when $\beta = 0$, our method degrades to multitask feature learning without considering the task-specific information.

B. Learning Algorithm

The composite problem in (8) is difficult to be solved since the hinge loss, scaled latent trace norm, and ℓ_1-norm are all nonsmooth and nondifferentiable. Fortunately, the problem in (8) is convex due to the convexity of each subcomponent and, thus, it has a global optimum. Here, we resort to the ADMM framework, which splits the original convex problem into several easier subproblems. The subproblems can then be solved by nonexpensive proximal operators. We first substitute (7) into (8) and use the augmented Lagrangian method to reformulate the constrained problem into an unconstrained one as

$$\begin{aligned} & \min_{\{\mathcal{W}^t, b^t, \mathcal{Q}^t, \mathcal{Y}^t\}_t^T, \{\mathcal{P}^{(k)}\}_{k=1}^K} L \\ & = \sum_{t=1}^T \sum_{i=1}^{m_t} \left[1 - y_i^t \left(<\mathcal{W}^t, \mathcal{X}_i^t> + b^t\right)\right]_+ + \frac{\lambda}{2} \sum_{t=1}^T S\left(\mathcal{W}^t\right) \\ & \quad + \sum_{k=1}^K \tau_k \left\| \mathbf{P}_{(k)}^{(k)} \right\|_* + \sum_{t=1}^T <\mathcal{Y}^t, \mathcal{W}^t - \sum_{k=1}^K \mathcal{P}^{(k)} - \mathcal{Q}^t> \\ & \quad + \sum_{t=1}^T \beta \|\mathcal{Q}^t\|_1 + \frac{\gamma}{2} \sum_{t=1}^T \left\| \mathcal{W}^t - \sum_{k=1}^K \mathbf{P}_{(k)}^{(k)} - \mathcal{Q}^t \right\|_F^2 \end{aligned} \tag{9}$$

where $\{\mathcal{Y}^t\}_{t=1}^T$ are a sequence of Lagrangian multipliers over multiple tasks, γ is a positive scalar, and $\tau_k = (\tau/\sqrt{n_k})$ is the scaled weight for the mode-k unfolding of the kth latent shared tensor.

Then, we decouple the unconstrained problem in (9) into several subproblems based on ADMM and solve them alternatively. The ADMM learning procedures are summarized in Algorithm 1. The key steps in Algorithm 1 are the computation of $\mathcal{W}^t$, b^t, $\mathcal{P}$, and $\mathcal{Q}^t$, the derivation of which is shown in the following theorems.

Theorem 1: For each task, given the shared structure $\mathcal{P}$ and task-specific information $\mathcal{Q}^t$, one of the solutions to the optimization problem

$$\begin{aligned} \arg\min_{\mathcal{W}^t, b^t} \ & \sum_{i=1}^{m_t} \left[1 - y_i^t \left(<\mathcal{W}^t, \mathcal{X}_i^t> + b^t\right)\right]_+ + \frac{\lambda}{2} S\left(\mathcal{W}^t\right) + \\ & <\mathcal{Y}^t, \mathcal{W}^t> + \frac{\gamma}{2} \left\| \mathcal{W}^t - \sum_{k=1}^K \mathcal{P}^{(k)} - \mathcal{Q}^t \right\|_F^2 \end{aligned} \tag{10}$$

is

$$\begin{aligned} \hat{\mathcal{W}}^t &= \frac{\gamma \left(\sum_{k=1}^K \mathcal{P}^{(k)} + \mathcal{Q}^t \right) - \mathcal{Y}^t + \sum_{i=1}^{m_t} \alpha_i^t y_i^t \mathcal{X}_i^t}{\lambda D^t + \gamma} \\ \hat{b}^t &= \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(y_i^t - <\mathcal{W}^t, \mathcal{X}_i^t> \right) \end{aligned} \tag{11}$$

Algorithm 1: Learning Algorithm for Our method

Input : Training data $\{(\mathcal{X}_i^t, y_i^t)\}_{1\le i\le m_t}^{1\le t\le T}$, input coefficients λ,τ and β
Output: $\{\mathcal{W}^t, b^t\}^{1\le t\le T}$
1 Initialize: $\{\mathcal{Y}^t\}_{t=1}^T = \mathbf{0}$,$\{\mathcal{P}^{(k)}\}_{k=1}^K = \mathbf{0}$,$\{\mathcal{Q}^t\}_{t=1}^T = \mathbf{0}$, $\gamma = 1$
while *not converge* **do**
2 **for** $t \leftarrow 1$ ***to*** T **do**
3 Calculate $\mathbf{M}$ and $\mathbf{q}$ with Eq. (13)
4 Calculate $\boldsymbol{\alpha}$ by solving Eq. (12)
5 Update $\{\mathcal{W}^t, b^t\}$ with Eq. (11)
6 **end**
7 **for** $k \leftarrow 1$ ***to*** K **do**
8 Update $\mathcal{P}^{(k)}$ with Eq. (15)
9 **end**
10 Calculate $\mathcal{P}$ by $\mathcal{P} = \sum_{k=1}^K \mathcal{P}^{(k)}$
11 **for** $t \leftarrow 1$ ***to*** T **do**
12 Update $\mathcal{Q}^t$ with Eq. (17)
13 Update Lagrangian multipliers $\mathcal{Y}^t$ by $\mathcal{Y}^t \leftarrow \mathcal{Y}^t + \gamma(\mathcal{W}^t - \mathcal{P} - \mathcal{Q}^t)$
14 **end**
15 **end**

where $\mathcal{B} = \{i : 0<\alpha_i^t < 1\}$, $|\mathcal{B}|$ is the number of elements in the subset $\mathcal{B}$, $D^t = \sum_{c=1}^2 \sum_{j=1}^{m_t^c} ||\mathcal{X}_j^t - \bar{\mathcal{X}}_c^t||_F^2 - \sum_{c=1}^2 m_t^c ||\bar{\mathcal{X}}_c^t - \bar{\mathcal{X}}^t||_F^2$ and $\{\boldsymbol{\alpha}^t\} \in \mathbb{R}^{m_t}$ are the dual variables of the following box constraint quadratic programming (QP):

$$\begin{aligned} \arg\max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{M}\boldsymbol{\alpha} + \mathbf{q}^T\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{0} \le \boldsymbol{\alpha} \le \mathbf{1} \\ & \sum_{i=1}^{m_t} \alpha_i^t y_i^t = 0. \end{aligned} \tag{12}$$

Specifically

$$\begin{aligned} M_{ij} &= \frac{y_i^t y_j^t <\mathcal{X}_i^t, \mathcal{X}_j^t>}{\lambda D^t + \gamma} \\ q_i &= 1 - \frac{<\gamma\left(\sum_{k=1}^K \mathcal{P}^{(k)} + \mathcal{Q}^t\right) - \mathcal{Y}^t, y_i^t \mathcal{X}_i^t>}{\lambda D^t + \gamma}. \end{aligned} \tag{13}$$

To figure out $\mathcal{W}^t$ directly may be difficult and time consuming, since $\mathcal{W}^t$ contains $N = \prod_{k=1}^K n_k$ variables. By Theorem 1, solving $\mathcal{W}^t$ and b^t can be replaced by solving (12) involving m_t variables, which is more efficient in the case of $m_t \ll N$.

Theorem 2: For positive τ_k and γ, the mode-k unfolding of the solution in the following problem:

$$\begin{aligned} \arg\min_{\mathbf{P}_{(k)}^{(k)}} \quad & \sum_{k=1}^K \tau_k \left\|\mathbf{P}_{(k)}^{(k)}\right\|_* - \sum_{t=1}^T \left(<\mathcal{Y}^t, \sum_{k=1}^K \mathcal{P}^{(k)}>\right) \\ & + \frac{\gamma}{2}\sum_{t=1}^T \left\|\mathcal{W}^t - \sum_{k=1}^K \mathcal{P}^{(k)} - \mathcal{Q}^t\right\|_F^2 \end{aligned} \tag{14}$$

is

$$\hat{\mathbf{P}}_{(k)}^{(k)} = \text{prox}_{\frac{\tau_k}{\gamma T}||\cdot||_*}\left(\frac{1}{T}\sum_{t=1}^T\left(\mathbf{W}_{(k)}^t - \mathbf{Q}_{(k)}^t + \frac{\mathbf{Y}_{(k)}^t}{\gamma}\right) - \sum_{j\ne k}^K \mathbf{P}_{(k)}^{(j)}\right). \tag{15}$$

By Theorem 2, we can obtain each $\mathcal{P}^{(k)}$ by tensorization of $\mathbf{P}_{(k)}^{(k)}$ in mode k.

Theorem 3: For positive β and γ, one of the solutions of the following problem:

$$\begin{aligned} \arg\min_{\mathcal{Q}^t} F(\mathcal{Q}^t) = \beta\sum_{t=1}^T \|\mathcal{Q}^t\|_1 - \sum_{t=1}^T <\mathcal{Y}^t, \mathcal{Q}^t> \\ + \frac{\gamma}{2}\sum_{t=1}^T \left\|\mathcal{W}^t - \sum_{k=1}^K \mathcal{P}^{(k)} - \mathcal{Q}^t\right\|_F^2 \end{aligned} \tag{16}$$

is

$$\hat{\mathcal{Q}}^t = \text{prox}_{\frac{\beta}{\gamma}||\cdot||_1}\left(\mathcal{W}^t + \frac{\mathcal{Y}^t}{\gamma} - \sum_{k=1}^K \mathcal{P}^{(k)}\right). \tag{17}$$

The proofs of all above theorems can be found in the Appendix.

In general, our learning algorithm is newly derived from the ADMM framework to efficiently address the multitask tensor classification problem. For each iteration in Algorithm 1, steps 2–6 solve multitask tensor classification parameters by the QP, steps 7–9 constrain the mode-k rank for the kth latent tensor with proximal operators for nuclear norm, and steps 11–14 are also different from the standard ADMM framework, which minimize the ℓ_1-norm of $\mathcal{Q}^t$ and update the Lagrangian multipliers in a multitask manner.

Time Complexity: The most exhaustive step for each iteration in Algorithm 1 is the QP in (12), which costs $O(m_t^2 N)$ floating-point operations to solve $\{\mathcal{W}^t, b^t\}$ for each task t. Therefore, $\{\mathcal{W}^t, b^t\}_{t=1}^T$ can be analytically computed in a time complexity of $O(\max(m_1, \ldots, m_T)^2 N)$, where m_t, m, N, and T denote the number of training samples in the t-th task, the total training sample, sample dimensionality, and the number of tasks, respectively. Other than the QP, it costs $O(\min(n_k N, [N^2/n_k]))$ to compute the eigen decomposition for each latent tensor $\mathcal{P}^{(k)}$. In addition, it costs $O(TN)$ to calculate the proximal operators for $\mathcal{Q}$, which can be negligible compared with those of QP. Therefore, the aforementioned optimization procedure takes a time complexity of $O(m^2 N)$.

V. Experiments

We validate the proposed method on motor-imagery (MI)-based EEG classification problems in the context of brain–computer interfaces (BCIs). The objective of the experiments is to detect two-class or multiclass motor activities from the EEG signals. For our experiments, we use three publicly available real EEG datasets: 1) dataset IVa of BCI competition III; 2) dataset IIb of BCI competition IV for binary classification; and 3) dataset IIa of BCI competition IV for multiclass classification.

A. Real EEG Datasets for Evaluation

1) Dataset IVa of BCI Competition III [43][2]*:* This dataset contains EEG signals from five subjects (subjects *al, aa, av, aw*, and *ay*) when they were performing right-hand and foot MI

[2]http://www.bbci.de/competition/iii/#data_set_iva

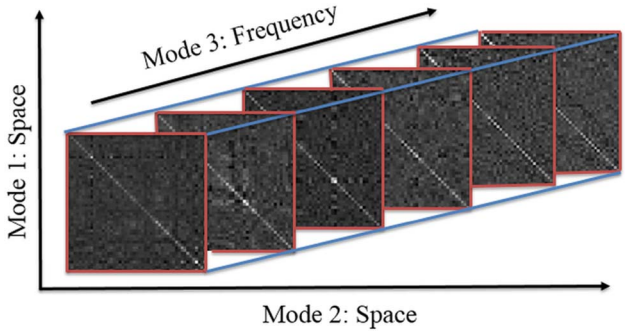

Fig. 1. Illustration of three-mode tensor of an EEG sample.

tasks. It recorded 280 samples for each subject. Following the work in [28], we also select 49 channels for succedent analysis.

2) Dataset IIb of BCI Competition IV [44][3]: This dataset recorded three bipolar-channel EEG signals from nine subjects (denoted as *B01–B09*) involving left-hand and right-hand MI tasks. There are about 720–760 samples for each subject.

3) Dataset IIa of BCI Competition IV [13][4]: The EEG data of this dataset were obtained from nine subjects (denoted as *S01–S09*), who were asked to perform four different MI tasks, that is, left hand, right hand, foot, and tongue. The EEG signals were measured by 22 electrodes. It contains 144 samples for per class per subject.

To model the tensor-form data for evaluation, we employ the following preprocessing steps. We use z ($z = 6$) nonoverlapped band-pass Butterworth filters with cutoff frequencies of $(8, 12)$, $(12, 16)$, $(16, 20)$, $(20, 24)$, $(24, 28)$, and $(28, 32)$ to filter out unrelated signals. Let $\mathbf{S}_i \in \mathbb{R}^{C\times p}$, where C denotes the number of channels, p is the number of sampled time points, and S_i be the matrix obtained by processing with the ith filter. Similar to [30], each covariance matrix $\mathbf{X}_i$ is defined as $\mathbf{X}_i = \hat{S}_i\hat{S}_i^T \in \mathbb{R}^{C\times C}$, where $\hat{S}_i = (1/\sqrt{p-1})S_i(\mathbf{I}_p - (1/p)\mathbf{1}\mathbf{1}^T)$ is the input signal after centering and scaling. Then, we arrange all $\mathbf{X}_i$, $i = 1, \ldots, z$ to form a sized $C \times C \times z$ tensor. Thus, samples in the above three datasets are of size $49 \times 49 \times 6$, $3 \times 3 \times 6$, and $22 \times 22 \times 6$, respectively. One of the samples in the first dataset is illustrated in Fig. 1.

B. Experimental Settings

There is no tensor-based MTL method for classification yet. To demonstrate the advantage of our method, we compare our method with the following state-of-the-art techniques.

1) *STL_Latent [30]:* A tensor-based classifier with the scaled latent trace norm for single-task learning (STL).
2) *STL_Overlap [30]:* A tensor-based classifier with the overlapped trace norm for STL. Both STL_Latent and STL_Overlap are chosen as benchmark, since they make no assumption on the relationships among tasks.
3) *MTL_Lasso [45]:* A vector-based MTL method with lasso regularization.
4) *MTL_Trace [21]:* A vector-based MTL method with low-rank regularization.
5) *MTL_L21 [46]:* A vector-based MTL method with $\ell_{2,1}$-norm regularization.
6) *RMTFL [47]:* A vector-based MTL method for joint feature selection and classification.
7) *GO-MTL [24]:* A vector-based MTL method for modeling task grouping and overlap.
8) *SSMM [41]:* A sparse support matrix machine for joint feature selection and classification, which is the state-of-the-art for EEG classification.

To illustrate the effect of the Fisher discriminant and tensor decomposition in the proposed method, we also perform experiments by setting hyperparameters λ and β to 0, respectively, which are denoted as Ours_λ0 and Ours_β0.

In each task, we randomly sample 80% for training and the remaining for testing. We also stack the tensor-form data into vectors before fitting the vector-based MTL methods. We first conduct the binary classification on the first two datasets and extend all methods for multiclass classification on the third dataset via the one-versus-the-rest (OvR) strategy. To measure the classification performance, we calculate the error rate and the lower error rate represents better performance.

C. Results

1) Parameter Sensitivity: We first test the influence of parameters for the proposed method. There are three parameters, namely λ, β, and τ, in our method, which are proposed to capture the feature selection, task-specific information, and shared-structural among tasks, respectively. To test the influence of them, we first fix the value of $\lambda = 0, 0.1, 1$, respectively, to validate the performance of feature selection. Then, we tune the parameters β and τ accordingly. We show the classification performance of our method on the dataset IVa of BCI Competition III with different parameter settings in Fig. 2. We highlight the lower error rate with darker blue. It can be observed from Fig. 2 that when either λ, β, or τ is set larger than 0, the performance is rather consistently better than when they are set to 0. Similar phenomena also occur when employing our method on other EEG datasets. Thus, it can be indicated that our method considering both feature selection and multitask tensor learning is promising in real EEG classification.

In our experiments, we empirically select the best hyperparameters λ, τ, and β with grid search via cross-validation. For fair comparison, all the hyperparameters involved in the compared methods are also fine tuned by cross-validation.

2) Binary Classification: Table I lists the testing error rates for each task and the mean error rates for all tasks of different methods on the first dataset. The lowest error rates are highlighted in boldface. It can be observed that MTL_Lasso obtains the highest error and our method achieves the lowest error among the mean error rates of different methods. This demonstrates the advantage of tensor representation in our method. Both STL_Overlap and STL_Latent are outperformed by our method, which indicates that the shared information among multiple related tasks are useful for performance improvement. The average error rates of Ours_λ0 and Ours_β0 are higher than the proposed method but lower than other compared

[3]http://www.bbci.de/competition/iv/#dataset2b
[4]http://www.bbci.de/competition/iv/#dataset2a

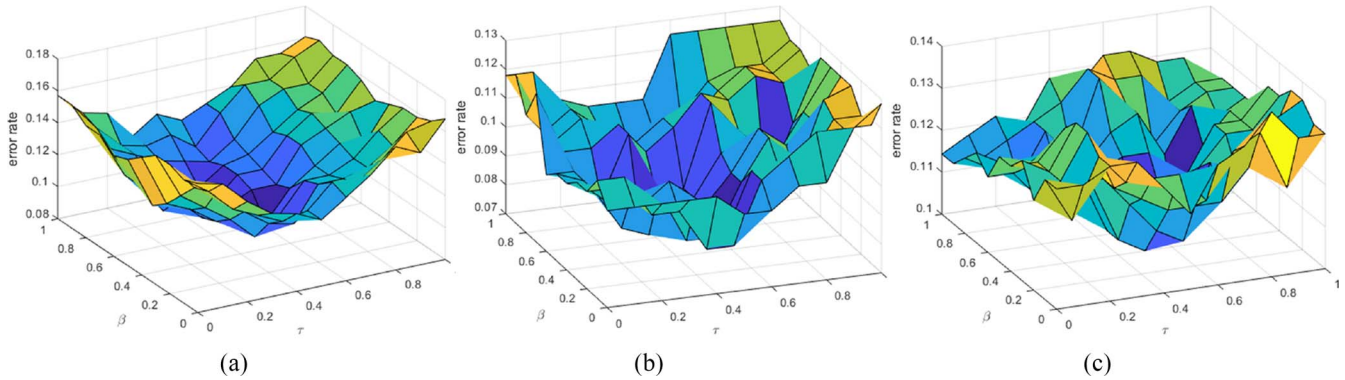

Fig. 2. Error rates of our method on the dataset BCI III IVa with different parameter settings.

TABLE I
TESTING ERROR OF DIFFERENT METHODS ON DATASET IVA OF BCI COMPETITION III

Tasks	STL_Overlap	STL_Latent	MTL_Lasso	MTL_Trace	MTL_L21	RMTFL	GO-MTL	SSMM	Ours_λ0	Ours_β0	Ours
aa	0.1429	0.1250	0.1429	0.1071	0.1071	**0.0893**	0.1429	0.1429	0.1429	0.1250	**0.0893**
al	0.0179	**0**	0.0179	0.0179	**0**	**0**	0.0179	**0**	**0**	**0**	**0**
av	0.3571	0.2679	0.3570	0.2857	0.2857	0.3036	0.2679	**0.1964**	0.2143	0.3036	0.2321
aw	0.0357	0.0179	0.0893	0.0179	0.0893	0.0357	0.0536	0.0536	0.0357	0.0179	**0**
ay	0.0893	0.1071	0.1250	0.0893	0.1250	0.0893	0.0714	0.0536	0.0714	**0.0357**	0.0714
avg	0.1286	0.1036	0.1500	0.1036	0.1214	0.1036	0.1107	0.0893	0.0929	0.0964	**0.0786**

TABLE II
TESTING ERROR OF DIFFERENT METHODS ON DATASET IIB OF BCI COMPETITION IV

Tasks	STL_Overlap	STL_Latent	MTL_Lasso	MTL_Trace	MTL_L21	RMTFL	GO-MTL	SSMM	Ours_λ0	Ours_β0	Ours
B01	0.1667	0.1944	0.1944	**0.1528**	0.1806	0.2014	0.2222	0.3056	0.1875	0.1736	0.1597
B02	0.4485	0.4412	0.4338	0.4779	0.4265	0.4338	**0.4191**	0.4779	0.4338	0.4265	0.4265
B03	0.4653	**0.4514**	0.4722	0.4772	0.4931	0.4583	0.4722	0.4653	**0.4514**	0.4722	**0.4514**
B04	0.0405	0.0473	0.0541	0.0473	0.0473	0.0405	0.0473	**0.0270**	0.0473	0.0473	0.0405
B05	0.1081	0.0946	0.1149	0.1284	0.1081	0.1014	0.1149	0.1081	0.1014	0.1081	**0.0878**
B06	0.2014	0.1806	0.2014	0.1875	0.2014	0.1944	**0.1667**	0.1944	0.1875	0.2083	0.1736
B07	0.2361	0.2431	0.2292	0.2292	0.2222	0.2361	0.2500	0.2500	0.2431	0.2153	**0.2083**
B08	0.1645	0.1842	0.1908	**0.1579**	0.1645	0.1776	0.1973	0.1908	0.1974	0.1842	0.1645
B09	0.1875	0.2014	0.2014	**0.1806**	0.1875	**0.1806**	0.1875	0.2014	0.1944	0.2153	**0.1806**
avg	0.2243	0.2265	0.2301	0.2232	0.2232	0.2224	0.2285	0.2439	0.2247	0.2255	**0.2103**

methods except for SSMM. The performance advantages of our method over Ours_λ0 and Ours_β0 methods demonstrate that considering both feature selection and sparse patterns do help for multitask classification modeling. It is worth noting that the average performance of the proposed method surpasses that of SSMM, the state-of-the-art for EEG classification. The SSMM is a sparse support matrix machine method for joint feature selection and classification. It is a matrix-based STL classifier. Compared with SSMM, the proposed method is a tensor-based MTL method. The advantages of the proposed method are twofold. On the one hand, tensor-based learning is capable of efficiently leverage the inherent structural information from high-order tensor samples. On the other hand, MTL jointly learns multiple related tasks so that sample size can be effectively increased and knowledge obtained from each task can be shared in the learning process. In this way, MTL is an effective learning paradigm to alleviate the labeled data deficiency issue and improve the generalization performance of multiple related tasks.

The testing error rates on the second dataset are shown in Table II. The proposed method again achieves the lowest mean error rate than all compared methods, which demonstrates the effectiveness and robustness of our joint feature selection and multitask classification model.

3) Multiclass Classification: We also evaluate our method in the multiclass classification problem on the third dataset. Table III reports the classification error rates. The results show that the STL methods (both STL_Overlap and STL_Latent) are outperformed by the MTL methods. It indicates that the shared patterns among multiple related tasks are useful in the learning process. Among the MTL methods, both RMTFL and our method jointly learn the discriminative features when fitting the classification models and achieve better overall results compared with the other methods. Therefore, jointly selecting the important features also benefits the model learning for classification. However, RMTFL still cannot beat ours because it requires data vectorization and ignores structural information unavoidably. Therefore, our method, which simultaneously leverages knowledge among all related tasks and learns discriminative tensor-based features, can achieve the best performance.

TABLE III
TESTING ERROR OF DIFFERENT METHODS ON DATASET IIA OF BCI COMPETITION IV

Tasks	STL_Overlap	STL_Latent	MTL_Lasso	MTL_Trace	MTL_L21	RMTFL	GO-MTL	SSMM	Ours_λ0	Ours_β0	Ours
S01	0.1897	0.2155	0.1552	0.2328	0.1466	0.1466	**0.1293**	0.1379	0.1810	0.1897	0.1724
S02	0.4052	0.4483	0.3448	0.3534	0.3783	0.3793	0.3621	0.3707	0.4483	0.3879	**0.3103**
S03	**0.1121**	0.1207	0.1552	0.1379	0.1379	0.1207	0.1552	0.1207	**0.1121**	0.1552	**0.1121**
S04	0.3276	0.2759	0.2759	0.2672	0.2845	0.2500	0.2586	0.2586	0.3017	0.2845	**0.2328**
S05	0.4224	0.4138	**0.3879**	0.4741	0.4655	0.4569	0.4310	0.3966	0.4483	0.4310	0.4224
S06	0.4052	0.4052	0.4397	0.3879	0.4397	0.3966	0.4569	0.4138	0.3966	0.3966	**0.3707**
S07	0.2069	0.1810	0.1638	0.1724	0.1897	0.2155	0.1810	**0.1207**	0.1897	0.2328	0.2241
S08	0.1293	0.1293	0.1724	0.1293	0.1207	0.1293	0.1552	0.1379	**0.1121**	0.1293	0.1379
S09	0.1724	0.2241	0.1638	0.1121	0.1638	0.1552	0.1552	0.1897	0.1638	**0.0948**	0.1379
avg	0.2634	0.2682	0.2510	0.2519	0.2586	0.2500	0.2538	0.2385	0.2605	0.2558	**0.2356**

TABLE IV
APVS FROM HOLM METHOD BETWEEN OUR METHOD AND OTHER COMPARED METHODS

Methods	STL_Overlap	STL_Latent	MTL_Lasso	MTL_Trace	MTL_L21	RMTFL	GO-MTL	SSMM	Ours_λ0	Ours_β0
APVs	0.0012	0.0063	1.6e-5	0.0125	0.0003	0.0207	0.0016	0.0086	0.0087	0.0055

TABLE V
TRAINING TIME ON THE THREE DATA SETS (IN SECONDS)

Dataset	STL_Overlap	STL_Latent	MTL_Lasso	MTL_Trace	MTL_L21	RMTFL	GO-MTL	SSMM	Ours
BCI III IVa	583.9	449.9	7.6	14.5	13.7	72.1	15.7	20.8	28.4
BCI IV IIb	420.3	137.0	1.7	3.8	3.3	16.2	6.5	223.7	58.4
BCI IV IIa	4308.3	1212.4	15.6	18.4	17.9	152.9	76.7	1990.0	94.3

4) Statistical Test: We further employ the hypothesis-testing technique to find the significant differences among the results obtained by our method and all compared methods. Specifically, we first employ the Friedman test as well as the Iman-Davenport test [48] to check whether there are significant differences in the performance among all methods. The p-values of the Friedman and Iman-Davenport statistic on the classification errors of all subjects are 1.378e-3 and 8.963e-4, respectively. The results indicate that the null hypothesis is rejected (p-values < 0.05); thus, there is a significant difference among the performance of all methods. We then employ the Holm method [49] to calculate the adjusted p-values (APVs) for the pairwise comparison between our method and each of the compared methods. We present the resulting APVs in Table IV. Since all of the APVs are smaller than 0.05, it can be safely concluded that the proposed method is statistically better than the compared method regarding the measures of classification error with a significant level of 0.05.

5) Computational Time: We further compare the computational efficiency of all classification models on the three datasets. Among all the classification models, STL_Overlap, STL_Latent, and our method work for data in tensor form, SSMM is for matrices and the remaining are for vectors. On a workstation with Intel Xeon E5-1620 3.70-GHz CPU and 16.0-GB RAM, and with a MATLAB implementation, the training time of all models on the three datasets is reported in Table V. There are three interesting observations. First, almost all of the STL methods (STL_Overlap, STL_Latent, and SSMM) are slower than MTL methods on these three datasets. This is because MTL methods, including ours, can train classifiers for all tasks at the same time, while single-task ones have to repeat the training process one by one. Therefore, our method is faster than the other two tensor classification models (STL_Overlap and STL_Latent). Second, compared with the vector-form MTL methods (MTL_Lasso, MTL_Trace, MTL_L21, RMTFL, and GO-MTL), our method takes relatively longer training time. The reason might be that our method takes high-order structural information as well as feature selection into consideration. Finally, the training time of our method is more stable than RMTFL and SSMM, which also consider the feature selection issue. This might benefit from the multitask training scheme as well as the efficient ADMM solver.

VI. DISCUSSION

The proposed method is the first multitask tensor classifier to train multiple related classification models for tensor-form EEG samples at the same time. During the learning process, it is capable of jointly selecting the discriminative features, and leveraging both shared and individual-specific structural information to improve the generalization performance. This benefits from the novel combination of the Fisher discriminant criterion, scaled latent trace norm, and ℓ_1-norm on the regression tensor.

Though previous work [30] also proposed a tensor classification model based on the scaled latent trace norm, it is motivated to consider each classification task independently and extracted the high-order structural information within tensors for each classifier. It just assumes the input EEG signals are noise free, and thus lacks robustness to outliers and noises in the real-world applications. In this regard, our method not

only extends [30] to an MTL paradigm but also performs feature selection to select discriminative features and account for the nonstationarity problem. In addition, to solve the resulting nondifferentiable optimization problem, we newly derive an efficient learning algorithm based on ADMM, which is also different from [30].

Furthermore, the proposed method is general enough to work for tensors in other domains except EEG classification if the following conditions are satisfied: 1) the multiple classification tasks are related and each has similar training samples and 2) the given training samples can be inherently arranged into tensors and the shared structure information is of low rank. The proposed method is able to learn multiple tasks simultaneously and explores the shared structural information to produce better generalization performance. It is also robust for nonstationarity cases.

VII. Conclusion

Traditional MTL methods are designed for vector-form data while many modern applications are producing high-dimensional data that can be naturally modeled as tensors. In this article, we have presented a regularized tensor-based MTL method to explore the inherent structural information. Different from the existing methods, our method integrated the feature selection and classification in a unified MTL framework. We employed the Fisher discriminant criterion to select discriminative features and control the nonstationarity among samples, which minimized the within-class variance and meanwhile maximized the between-class distances. To leverage the relationships among all related tasks, we decomposed the regression tensor of each task into a linear combination of a shared tensor and a task-specific tensor. Specifically, we investigated the scaled latent trace norm to model the common structure within the shared tensor, and ℓ_1-norm to regularize the sparsity for the task-specific structures. To solve the resulting optimization problem, we developed an efficient algorithm based on the ADMM framework and proximal operators. We have validated the efficacy of our method on three real EEG datasets for binary and multiclass classification. The experimental results have demonstrated that our method outperformed the state-of-the-art techniques.

Appendix
Proof

Proof of Theorem 1: Optimizing (10) is equivalent to solve the following problem:

$$\begin{aligned}\underset{\mathcal{W}^t,b^t,\boldsymbol{\xi}^t}{\arg\min}\quad & \sum_{i=1}^{m_t}\xi_i^t+\frac{\lambda}{2}\mathcal{S}\big(\mathcal{W}^t\big)+<\mathcal{Y}^t,\mathcal{W}^t>+\\ & \frac{\gamma}{2}||\mathcal{W}^t-\sum_{k=1}^{K}\mathcal{P}^{(k)}-\mathcal{Q}^t||_F^2\\ \text{s.t.}\quad & y_i^t\big(<\mathcal{W}^t,\mathcal{X}_i^t>\big)\geq 1-\xi_i^t\\ & \xi_i^t\geq 0, i=1,\ldots,m_t\end{aligned}\tag{18}$$

where ξ_i^t is the slack variable. To solve the above constrained problem in (18), we formulate the following Lagarange function:

$$\begin{aligned}L\big(\mathcal{W}^t,b^t,\boldsymbol{\xi}^t,\boldsymbol{\alpha},\boldsymbol{\delta}\big)=&\sum_{i=1}^{m_t}\xi_i^t+\frac{\lambda}{2}\mathcal{S}\big(\mathcal{W}^t\big)+<\mathcal{Y}^t,\mathcal{W}^t>\\ &+\frac{\gamma}{2}||\mathcal{W}^t-\sum_{k=1}^{K}\mathcal{P}^{(k)}-\mathcal{Q}^t||_F^2-\sum_{i=1}^{m_t}\delta_i\xi_i^t\\ &-\sum_{i=1}^{m_t}\alpha_i\big\{y_i^t\big(<\mathcal{W}^t,\mathcal{X}_i^t>+b_i^t\big)-1+\xi_i^t\big\}.\end{aligned}\tag{19}$$

Setting the derivative of L with respect to $\mathcal{W}^t$, $\boldsymbol{\xi}^t$, and b^t to be 0, respectively, we have

$$\begin{aligned}\hat{\mathcal{W}}^t&=\frac{\gamma\Big(\sum_{k=1}^{K}\mathcal{P}^{(k)}+\mathcal{Q}^t\Big)-\mathcal{Y}^t+\sum_{i=1}^{m_t}\alpha_i^t y_i^t\mathcal{X}_i^t}{\lambda D^t+\gamma}\\ \delta_i&=1-\alpha_i\geq 0, i=1,\ldots,m_t\\ \sum_{i=1}^{m_t}\alpha_i y_i^t&=0\end{aligned}\tag{20}$$

where $D^t=\sum_{c=1}^{2}\sum_{j=1}^{m_t^c}||\mathcal{X}_j^t-\bar{\mathcal{X}}_c^t||_F^2-\sum_{c=1}^{2}m_t^c||\bar{\mathcal{X}}_c^t-\bar{\mathcal{X}}^t||_F^2$.

Substituting (20) into (19), we obtain the following dual function

$$\begin{aligned}&L\big(\mathcal{W}^t,b^t,\boldsymbol{\xi}^t,\boldsymbol{\alpha},\boldsymbol{\delta}\big)\\ &=\sum_{i=1}^{m_t}\left(1-\frac{<\gamma\Big(\sum_{k=1}^{K}\mathcal{P}^{(k)}+\mathcal{Q}^t\Big)-\mathcal{Y}^t,y_i^t\mathcal{X}_i^t>}{\lambda D^t+\gamma}\right)\alpha_i\\ &\quad-\frac{1}{2(\lambda D^t+\gamma)}\sum_{i=1}^{m_t}\sum_{j=1}^{m_t}\alpha_i\alpha_j y_i^t y_j^t<\mathcal{X}_i^t,\mathcal{X}_j^t>+\ \text{Const}\end{aligned}\tag{21}$$

where Const is a constant variable. Thus, the solution of (10) is equivalent to that of (12). Once the optimal $\boldsymbol{\alpha}$ is obtained, we can calculate optimal $\mathcal{W}^t$ by (20). The KKT conditions also provide

$$\begin{aligned}\alpha_i\big\{y_i^t\big(<\mathcal{W}^t,\mathcal{X}_i^t>+b_i^t\big)-1+\xi_i^t\big\}&=0\\ \delta_i\xi_i^t&=0.\end{aligned}\tag{22}$$

Then, the optimal b can be calculated by

$$\hat{b}^t=y_i^t-<\mathcal{W}^t,\mathcal{X}_i^t>.\tag{23}$$

To obtain a more stable result, we calculate the average solution by

$$\hat{b}^t=\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\big(y_i^t-<\mathcal{W}^t,\mathcal{X}_i^t>\big).\tag{24}$$

■

Proof of Theorem 2: For the sake of easy description, we denote (14) by $\arg\min\, g(\mathbf{P}_{(k)}^{(k)})$ and $(1/T)\sum_{t=1}^{T}(\mathbf{W}_{(k)}^t-\mathbf{Q}_{(k)}^t+[\mathbf{Y}_{(k)}^t/\gamma])-\sum_{j\neq k}^{K}\mathbf{P}_{(k)}^{(j)}$ by $\mathbf{H}$.

If (15) is satisfied, we need to prove $\mathbf{0}\in\partial g(\hat{\mathbf{P}}_{(k)}^{(k)})$.

We decompose $\mathbf{H}$ into two components as

$$\mathbf{H}=\mathbf{U}_0\boldsymbol{\Sigma}_0\mathbf{V}_0^T+\mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^T\tag{25}$$

where $\mathbf{\Sigma}_0$ is the diagonal matrix whose diagonal entries are greater than $(\tau_k/\gamma T)$; $\mathbf{U}_0$ and $\mathbf{V}_0$ are matrices of the corresponding left and right singular vectors; $\mathbf{\Sigma}_1$, $\mathbf{U}_1$, and $\mathbf{V}_1$ are the remaining parts of SVD whose singular values are less than or equal to $(\tau_k/\gamma T)$.

Thus,

$$\hat{\mathbf{P}}_{(k)}^{(k)} = \text{prox}_{\frac{\tau_k}{\gamma T}||\cdot||_*}(\mathbf{H}) = \mathbf{U}_0\left(\mathbf{\Sigma}_0 - \frac{\tau_k}{\gamma T}\right)\mathbf{V}_0^T. \tag{26}$$

Substituting (25) and (26) into the derivative of $g(\mathbf{P}_{(k)}^{(k)})$, we have

$$\begin{aligned}
\partial g\left(\hat{\mathbf{P}}_{(k)}^{(k)}\right) &= \tau_k \partial \left\|\hat{\mathbf{P}}_{(k)}^{(k)}\right\|_* - \sum_{t=1}^{T}\mathbf{Y}_{(k)}^t \\
&\quad + \gamma\sum_{t=1}^{T}\left(\hat{\mathbf{P}}_{(k)}^{(k)} + \sum_{j\neq k}^{K}\mathbf{P}_{(k)}^{(j)} + \mathbf{Q}_{(k)}^t - \mathbf{W}_{(k)}^t\right) \\
&= \tau_k \partial \left\|\hat{\mathbf{P}}_{(k)}^{(k)}\right\|_* - \gamma T\mathbf{H} + \gamma T\hat{\mathbf{P}}_{(k)}^{(k)} \\
&= \tau_k \partial \left\|\hat{\mathbf{P}}_{(k)}^{(k)}\right\|_* - \gamma T\left(\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^T + \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T\right) \\
&\quad + \gamma T\mathbf{U}_0\left(\mathbf{\Sigma}_0 - \frac{\tau_k}{\gamma T}\mathbf{I}\right)\mathbf{V}_0 \\
&= \tau_k \partial \left\|\hat{\mathbf{P}}_{(k)}^{(k)}\right\|_* - \tau_k\mathbf{U}_0\mathbf{V}_0 - \gamma T\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1
\end{aligned} \tag{27}$$

where $\partial\|\hat{\mathbf{P}}_{(k)}^{(k)}\|_*$ is the set of subgradients of the nuclear norm. Let $\mathbf{P}_{(k)}^{(k)}$ be an arbitrary matrix and denote its SVD as $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. It follows from [50]:

$$\partial\left\|\hat{\mathbf{P}}_{(k)}^{(k)}\right\|_* = \left\{\mathbf{U}\mathbf{V}^T + \mathbf{Z} : \mathbf{U}^T\mathbf{Z} = \mathbf{0}, \mathbf{Z}\mathbf{V} = \mathbf{0}, ||\mathbf{Z}||_F \leq 1\right\}. \tag{28}$$

In this regard, we define $\mathbf{U} = \mathbf{U}_0$, $\mathbf{V} = \mathbf{V}_0$, and $\mathbf{Z} = (\gamma T/\tau_k)\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1$. It is obvious to verify that $\mathbf{U}_0^T\mathbf{Z} = \mathbf{0}$, $\mathbf{Z}\mathbf{V}_0 = \mathbf{0}$, and $||\mathbf{Z}||_F \leq 1$. Thus, we have $\mathbf{0} \in \partial g(\hat{\mathbf{P}}_{(k)}^{(k)})$. ■

Proof of Theorem 3: To prove Theorem 3, we first reformulate the optimization problem in (16) as

$$\arg\min_{\mathcal{Q}^t}\frac{\gamma}{2}\left\|\mathcal{Q}^t + \sum_{k=1}^{K}\mathcal{P}^{(k)} - \mathcal{W}^t - \frac{\mathcal{Y}^t}{\gamma}\right\| + \beta||\mathcal{Q}^t||_1 + \text{Const}' \tag{29}$$

where Const′ is a constant unrelated to the solution. Based on the proximal operator in [33], the solution of optimization problem in (29) can be easily derived with (17). ■

REFERENCES

[1] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 1017–1027, Apr. 2017.

[2] D. Iacoviello, A. Petracca, M. Spezialetti, and G. Placidi, "A classification algorithm for electroencephalography signals by self-induced emotional stimuli," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3171–3180, Dec. 2016.

[3] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1444–1452.

[4] Y. Luo, D. Tao, B. Geng, C. Xu, and S. J. Maybank, "Manifold regularized multitask learning for semi-supervised multilabel image classification," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 523–536, Feb. 2013.

[5] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.

[6] T. Liu, D. Tao, M. Song, and S. J. Maybank, "Algorithm-dependent generalization bounds for multi-task learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 227–241, Feb. 2017.

[7] X. Tian, Y. Li, T. Liu, X. Wang, and D. Tao, "Eigenfunction-based multitask learning in a reproducing kernel Hilbert space," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1818–1830, Jun. 2019.

[8] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2853–2884, 2016.

[9] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1415–1428, Aug. 2009.

[10] H. Lee, Y.-D. Kim, A. Cichocki, and S. Choi, "Nonnegative tensor factorization for continuous EEG classification," *Int. J. Neural Syst.*, vol. 17, no. 4, pp. 305–317, 2007.

[11] X. Gu, F.-L. Chung, H. Ishibuchi, and S. Wang, "Multitask coupled logistic regression and its fast implementation for large multitask datasets," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1953–1966, Sep. 2015.

[12] H. Zeng and A. Song, "Optimizing single-trial EEG classification by stationary matrix logistic regression in brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2301–2313, Nov. 2016.

[13] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, "Seperability of four-class motor imagery data using independent components analysis," *J. Neural Eng.*, vol. 3, no. 3, p. 208, 2006.

[14] Y. Yang, Y. Feng, and J. A. K. Suykens, "Robust low-rank tensor recovery with regularized redescending M-estimator," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1933–1946, Sep. 2016.

[15] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured Schatten norm regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1331–1339.

[16] K. Wimalawarne, M. Sugiyama, and R. Tomioka, "Multitask learning meets tensor factorization: Task imputation via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2825–2833.

[17] Q. Zheng, F. Zhu, and P.-A. Heng, "Robust support matrix machine for single trial EEG classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 551–562, Mar. 2018.

[18] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Trans. Cybern.*, vol. 45, no. 6, pp. 1209–1221, Jun. 2015.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[20] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[21] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[22] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2004, pp. 109–117.

[23] Y. Li, X. Tian, T. Liu, and D. Tao, "Multi-task model and feature joint learning," in *Proc. IJCAI*, 2015, pp. 3643–3649.

[24] A. Kumar and H. Daume, III, "Learning task grouping and overlap in multi-task learning," *arXiv preprint arXiv:1206.6417*, 2012.

[25] Y. Luo, Y. Wen, and D. Tao, "Heterogeneous multitask metric learning across multiple domains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4051–4064, Sep. 2018.

[26] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, 2011.

[27] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[28] R. Tomioka and K. Aihara, "Classifying matrices with a spectral regularization," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 895–902.

[29] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 208–220, Jan. 2013.

[30] K. Wimalawarne, R. Tomioka, and M. Sugiyama, "Theoretical and experimental analyses of tensor-based regression and classification," *Neural Comput.*, vol. 28, no. 4, pp. 686–715, Apr. 2016.

[31] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[32] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[33] N. Parikh and S. P. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, 2014.

[34] M. Alamgir, M. Grosse-Wentrup, and Y. Altun, "Multitask learning for brain-computer interfaces," in *Proc. 13th Int. Conf. Artif. Intell. Stat.*, 2010, pp. 17–24.

[35] G. Panagopoulos, "Multi-task learning for commercial brain computer interfaces," in *Proc. IEEE 17th Int. Conf. Bioinformat. Bioeng. (BIBE)*, 2017, pp. 86–93.

[36] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Bilinear classifiers for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1482–1490.

[37] L. Luo, Y. Xie, Z. Zhang, and W.-J. Li, "Support matrix machines," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 938–947.

[38] H. Zhou and L. Li, "Regularized matrix regression," *J. Roy. Stat. Soc. B (Stat. Methodol.)*, vol. 76, no. 2, pp. 463–483, 2014.

[39] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Aggregation of sparse linear discriminant analyses for event-related potential classification in brain–computer interface," *Int. J. Neural Syst.*, vol. 24, no. 1, 2014, Art. no. 1450003.

[40] Y. Zhang, G. Zhou, J. Jin, Q. Zhao, X. Wang, and A. Cichocki, "Sparse Bayesian classification of EEG for brain–computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2256–2267, Nov. 2016.

[41] Q. Zheng, F. Zhu, J. Qin, B. Chen, and P.-A. Heng, "Sparse support matrix machine," *Pattern Recognit.*, vol. 76, pp. 715–726, Apr. 2018.

[42] X. Guo, Q. Yao, and J. T.-Y. Kwok, "Efficient sparse low-rank tensor completion using the Frank–Wolfe algorithm," in *Proc. AAAI*, 2017, pp. 1948–1954.

[43] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Muller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, Jun. 2004.

[44] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain–computer communication: Motivation, aim, and impact of exploring a virtual apartment," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 473–482, Dec. 2007.

[45] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-Task Learning Via Structural Regularization*, Arizona State Univ., Tempe, AZ, USA, 2011. [Online]. Available: http://www.MALSAR.org

[46] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 41–48.

[47] P. Gong, J. Ye, and C. Zhang, "Robust multi-task feature learning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2012, pp. 895–903.

[48] S. García, A. Fernández, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power," *Inf. Sci.*, vol. 180, no. 10, pp. 2044–2064, 2010.

[49] Y. Wen, H. Xu, and J. Yang, "A heuristic-based hybrid genetic-variable neighborhood search algorithm for task scheduling in heterogeneous multiprocessor system," *Inf. Sci.*, vol. 181, no. 3, pp. 567–581, 2011.

[50] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, 2009.

Qingqing Zheng (M'14) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong.

She is currently a Researcher with Tencent Youtu Lab, Shenzhen, China. Her current research interests include machine learning, computer vision, and brain computer interfaces.

Yi Wang received the Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2017.

He is currently serving as an Assistant Professor with the School of Biomedical Engineering, Shenzhen University, Shenzhen, China. He values practical techniques that can be transferred to clinically applicable systems. His current research interests include medical image computing (especially, in medical image registration), medical imaging, computer vision, image processing, and machine learning.

Pheng Ann Heng (SM'06) received the Ph.D. degree in computer science from Indiana University, Indianapolis, IN, USA.

He is currently a Professor with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. He is also the Director of the Research Center for Human–Computer Interaction, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His current research interests include visual reality, visualization, medical imaging, human–computer interfaces, rendering and modeling, interactive graphics, and animation.