Traffic Sign Recognition via Multi-Modal Tree-Structure Embedded Multi-Task Learning

Xiao Lu, Yaonan Wang, Xuanyu Zhou, Zhenjun Zhang, and Zhigang Ling

***Abstract*—Traffic sign recognition is a rather challenging task for intelligent transportation systems since signs in different subsets, e.g., speed limit signs, prohibition signs, and mandatory signs, are very different from each other in color or shape, whereas they share some similarities to the ones in the same subset. Therefore, it is important to integrate different modalities of visual features, such as color and shape, and select discriminative features for better sign description; in addition, it benefits to explore the correlations between the classes of traffic signs to learn the classifiers jointly to improve the generalization performance. In this paper, we propose Multi-Modal tree-structure embedded Multi-Task Learning called M^2-tMTL to select discriminative visual features both between and within modalities, as well as the correlated features shared by similar classification tasks. Our method simultaneously introduces two structured sparsity-induced norms into a least squares regression. One of the norms can be used not only to select modality of features but also to conduct within-modality feature selection. Moreover, the hierarchical correlations among the classification tasks are well represented by a tree structure, and therefore, the tree-structure sparsity-induced norm is used for learning the regression coefficients jointly to boost the performance of multi-class traffic sign recognition. Alternating direction method of multipliers (ADMM) is used to efficiently solve the proposed model with guaranteed convergence. Extensive experiments on public benchmark data sets demonstrate that the proposed algorithm leads to a quite interpretable model, and it has better or competitive performance with several state-of-the-art methods but with less computational and memory cost.**

***Index Terms*—traffic sign recognition, multi-modal feature learning, tree-structure embedded multi-task learning, structured sparsity-induced norm, ADMM.**

I. Introduction

TRAFFIC sign recognition systems have been given increasing importance as their applications in autonomous intelligent vehicle and intelligent assistance driving systems in recent years [1]–[3]. However, to recognize the traffic signs effectively may confront with many difficulties, e.g., unpredictable complex scenes, changing lighting conditions, variations of the angle of view, partially occlusion, camera noise, low resolution, motion blur and so on. In addition to the above external factors, traffic sign classification is a rather challenging pattern recognition problem, as there are so many categories of signs, and some sets of them share a high level of similarity. In the German Traffic Sign Recognition Benchmark (GTSRB), there are 43 different classes of signs to recognize and they can be split into subsets according to their different warning or guiding meanings, such as speed limit signs and mandatory signs, etc. The signs in the same subset are very similar to each other in color and shape. Moreover, the traffic sign classification is a multi-class problem with rather unbalanced class frequencies, as some signs are rare in practice, in the training set of GTSRB, the least training number is 210 versus the most one 2250.

Traffic signs are designed to be attractive to human drivers, the ones with the same general meaning share the same color and shape, while they are observably different in color or shape between the subsets with different meanings. Humans usually determine the meanings of a test sign by its color and shape, e.g., "It is a speed limit sign," and then recognize it according to the icons and text, which conforms to the cognitive science of human brains. The result of best individual human performance in the GTSRB competition [4] shows that there is rare miss-classification between subsets with different meanings.

Considering that for high-dimensional heterogeneous input features such as color, texture and shape, different kinds (modalities) of features have different intrinsic discriminative power, it is rather important to design an algorithm that can select a few discriminative features for better recognition. Moreover, as signs within the same subset are more similar than others in different subsets, correspondingly, the classifiers for the signs in the same subset should be more correlated than others in different subsets, thus, it is also important to model the correlations of the classification tasks to improve the generalization performances.

Multi-modal/multi-view feature learning has been a hot research topic to enhance the visual understanding. Multiple kernel learning (MKL) methods [5], [6] have been widely studied to integrate heterogeneous features and select multi-modal features. However, these methods evaluate the importance of a feature-modality by a simple weight, and all features in that modality are equally weighted, which often causes low performance [7]. In recent research, sparse regularization [8] has been widely investigated and applied in machine learning studies. The sparse feature selection can be achieved by imposing non-smooth norms as regularizer in the optimization problems,

Manuscript received June 7, 2015; revised December 3, 2015 and May 3, 2016; accepted July 21, 2016. Date of publication August 24, 2016; date of current version March 27, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 751202102, 61471166 and in part by Hunan Provincial Natural Science Foundation under Grant 14JJ2052. The Associate Editor for this paper was S. S. Nedevschi. *(Corresponding author: Xiao Lu.)*

X. Lu is with the College of Engineering and Design, Hunan Normal University, Changsha 410006, China (e-mail: xlu_hnu@163.com).

Y. Wang, Z. Zhang, and Z. Ling are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: yaonan@hnu.edu.cn; zhenjun@hnu.edu.cn; zgling_hunan@126.com).

X. Zhou is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: zhouxuanyu@whu.edu.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/TITS.2016.2598356

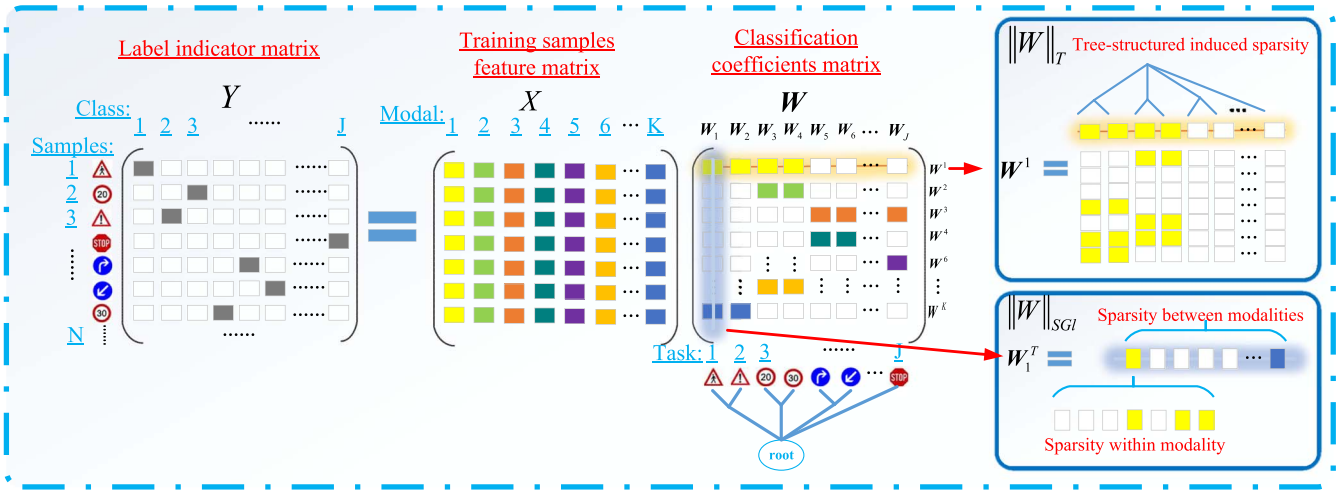


Fig. 1. Algorithm overview of the proposed framework. Different colors of rectangles in data matrix $\mathbf{X}$ indicate different types of heterogeneous visual features. Blank rectangles indicate zero values in coefficient matrix $\mathbf{W}$ and label matrix $\mathbf{Y}$. In coefficient matrix $\mathbf{W}$, rectangles with colors according to data matrix $\mathbf{X}$ indicate the nonzero coefficients, and each rectangle indicates a coefficient block of a feature modality. Two kinds of structured sparsity-induced norms are imposed on the regression coefficient matrix $\mathbf{W}$, in which $\|\mathbf{W}\|_T$ is the tree-structure sparsity-induced norm; it is imposed to capture the correlations of each classifiers, and $\|\mathbf{W}\|_{SGl}$ is imposed to select discriminative features within and between modalities simultaneously. $\mathbf{W}^1$ and $\mathbf{W}_1$ are the coefficients of the first feature modality and the first classification task, respectively.

e.g., the ℓ_1-norm of the popular lasso [9]. Moreover, the structured sparsity regularizer [10] can capture the structures existing in features, e.g., the group lasso [11] regularizer used to learn the feature importance from group-wise is proved to be useful in discovering the underlying patterns.

Multi-task learning (MTL) has been attractive in machine learning as it is a statistical learning framework which seeks to learn several models in a joint manner [12], [13]. The idea behind this paradigm is that, when the tasks to be learned are similar enough or are correlated in some sense, it may be advantageous to take into account these correlations between tasks. Several works have provided empirical evidence on the benefit of such a framework. Multi-task feature learning [12], [14] generalize the ℓ_1-norm for single-task case to learn a few features common across tasks by leveraging a mixed norm which both couples the tasks and enforces sparsity.

Inspired by the above work, in this paper, we propose to model the hierarchical correlations between different sign classes as a tree structure and deal with the multi-class classification problem as a multi-task regression problem. The classes belonging to the same tree node are more similar to each other, and the similarity between two nodes is related to the depth of the tree node. Thus, instead of learning each classifier independently, all the correlated classifiers can be learnt jointly by extracting and utilizing appropriate shared information across tasks. By imposing a tree guided structured sparsity regularizer, similar classes are encouraged to select similar features to boost the prediction performance. Furthermore, to leverage the different descriptions of multi-modal features, we implement sparse group lasso regularizer to select the sparse discriminative features between modalities and within modalities simultaneously. Specifically, the group ℓ_1-norm [11] is used to enforce the sparsity between different modalities, and the ℓ_1-norm [9] imposes sparsity within each modality. The framework of our proposed method for learning the classification coefficient matrix is presented in Fig. 1. Lastly, ADMM [15] is used to solve the convex but highly non-smooth optimization problem with guaranteed convergence. Experiments on benchmark data sets demonstrate that the proposed algorithm exhibits better or competitive performance compared to some state-of-the-art methods, furthermore, it leads to a quite interpretable model.

The contributions of this paper are summarized as follows:

1) To the best of our knowledge, this is the first work to utilize the tree structure MTL method to deal with traffic sign recognition problem, which models the hierarchical correlations of the classification tasks and learns classifiers across tasks in a joint manner.
2) Multi-modal features, i.e., color histogram features, histogram of orientation gradient (HOG) features [16] and Haar features [17] with different types and sizes, are used to complement with each other for better describing a sign, and a mixed norm is introduced as a regularizer to select a small portion of discriminative features.
3) ADMM is implemented to solve this high dimensional, convex but highly non-smooth optimization problem with guaranteed convergence.

The rest of the paper is organized as follows: we first introduce some works related to traffic sign recognition and some preliminaries of structured sparsity-induced norm in Section II. Section III presents the proposed M^2-tMTL method for traffic sign recognition, and the ADMM method for solving the optimization problem is presented in Section IV. The details of the benchmark data sets used in our experiment are introduced in Section V. Section VI presents the experimental results and finally Section VII draws the conclusion of this paper.

II. Related Work

In this section, we first review some related traffic sign recognition methods. Then, we introduce the definition of structured

sparsity-induced norm, based on which we will propose our M^2-tMTL method.

A. Related Traffic Sign Recognition Methods

Many state-of-the-art machine learning methods were employed in the competition of GTSRB [4], which was held by the International Joint Conference on Neural Networks 2011 (IJCNN 2011). The neural-network based methods achieved perfect performance. Both the multiscale convolutional neural networks (CNN) proposed by Sermanet *et al.* [18] and the multicolumn deep neural networks (MCDNN) proposed by Ciresan *et al.* [19] were claimed to outperform the human test person, although they are high in computational cost. Linear Discriminative Analysis (LDA) served as the baseline method in the competition of GTSRB for its simplicity [4], however, it gives surprisingly good results in practice. In addition to these two kinds of methods, support vector machine (SVM), ensemble learning methods, and subspace learning methods were implemented and achieved high performance.

The high classification accuracy achieved by multi-scale CNN benefits from the multi-layer feed-forward structure which can learn task-specific invariant features in a hierarchical manner, and the implementation of the combination of a rectified sigmoid followed by subtractive and divisive local normalization further improves the performance [18]. Recently, Jin *et al.* [20] proposed the hinge loss trained CNN method, and achieved the best accuracy of 99.65% over the best record 99.46% kept by Dan Ciresan. These methods are high in computational cost, as in the multi-scale CNN, all feature extraction stages are fed into the classifier, and a high-performance GPU is usually needed for efficiency. Besides the neural network based methods, some other methods also exhibit good performance. In [21], a kernel rule is developed for road sign classification using the Laplace probability density, and in [22] Saturnino *et al.* perform the content recognition by employing Gaussian kernel SVMs. Kernel based methods learn the classifier by projecting the features into a high dimensional space, however, the kernel functions and parameters are usually hard to tune and determine. On the contrary, subspace learning methods try to learn a more compact and representative feature space by reducing the dimensionality. Sparse representation classification (SRC) [23] is also used for traffic sign recognition as its successful implementation on face recognition. Sparse-representation-based linear projections (SRLPs) [24] and sparse-representation-based graph embedding (SRGE) [25] are performed on GTSRB. Although they achieved high performance, they suffer from the intrinsic flaw of SRC, e.g., large number of training samples and sensitive to variations of view angle.

Instead of using machine learning algorithms for recognition directly, some researchers tried to design features fit for traffic signs. In [26], the corner masks are designed for triangular, rectangular and circular respectively. However, the corner detector is sensitive to noise. Saturnino *et al.* [22] define the distance to border (DtB), which is extracted as the feature for shape classification by support vector machines. Khan *et al.* [27] plot the distances of the boundary points from the centroid to find the number of sides and it is then combined with the Peri2Area(Perimeter^2/Area) to classify the shape. Similarly, in [28], the distance from the mass center to the boundary of the object is defined as a function of the angle, then it is used as the mainly feature for shape classification. All these features are contour-based, which are susceptible to occlusion and deformation. In [27], the shape analysis and reassortment technique for the rectangular shapes are proposed to allow the recognition block to focus the search on a smaller range of possible signs. However, only the tilted effect is considered. Both in [28] and [29], the shape rectification regulations are designed for circular, triangular and rectangular respectively, which means that the rectification must be based on the correct shape classification. These shape features should be followed by some algorithms for further classification, the framework works as a pipeline, and hence the different features cannot complement with each other.

B. Structured Sparsity-Induced Norm

In this section, we first provide notations used in the rest of this paper. Then we introduce the definition of the structured sparsity-induced norm. Finally, we briefly introduce some instances of structured sparsity-induced norm for multi-modal feature learning and multi-task learning.

1) Notations: In this paper, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. For matrix $\mathbf{A}$, its ith row and jth column are denoted by $\mathbf{A}_{(i,:)}$ and $\mathbf{A}_{(:,j)}$, respectively. Assuming we are given N training samples from J classes, with n_j samples in the jth class ($N = \sum_{j=1}^{J} n_j$), let $\mathbf{X}^T = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathcal{R}^{D\times N}$ be the data matrix, where $\mathbf{x}_i = [(\mathbf{x}_i^1)^T, \ldots, (\mathbf{x}_i^K)^T]^T \in \mathcal{R}^D$ is the feature vector for the ith sample, including all features from a total of K modalities and the kth modality has d_k features ($D = \sum_{k=1}^{K} d_k$). Matrix $\mathbf{Y} \in \{0,1\}^{N\times J}$ represents the label indicator matrix of samples: $y_{ij} = 1$ if the ith sample belongs to the jth class and $y_{ij} = 0$ if otherwise. We directly learn a $D \times J$ coefficient matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_J] \in \mathcal{R}^{D\times J}$, and $\mathbf{w}_j = [(\mathbf{w}_j^1)^T, \ldots, (\mathbf{w}_j^K)^T]^T$ where $\mathbf{w}_p^q \in \mathcal{R}^{d_q}$ indicates the coefficients of all features from the qth modality of the pth class. Given a D-dimensional vector α, let α_i be its ith element, the ℓ_1-norm and ℓ_2-norm are defined as $\|\alpha\|_1 = \sum_{i=1}^{D} |\alpha_i|$ and $\|\alpha\|_2 = \sqrt{\sum_{i=1}^{D} \alpha_i^2}$, respectively. Throughout this paper, we assume that the label matrix $\mathbf{Y}$ is centered and feature matrix $\mathbf{X}$ is centered and standardized.

2) Structured Sparsity-Induced Norm: Jenatton *et al.* proposed a general definition of the structured sparsity-induced norm in [10], based on which many sparsity penalties such as lasso [9], group lasso [11], and tree-guided group lasso [30], [31] can be instantiated.

Given a D-dimensional input vector $\mathbf{x}$, let us assume that the set of groups of inputs $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ is defined as a subset of the power set of $\{1, \ldots, D\}$, and w_g is a positive scalar weight indexed by group g, the structured sparsity-induced norm $\Omega(\mathbf{x})$ is defined as

$$\Omega(\mathbf{x}) \equiv \sum_{g\in\mathcal{G}} w_g \|\mathbf{x}_g\|_q \tag{1}$$

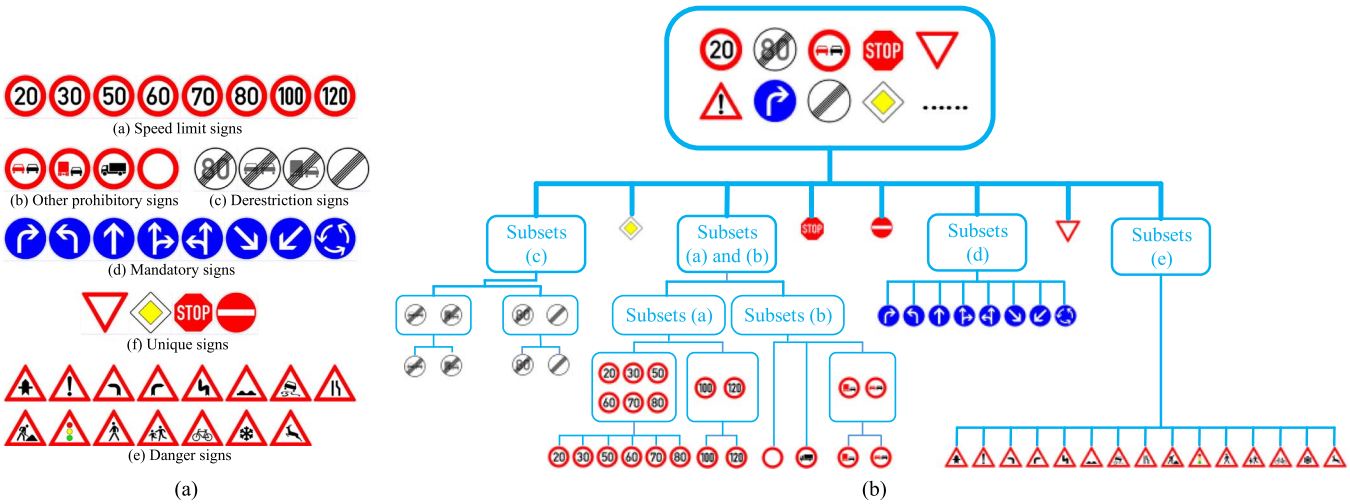


Fig. 2. The 43 categories of signs in GTSRB are split into six subsets shown in (a); all the classification tasks are embedded in a tree structure $\mathcal{T}_c$ shown in (b), where each leaf node represents a classification task, and each internal node represents a grouping of the tasks located at the leaves of the subtree. This tree structure is built according to the similarity of the signs; the signs in the same subset share the same shape and color. Moreover, some signs in a subset are more similar than others, e.g., speed limit signs 100 and 120 are more similar than others, the set with unique signs contains elements that are neither similar to each other in the subset nor similar to other signs, and they are leaf nodes of the root node. (a) Subsets of traffic signs. (b) Tree structure $\mathcal{T}_c$ embedded in classification tasks.

where $q \in (1, \infty]$, $\mathbf{x}_g \in \mathcal{R}^{|g|}$ is the subvector of $\mathbf{x}$ for the input index in group g. This norm is usually referred to as a mixed ℓ_1/ℓ_q-norm, and in practice, popular choices for $q = 2$. Specifically, if $\mathcal{G}$ is the set of singleton, i.e., $\mathcal{G} = \{\{1\}, \{2\}, \ldots, \{D\}\}$ and all elements are equally weighted, $\Omega(\mathbf{x})$ is instantiated to be an ℓ_1-norm of vector $\mathbf{x}$.

3) Disjoint Groups of Variables Selection: The most natural form of structured sparsity is arguably disjoint group sparsity, with the prior knowledge that pre-specified disjoint blocks of variables should be selected or ignored simultaneously. In this case, $\mathcal{G}$ is divided into K disjoint groups and the group sparsity norm is imposed to encourage the kth group of features to be selected or ignored. In the context of least-squares regression, this regularization is known as the group lasso [11]. If we use a matrix $\mathbf{X}_k \in \mathcal{R}^{N \times d_k}$ to represent the features of training samples from the kth modality, for the jth class, the group lasso can be formulated as:

$$\min_{\mathbf{w}_j} \left\| \mathbf{Y}_{(:,j)} - \sum_{k=1}^{K} \mathbf{X}_k \mathbf{w}_j^k \right\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{w_k} \left\| \mathbf{w}_j^k \right\|_2 \quad (2)$$

where $\lambda > 0$ is the regularization parameter.

4) Multi-Task Feature Learning: Applications of the structured sparsity regularization scheme arise also in the context of multi-task learning. The mixed ℓ_1/ℓ_2-norm is imposed on the coefficient matrix $\mathbf{W}$ to account for features shared across tasks [14]:

$$\min_{\mathbf{W} \in \mathcal{R}^{N \times J}} \ell(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \rho \sum_{i=1}^{D} \|\mathbf{W}_{(i,:)}\|_2 \quad (3)$$

where $\rho > 0$ is the regularization parameter, and $\ell(\cdot)$ is a smooth convex loss function such as the least square loss and the logistic loss. The coefficients corresponding to the ith feature are grouped together via the ℓ_2-norm of $\mathbf{W}_{(i,:)}$. Thus, the ℓ_1/ℓ_2-norm regularization tends to select features based on the strength of the input variables of the J tasks jointly rather than on the strength of individual input variables as in the case of single task learning.

III. M^2-tMTL for Traffic Sign Recognition

According to the meanings of the signs, the samples from 43 classes in GTSRB are split into six subsets shown in Fig. 2(a). To explore the correlations between the 43 classification tasks for learning a few features common across the similar classes, all the tasks can be equipped with a hierarchical tree structure $\mathcal{T}_c$ with the set of vertices $\mathbf{V}$ of size $|\mathbf{V}|$. As shown in Fig. 2(b), each of the J leaf nodes of $\mathcal{T}_c$ represents a classification task, it is associated with an output variable, and the internal nodes represent groupings of the tasks located at the leaves of the subtree rooted at the given internal node. The tasks belonging to the same node are similar to each other, e.g., the tasks for classification speed limit signs belong to the same node, they are more similar to each other than the tasks for classification mandatory signs. Moreover, the similarity between two nodes is related to the depth of the tree node, e.g., within the speed limit signs node, sign 100 is more similar to sign 120 compared to other speed limit signs.

In this case, a subset of highly correlated tasks may share a common set of relevant features, whereas weakly related tasks are less likely to be affected by the same features, therefore, it is not appropriate to impose the ℓ_1/ℓ_2-norm on $\mathbf{W}$. Given the tree $\mathcal{T}_c$ over the tasks, we get a tree-structured set of groups $\mathcal{G}_T = \{v_1, \ldots, v_{|V|}\}$, where v_i $(i = 1, \ldots, |V|)$ represents a group of tasks corresponding to node $v_i \in V$. For example, in Fig. 2(b), we have $v_1 = \{\text{signs in the derestriction subset}\}$ and $v_8 = \{\text{signs in the danger subset}\}$.

Our M^2-tMTL method for learning $\mathbf{W}$ is defined as the following optimization problem:

$$\min_{\mathbf{W}} \ell(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \lambda_1 \|\mathbf{W}\|_T + \lambda_2 \|\mathbf{W}\|_{\text{SGl}} \quad (4)$$

where $\lambda_1, \lambda_2 > 0$ are regularization parameters, $\ell(\mathbf{X}, \mathbf{Y}, \mathbf{W})$ is the loss function, and we use the least-square loss in this paper.

The tree-structured sparsity-induced norm [10] $\|\mathbf{W}\|_T$ is imposed to recover the common features of relevant classification tasks, it is defined as:

$$\|\mathbf{W}\|_T = \sum_{d=1}^{D} \sum_{v \in \mathcal{G}_T} w_v \left\|\mathbf{W}_v^d\right\|_2 \tag{5}$$

where $\mathbf{W}^d = \mathbf{W}_{(d,:)}$ for ease of notation. Apparently, the tree-structured sparsity-induced norm $\sum_{v \in \mathcal{G}_T} w_v \|\mathbf{W}_v^d\|$ is a special example of the definition $\Omega(\mathbf{W}^d)$ in (1) and it generalizes the ℓ_1/ℓ_2-norm in (3) for tree-structure embedded multi-task feature learning. The weight w_v in (5) is defined as:

$$w_v = \begin{cases} g_v \prod_{m \in \text{Ancestor(v)}} s_m, & v \text{ is an internal node,} \\ \prod_{m \in \text{Ancestor(v)}} s_m, & v \text{ is a leaf node.} \end{cases} \tag{6}$$

where the two quantities s_v and g_v are associated with the internal node v of the tree $\mathcal{T}_c$ with condition $s_v + g_v = 1$ satisfied. Furthermore, s_v represents the weight for selecting the labels associated with each of the children of node v separately, and g_v represents the weight for selecting them jointly. More details of the weighting scheme can be referred to [31].

Furthermore, the **S**parse **G**roup **l**asso regularizer [32] $\|\mathbf{W}\|_{\text{SGl}}$ is imposed to select discriminative features for each classification task, which is defined as:

$$\|\mathbf{W}\|_{\text{SGl}} = \sum_{j=1}^{J} \sum_{k=1}^{K} w_k \|\mathbf{W}_j^k\|_2 + \alpha \sum_{j=1}^{J} \|\mathbf{W}_j\|_1 \tag{7}$$

where $\mathbf{W}_j = \mathbf{W}_{(:,j)}$ for ease of notation. It can be seen that for each classifier $\mathbf{W}_j$, (7) includes two norms. The first group sparsity norm enforces the sparsity between different modalities, i.e., if one modality of features are not discriminative for certain tasks, the objective in (2) will assign zeros to them for corresponding tasks; otherwise, their values are large. Thus, this group sparsity regularizer captures the global relationships between modalities. The latter ℓ_1-norm can induce sparsity on the whole coefficient vector $\mathbf{W}_j$ to select a few highly discriminative features within a modality. $\alpha > 0$ is regularization parameter to balance the effect of the two norms. Thus, $\mathbf{W}_{\text{SGl}}$ in (4) can be used to perform feature selection between and within modalities simultaneously.

IV. ADMM for Solving M^2-tMTL

It can be seen that both $\|\mathbf{W}\|_T$ in (5) and $\|\mathbf{W}\|_{\text{SGl}}$ in (7) are convex but non-smooth, and the loss function $\ell(\cdot)$ is convex and smooth. Moreover, note that the two norms are imposed on different directions of the coefficient matrix $\mathbf{W}$, i.e., $\|\mathbf{W}\|_T$ is imposed on the rows and $\|\mathbf{W}\|_{\text{SGl}}$ on the columns, which results in a convex but highly non-smooth optimization problem and thus could not be easily solved. In this section, we introduce ADMM for solving our M^2-tMTL problem. Although ADMM was originally proposed in the mid-1970 s by Glowinski *et al.* [33] and Gabay *et al.* [34], it has been widely known until recently with the development of large-scale distributed computing systems [15].

Introducing a matrix variable $\mathbf{Z}$, we rewrite the optimization problem (4) as the following linearly-constrained problem:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{Z}} \; & \ell(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \lambda_1 \|\mathbf{W}\|_T + \lambda_2 \|\mathbf{Z}\|_{\text{SGl}} \\ \text{s.t. } & \mathbf{W} = \mathbf{Z}. \end{aligned} \tag{8}$$

The augmented Lagrangian related to this problem is

$$\begin{aligned} \mathcal{L}_\mu(\mathbf{W}, \mathbf{Z}, \mathbf{B}) = \ell(\mathbf{X}, \mathbf{Y}, \mathbf{W}) + \lambda_1 \|\mathbf{W}\|_T + \lambda_2 \|\mathbf{Z}\|_{\text{SGl}} \\ + \langle \mathbf{B}, \mathbf{W} - \mathbf{Z} \rangle + \frac{\mu}{2} \|\mathbf{W} - \mathbf{Z}\|_F^2 \end{aligned} \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\mathbf{B}$ is a matrix of Lagrangian multipliers related to the equality constraint and μ is a parameter weighting the quadratic penalty. After rearranging the terms, one can show that the augmented Lagrangian is

$$\begin{aligned} \mathcal{L}_\mu(\mathbf{W}, \mathbf{Z}, \mathbf{A}) = \ell(\mathbf{W}) + \lambda_1 \|\mathbf{W}\|_T + \lambda_2 \|\mathbf{Z}\|_{\text{SGl}} \\ + \frac{\mu}{2} \|\mathbf{W} - \mathbf{Z} + \mathbf{A}\|_F^2 - \frac{\mu}{2} \|\mathbf{A}\|_F^2 \end{aligned} \tag{10}$$

where $\mathbf{A} = \mathbf{B}/\mu$ is the scaled dual variable. The ADMM that solves our original problem (8) looks for a saddle point of the augmented Lagrangian by solving alternatively at iteration t the following problems:

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} \mathcal{L}_\mu(\mathbf{W}, \mathbf{Z}_t, \mathbf{A}_t) \tag{11}$$

$$\mathbf{Z}_{t+1} = \arg\min_{\mathbf{Z}} \mathcal{L}_\mu(\mathbf{W}_{t+1}, \mathbf{Z}, \mathbf{A}_t) \tag{12}$$

$$\mathbf{A}_{t+1} = \mathbf{A}_t + \mathbf{W}_{t+1} - \mathbf{Z}_{t+1}. \tag{13}$$

All the challenges of the algorithm now resides essentially in the solution of these subproblems.

A. Solving Problem (11)

The optimization problem related to $\mathbf{W}$ can be restated as

$$\min_{\mathbf{W}} \; \ell(\mathbf{W}) + \frac{\mu}{2} \|\mathbf{W} - \mathbf{Z} + \mathbf{A}\|_F^2 + \lambda_1 \|\mathbf{W}\|_T . \tag{14}$$

Defining $f(\mathbf{W}) = \ell(\mathbf{W}) + (\mu/2)\|\mathbf{W} - \mathbf{Z} + \mathbf{A}\|_F^2$, $f(\mathbf{W})$ is convex and smooth, thus we can resort to the widely applied Accelerated Proximal Gradient (APG) method [35], [36] to optimize (14). In the seminal work [35], Nesterov considered the minimization problem with the objective function composed of a smooth convex part and a non-smooth convex part, furthermore, there exists a closed form minimizer of the sum of the non-smooth part with a quadratic auxiliary function. The APG algorithm proposed in [35] achieves $\mathcal{O}(1/t^2)$ convergence rate. Independently, Beck *et al.* [37] proposed the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) for solving linear inverse problem with the same convergence rate.

We adopt framework in [37] to provide a fast convergence rate algorithm for solving (14). Firstly, we define the *generalized gradient update* step as:

$$\begin{aligned} Q_L(\mathbf{W}, \mathbf{W}_t) = f(\mathbf{W}_t) + \langle \mathbf{W} - \mathbf{W}_t, \nabla f(\mathbf{W}_t) \rangle \\ + \frac{L}{2} \|\mathbf{W} - \mathbf{W}_t\|_F^2 + \lambda_1 \|\mathbf{W}\|_T \\ q_L(\mathbf{W}_t) = \arg\min_{\mathbf{W}} Q_L(\mathbf{W}, \mathbf{W}_t) \end{aligned} \tag{15}$$

where L is the Lipschitz constant of ∇f. Now, we focus on how to solve the generalized gradient update efficiently. After rewriting (15), we obtain

$$q_L(\mathbf{W}_t) = \arg\min_{\mathbf{W}} \frac{1}{2}\left\|\mathbf{W} - \left(\mathbf{W}_t - \frac{1}{L}\nabla f(\mathbf{W}_t)\right)\right\|_F^2 + \frac{\lambda_1}{L}\|\mathbf{W}\|_T. \quad (16)$$

For the sake of simplicity, denote $(\mathbf{W}_t - (1/L)\nabla f(\mathbf{W}_t))$ as $\mathbf{V}$ and λ_1/L as $\tilde{\lambda}_1$. Combining (5) and (16) we obtain the following form:

$$q_L(\mathbf{W}_t) = \arg\min_{\mathbf{W}} \left(\frac{1}{2}\|\mathbf{W} - \mathbf{V}\|_F^2 + \tilde{\lambda}_1\|\mathbf{W}\|_T\right) = \arg\min_{\mathbf{W}^1,\ldots,\mathbf{W}^D} \sum_{i=1}^{D}\left(\frac{1}{2}\|\mathbf{W}^i - \mathbf{V}^i\|_F^2 + \tilde{\lambda}_1\|\mathbf{W}^i\|_T\right) \quad (17)$$

where $\mathbf{W}^i$, $\mathbf{V}^i$ denotes the ith row of the matrix $\mathbf{W}$, $\mathbf{V}$ respectively. Therefore, (17) can be decomposed into D separate subproblems.

For ease notation, we denote $\mathbf{W}^i$ and $\mathbf{V}^i$ as $\mathbf{w}$ and $\mathbf{v}$ respectively. For each subproblem with the tree-structured sparsity regularization, Liu *et al.* showed that there exists an analytical solution in [30]. Firstly, we introduce a definition of the so-called index tree [30]:

Definition 1: For an index tree $\mathcal{T}$ of depth p, we let $\mathcal{T}^i = \{\mathcal{G}_1^i, \mathcal{G}_2^i, \ldots, \mathcal{G}_{n_i}^i, \}$ contain all the nodes corresponding to depth i, where $n_0 = 1$, $\mathcal{G}_1^0 = \{1, 2, \ldots, D\}$ and $n_i \geq 1$, $i = 1, 2, \ldots, p$. The nodes satisfy the following conditions: 1) the nodes from the same depth level have non-overlapping indices, i.e., $\mathcal{G}_j^i \cap \mathcal{G}_k^i = \emptyset, \forall\, i = 1, \ldots, p, j \neq k, 1 \leq j, k \leq n_i$; and 2) let $\mathcal{G}_{j0}^{i-1}$ be the parent node of a non-root node $\mathcal{G}_j^i$, then $\mathcal{G}_j^i \subseteq \mathcal{G}_{j0}^{i-1}$.

Then the tree structured sparsity norm for each row of $\mathbf{W}$ can be rewritten as $\|\mathbf{w}\|_T = \sum_{i=0}^{p}\sum_{j=1}^{n_i} w_j^i \|\mathbf{w}_{\mathcal{G}_j^i}\|$, and each subproblem is

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{v}\|_2^2 + \tilde{\lambda}_1 \sum_{i=0}^{p}\sum_{j=1}^{n_i} w_j^i \|\mathbf{w}_{\mathcal{G}_j^i}\| \quad (18)$$

(18) admits an analytical solution. By maintaining a working variable $\mathbf{u}$ initialized with $\mathbf{v}$, $\mathbf{u}$ can be updated by traversing the index tree $\mathcal{T}$ in the bottom-to-up breadth-first order according to the following operation:

$$\mathbf{u}_{\mathcal{G}_j^i}^i = \begin{cases} \mathbf{0}, & \|\mathbf{u}_{\mathcal{G}_j^i}^{i+1}\| \leq \tilde{\lambda}_j^i \\ \frac{\|\mathbf{u}_{\mathcal{G}_j^i}^{i+1}\| - \tilde{\lambda}_j^i}{\|\mathbf{u}_{\mathcal{G}_j^i}^{i+1}\|}\mathbf{u}_{\mathcal{G}_j^i}^{i+1}, & \|\mathbf{u}_{\mathcal{G}_j^i}^{i+1}\| > \tilde{\lambda}_j^i \end{cases} \quad (19)$$

where $\tilde{\lambda}_j^i = \tilde{\lambda}_1 w_j^i$. Then the output $\mathbf{u}^0$ is the solution of the subproblem (18). The APG method for solving (11) is presented in Algorithm 1.

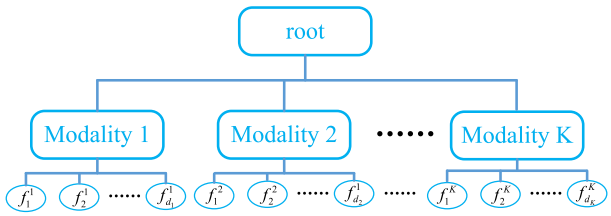


Fig. 3. Tree structure $\mathcal{T}_m$ induced by sparsity between and within modalities.

Algorithm 1 APG method for solving (11)

Input: Feature matrix $\mathbf{X}$, label indicator matrix $\mathbf{Y}$, regularization parameter λ_1, task index tree $\mathcal{T}_t$ ($\mathcal{T}_t^i = \{\mathcal{G}_1^i, \mathcal{G}_2^i, \ldots, \mathcal{G}_{n_i}^i, \}$) with weights w_j^i, parameter μ, matrix $\mathbf{Z}$ and $\mathbf{A}$.
Initialization: $L_0 > 0$, $\eta > 1$, $\mathbf{W}_0 \in \mathcal{R}^{D\times J}$, $\mathbf{V}_0 = \mathbf{W}_0$, $\alpha_0 = 1$ and $t = 0$;
while stopping criterion not met **do**
  //The generalized gradient mapping step:
  Calculate the gradient $\nabla f(\mathbf{W}_t)$, $\mathbf{V}_t = \mathbf{W}_t - (1/L)\nabla f(\mathbf{W}_t)$;
  Set $L = L_t$;
  while $F(q_L(\mathbf{V}_t)) > Q_L(q_L(\mathbf{V}_t), \mathbf{V}_t)$ **do**
    $L = \eta L$;
    $\mathbf{V}_t = \mathbf{W}_t - (1/L)\nabla f(\mathbf{W}_t)$;
  end while
  Set $L_{t+1} = L$;
  //Solve $\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}} Q_{L_{t+1}}(\mathbf{W}, \mathbf{V}_t)$
  For each row of $\mathbf{W}_{t+1}$, solve (18) according to (19) by traversing $\mathcal{T}_t$ in the bottom-to-up breadth-first order using $\mathbf{V}_t$;
  //The aggregation step:
  Set $\alpha_{t+1} = 2/(t+3)$;
  Update $\mathbf{V}_{t+1} = \mathbf{W}_{t+1} + (((1-\alpha_t)\alpha_{t+1})/\alpha_t)(\mathbf{W}_{t+1} - \mathbf{W}_t)$;
  Set $t = t + 1$;
end while
Output: $\mathbf{W}$

B. Solving Problem (12)

Now supposing that $\mathbf{W}$ and the Lagrangian multipliers $\mathbf{A}$ are fixed in the Lagrangian (10), the optimization problem related to (12) transforms into

$$\mathbf{Z}_{t+1} = \arg\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{Z} - \mathbf{R}\|_F^2 + \tilde{\lambda}_2\|\mathbf{Z}\|_{\mathrm{SGl}} \quad (20)$$

with $\mathbf{R} = \mathbf{W} + \mathbf{A}$, and $\tilde{\lambda}_2 = \lambda_2/\mu$. From the definition in (7) and (20) can be decomposed into J separate subproblems, and they take the following form:

$$\arg\min_{\mathbf{Z}_1,\ldots,\mathbf{Z}_J} \sum_{i=1}^{J}\left(\frac{1}{2}\|\mathbf{Z}_i - \mathbf{R}_i\|_F^2 + \tilde{\lambda}_2\|\mathbf{Z}_i\|_{\mathrm{SGl}}\right) \quad (21)$$

where $\mathbf{Z}_i$, $\mathbf{R}_i$ denotes the ith column of the matrix $\mathbf{Z}$ and $\mathbf{R}$, respectively. For each subproblem, denote $\mathbf{Z}_i$, $\mathbf{R}_i$ as $\mathbf{z}$ and $\mathbf{r}$, respectively. Since the K modalities are groups which partition $\mathbf{z}$ disjointedly, and each feature in a modality forms a singleton, according to the definition in (1), the set of groups and the

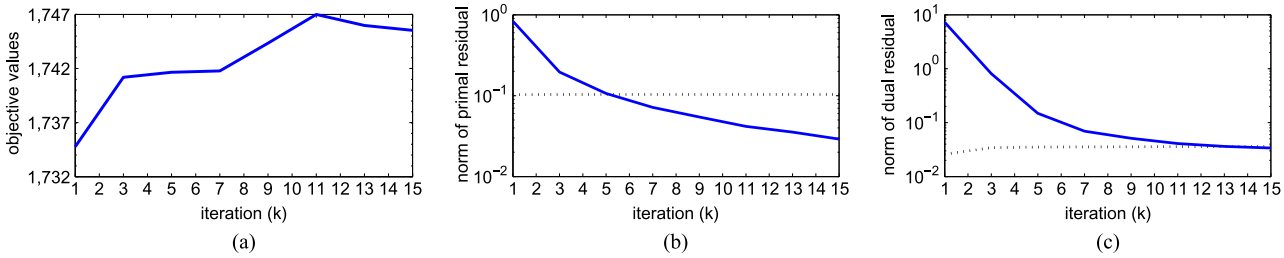


Fig. 4. An example of M^2-tMTL process on GTSRB with $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\alpha = 0.1$. (a) Curve of objective values. Norms of (b) primal residual and (c) dual residual versus iteration; the dashed lines show the convergence threshold.

singletons form an index tree structure $\mathcal{T}_m$, as shown in Fig. 3. Then, each subproblem is:

$$\mathbf{z} = \arg\min_{\mathbf{z}} \left(\frac{1}{2}\|\mathbf{z} - \mathbf{r}\|_F^2 + \tilde{\lambda}_2 \left(\sum_{k=1}^{K} w_k \|\mathbf{z}_k\| + \alpha \|\mathbf{z}\|_1 \right) \right) \quad (22)$$

It is a special case of (18) induced by the index tree $\mathcal{T}_m$, which can be solved by the same scheme as that of (18).

C. Convergence Analysis

The algorithm for solving our proposed M^2-tMTL using ADMM is summarized in Algorithm 2, the convergence of this algorithm builds upon classical convergence results of ADMM or Douglas-Rachford splitting algorithm [38]. Eckstein *et al.* proved the convergence of the generalized ADMM problem $\min_{\mathbf{x}\in\mathcal{R}^n} f(\mathbf{x}) + g(\mathbf{M}\mathbf{x})$ in Theorem 8 in [38], given that $\mathbf{M}$ has full column rank. Indeed, a direct application of this theorem tells us that our algorithm converges for any $\mu > 0$. Practically, we use the optimality conditions and stopping criterion proposed in [15], where the dual residual and primal residual at iteration $t+1$ are $s_{t+1} = -\mu(\mathbf{Z}_{t+1} - \mathbf{Z}_t)$ and $r_{t+1} = \mathbf{W}_{t+1} - \mathbf{Z}_{t+1}$, respectively. Fig. 4(a) shows the objective values versus iteration, Fig. 4(b) and (c) show that the norms of primal residual and dual residual decrease as the iteration number increases, which illustrates the convergence of M^2-tMTL in a few iterations.

Algorithm 2 ADMM for solving M^2-tMTL

Input: Feature matrix $\mathbf{X}$, label indicator matrix $\mathbf{Y}$, regularization parameter λ_1, λ_2, α. Task index tree $\mathcal{T}_t$ with weights w_t, modality index tree $\mathcal{T}_m$ with weights w_m.
Initialization: $\mu > 0$, $t = 0$, $\mathbf{W}_0$, $\mathbf{Z}_0$, $\mathbf{A}_0 \in \mathcal{R}^{D\times J}$ are matrices with 0 elements;
1: **while** stopping criterion not met **do**
2: Solving (11) according to Algorithm 1 to obtain $\mathbf{W}_{t+1}$ induced by $\mathcal{T}_t$ and weights w_t;
3: $\mathbf{R} = \mathbf{W}_{t+1} + \mathbf{A}_t$;
4: Solving (22) to obtain each column of $\mathbf{Z}_{t+1}$ according to (19) by traversing $\mathcal{T}_m$ in the bottom-to-up breadth-first order using $\mathbf{R}$.
5: $\mathbf{A}_{t+1} = \mathbf{A}_t + \mathbf{W}_{t+1} - \mathbf{Z}_{t+1}$;
6: Set $t = t + 1$;
7: **end while**
Output: Coefficient matrix $\mathbf{W}$.

D. Computational Complexity

The two most computationally demanding parts of our algorithm are solving problem (11) and problem (12). For problem (11), the computational complexity of Algorithm 1 depends on step 11, in which the time complexity for Moreau-Yosida regularization for grouped tree structure for solving (17) is $\mathcal{O}(Jp)$, then Algorithm 1 has a complexity of $\mathcal{O}(DJp_t)$ (p_t is the depth of $\mathcal{T}_t$), which mainly depends on the dimension of feature D. Similarly, (22) has a complexity of $\mathcal{O}(JD\log D)$ (as $\mathcal{T}_m$ is a balanced tree, $p_m = \mathcal{O}(\log D)$).

V. Benchmark Data Sets

To our best knowledge, there are six publicly available traffic sign data sets:

- German TSR Benchmark (GTSRB)[1] [4], [39]
- KUL Belgium Traffic Sign Classification Data set (BelgiumTSC)[2] [40]
- Swedish Traffic Signs Data set (STS Data set)[3] [41]
- RUG Traffic Sign Image Database (RUG Data set)[4] [42]
- Stereopolis Database[5] [43]
- LISA Traffic Sign Data set (LISA Data set)[6] [1].

Among the above data sets, GTSRB is the most widespread and it is primarily gathered for classification task. The remaining five data sets are primarily used for detection as they all contain images of the whole scenes from the videos or include videos directly, which means that tracking system can be used to make detection results more robust. Although these five data sets are not yet widely used for classification, to show more informative of the relative merits of methods performance, we choose other two data sets: BelgiumTSC and LISA data sets, for evaluating and comparing the performance of our proposed method with state-of-the-art methods. More information about all the six data sets can be referred to Ref. [1].

Some details of the three data sets used in this paper are presented in Table I, note that the training and testing sets are selected by us in LISA data set, as it does not partition the annotations into training and testing sets. Some sample signs

[1] http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset
[2] http://btsd.ethz.ch/shareddata/
[3] http://www.cvl.isy.liu.se/research/datasets/traffic-signs-dataset/
[4] http://www.cs.rug.nl/~imaging/databases/traffic_sign_database/traffic_sign_database.html
[5] http://www.itowns.fr/benchmarking.html
[6] http://cvrr.ucsd.edu/LISA/lisa-traffic-sign-dataset.html

TABLE I
DETAILS OF DATA SETS USED IN EXPERIMENTS

	♯ classes	♯ annotations	♯ training samples	♯ testing samples	features provided	color or gray	sign sizes(pixels)
GTSRB	43	51,840	39,209	12,631	YES	color	15×15-222×193
BelgiumTSC	62	13,444	4,591	2,534	NO	color	100×100-$1,628 \times 1,236$
LISA Dataset	47	7,855	4,000	3,855	NO	both	6×6-167×158

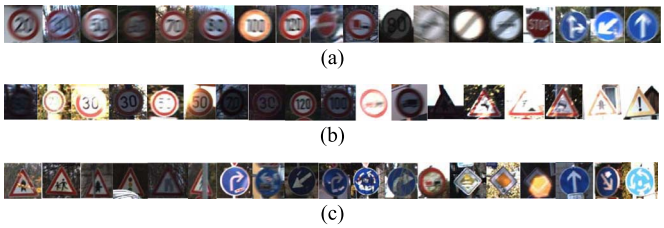


Fig. 5. Some sample signs in GTSRB with different recognition challenges. (a) Traffic signs with motion blurring and erosion. (b) Traffic signs with various lighting conditions. (c) Traffic signs with partial occlusion.

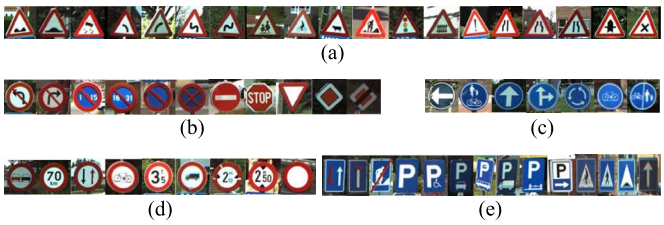


Fig. 6. Five subsets of similar traffic signs in Belgium Traffic Sign Classification Benchmark. (a) Danger signs. (b) Unique signs. (c) Circle mandatory signs. (d) Limit signs. (e) Rectangle mandatory signs.

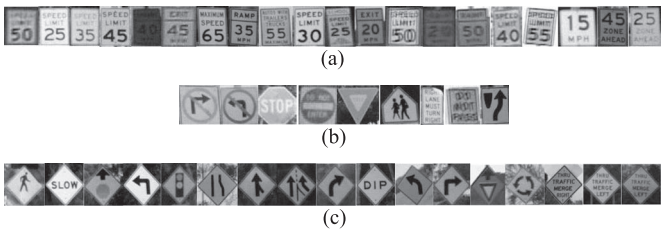


Fig. 7. Three subsets of similar traffic signs in LISA data set. (a) Limit signs. (b) Unique signs. (c) Mandatory signs.

of the three data sets are shown in Figs. 5–7, respectively. For GTSRB, we use the pre-extracted features provided for training the M^2-tMTL model, the features as well as their details are presented in Table II. Among them, the HOG modality uses a cell size 5×5 pixels, block size 2×2 cells, and orientation bins 8 to form a 1568 dims vector. The Haar features are extracted by using five different types (see Table II: HE = "horizontal edge," VE = "vertical edge," HB = "horizontal bar," VB = "vertical bar," CB = "chessboard") in different sizes to form a total of 12 modalities. However, there are no pre-extracted features provided in BelgiumTSC and LISA data sets, so for these two data sets, we rescale the signs to 40×40 and extract 5 modalities of features. The details of the features for these two data sets are presented in Tables III and IV, respectively.

VI. EXPERIMENTS

In this section, we present the results of our method on parameter setting and feature selection on GTSRB, and compare the performance on recognition accuracy with state-of-the-art methods on GTSRB, BelgiumTSC and LISA. All our experiments are conducted on an Intel(R) Core(TM) i5-4590(3.3 GHz) with 32 GB DDR4.

A. Parameter Setting and Tuning

In this section, we try to investigate the influence of the three parameters λ_1, λ_2 and α on the recognition performance. Take GTSRB for example, for the training data set, we use 10 different values for λ_1, λ_2, i.e., $\lambda_1, \lambda_2 \in \{0.05, 0.1, 0.5, 1, 2.5, 5, 8, 10, 15, 20\}$, and 9 different values for α, i.e., $\alpha \in \{0.05, 0.1, 0.5, 0.75, 1, 2.5, 5, 10, 20\}$, resulting in a total of 900 triples of $(\lambda_1, \lambda_2, \alpha)$. For a fixed α, we get the recognition accuracy on each pair of (λ_1, λ_2), and the recognition accuracy is an average of fivefold cross validation on the training data set. In Fig. 8, we depict three examples of parameter tuning by the fivefold cross validation with respect to recognition accuracy versus λ_1 and λ_2 when α varies. We found that the highest performance is achieved at some intermediate values of λ_1 and λ_2 when $\alpha = 2.5$. Therefore, in our experiment, we set $\alpha = 2.5$, $\lambda_1 = 8$, and $\lambda_2 = 5$.

B. Feature Selection on GTSRB

To explore the feature selection ability of our method, we output the coefficient vectors to investigate the results of feature selection for both between and within modalities. In Fig. 9(a)–(c), the coefficient vectors of the HOG features for three classes, signs 20 and 30 from the speed limit subset and the first sign in the danger subset presented in Fig. 2(a), are plotted to show the results of our method for within modality feature selection. Fig. 9(d) depicts the dimensionality of features in different modalities for these three classes. We can observe that, from Fig. 9(a)–(d), M^2-tMTL successfully induces sparsity of coefficient values both between and within modalities, only a small portion of the features are selected for classification. Furthermore, Fig. 9(a)–(c) show that, within the HOG feature modality, our method tends to select the same features for the signs 20 and 30 in speed limit subset. In addition, from Fig. 9(d), we can see the coefficients of first two classes have the same non-zero feature modalities, which are very different from that of the sign from danger subset, since they are more similar to each other than the signs in danger subset. That is to say, the coefficients for the classes from the same subset are more correlated than that of the classes from the different subset.

TABLE II
DETAILS OF THE FEATURES IN GTSRB

Feature modality	HOG	Hue	Haar1	Haar2	Haar3	Haar4	Haar5	Haar6	Haar7	Haar8	Haar9	Haar10	Haar11	Haar12
cell size/feature type	5×5	–	HE	VE	HE	VE	HE	VE	HB	HB	VB	VB	CB	CB
block size	2×2	–	4×4	4×4	6×6	6×6	8×8	8×8	6×6	6×6	9×9	9×9	6×6	10×10
dimensionality	1,568	256	1,156	1,156	1,024	1,024	900	900	1,024	1,024	784	784	1,024	784

TABLE III
DETAILS OF THE FEATURES IN BELGIUMTSC

Feature modality	HOG	Hue	Haar1	Haar2	Haar3
cell size/feature type	5×5	–	HE	VE	HB
block size	2×2	–	4×4	8×8	9×9
dimensionality	1,568	256	1,156	900	784

TABLE IV
DETAILS OF THE FEATURES IN LISA

Feature modality	HOG	Gray value	Haar1	Haar2	Haar3
cell size/feature type	5×5	–	HE	VE	HB
block size	2×2	–	4×4	8×8	9×9
dimensionality	1,568	1,600	1,156	900	784

To show the performance on feature selection more clearly, we also output a row vector of the coefficient matrix in Fig. 9(e), which shows that our method successfully selects the 322th feature in the HOG modality for the classification tasks in the speed limit and other prohibitive sign subsets. Except for one sign in the derestriction subset and one sign in the unique subset, the coefficients of other classifiers are all zeros on this feature as the signs in these subsets are very different from speed limit signs and other prohibitive signs.

We also compare our method with a simple lasso regression for feature selection, and the results are plotted in Fig. 10. It can be seen that although lasso regression can successfully induce sparsity of coefficients within modalities [see Fig. 10(a)–(c)], it cannot produce sparse groups of coefficients [see Fig. 10(d)]. The sparsity degree of lasso regression is 58.6%, comparing with M^2-tMTL 80.8%, which indicates that our algorithm achieves better performance for feature selection. Moreover, there are no coefficients correlations between similar tasks [see Fig. 10(e)], and thus, the feature selection results are more consistent and interpretable in our method.

C. Recognition Accuracy on GSTRB

For better understanding the performance of our method, we give some recognition results of other state-of-the-art methods, including: Committee of CNNs [19], Multi-Scale CNN [18], Human Performance (both best individual and average) [4], Random Forests [4], LDA [4], SRLP [24], SRGE [25], and linear SVM (LSVM). The overall recognition accuracy results for all 43-class and the individual results for each subset are listed in Table V.

From the overall recognition accuracy presented in Table V, we can see that our method outperformed LDA, Random Forest and SRLP, and gives a slightly higher accuracy than SRGE. The CNN-based methods exhibit superior performance over other methods, and even it outperformed humans. However, this high recognition accuracy is at the cost of a large amount of computing time for training the model. The ensemble CNNs [20] achieved an accuracy of 99.65%, which outperformed the best record achieved by the IJCNN 2011 recognition competition's winner Dan Ciresan. It costs several minutes to train the model on two Tesla C2075 GPUs and a 12 core Intel(R) Core i7-3960X 3.3-GHz computer, while the CCNNs [19] costs about 37 hours for training on a system with a Core i7-950 (3.33-GHz), and four graphics cards of type GTX 580. However, our method only spends 0.5 hours for the whole data set on a Core i5-4590 CPU with 32 GB DDR3.

The details about required computation cost of CCNNs and our method are presented in Table VI, where the required hardware platform and training time are listed, and the test cost is evaluated in multiplication operations. Note that because of the sparse feature selection ability of our algorithm, only about a hundred thousand times (weight matrix dims: 13,408 $\times$ 43, sparsity degree 80.8%) of multiplication are needed for testing a sample, which is much less than that of CCNNs, as the computation time complexity of CNN for evaluating a test sample depends closely on the trained architecture, such as the number of layers, the maps on each layer and the convolution kernel size (the 8-layer DNN architecture can be referred to [19]). Specifically, it costs only about 0.0998 s for testing a sample using our platform, which indicates that our system can be used for real-time traffic sign recognition in real automotive applications.

Our method gives slightly higher accuracy than SRGE on the whole data set, and the results of individual subsets show that our method performs better than SRGE in three subsets (speed limits, derestriction and unique signs). The competitive performance of SRGE and our method may originate from the relations of these two algorithms. The success of SRGE comes from the graph structure used to model between-class discriminative information, while in our method, the tree structure is used to model the correlations of all the classification tasks. On the other hand, both LDA and SRGE are dimensionality reduction-based methods, they project the original feature space to a reduced space by learning a linear space. However, they need a classifier when performing recognition, either k-Nearest Neighbor classifier or other parameterized classifiers, the former has a high memory usage, which needs to save all the training samples; while the latter needs another training algorithm, e.g., SVM. On the contrary, our method is parameterized, only the regression coefficients needed to be saved for classification, and it can also be used for feature reduction, as only a small portion of features have non-zero coefficients.

We also test the performance of our method in reducing training samples situations, and compare the results with the

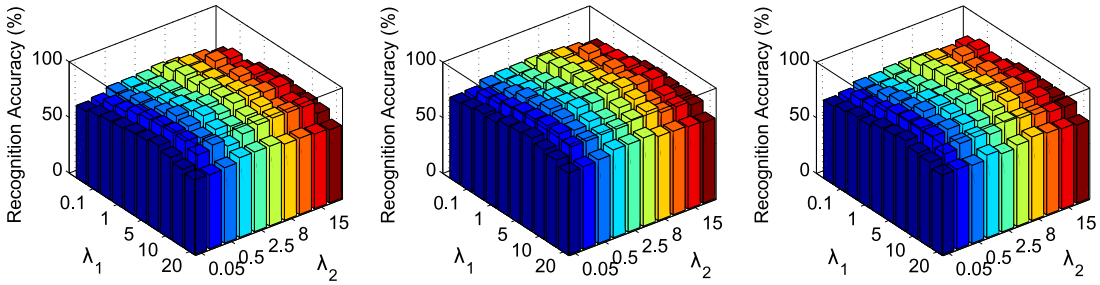


Fig. 8. Recognition accuracy rate (in %) versus λ_1 and λ_2 when (a) $\alpha = 0.05$, (b) $\alpha = 0.75$, and (c) $\alpha = 5$.

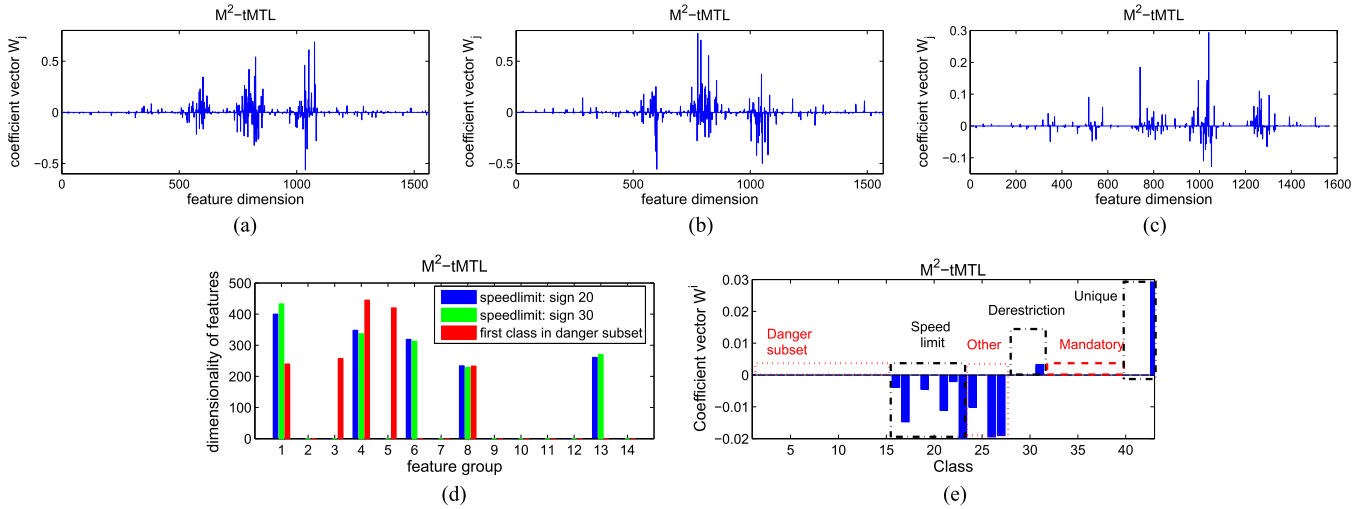


Fig. 9. M^2-tMTL results for feature selection: coefficient vectors of the HOG features for (a) signs 20; (b) signs 30 from the speed limit subset and (c) the first sign in the danger subset, which show the results for feature selection within modality; (d) results of feature selection between modalities, where different colors of histograms indicate the dimensionality of features in different modalities; and (e) the 322th row vector of the HOG feature coefficient, in which the groups for the six subsets are marked.

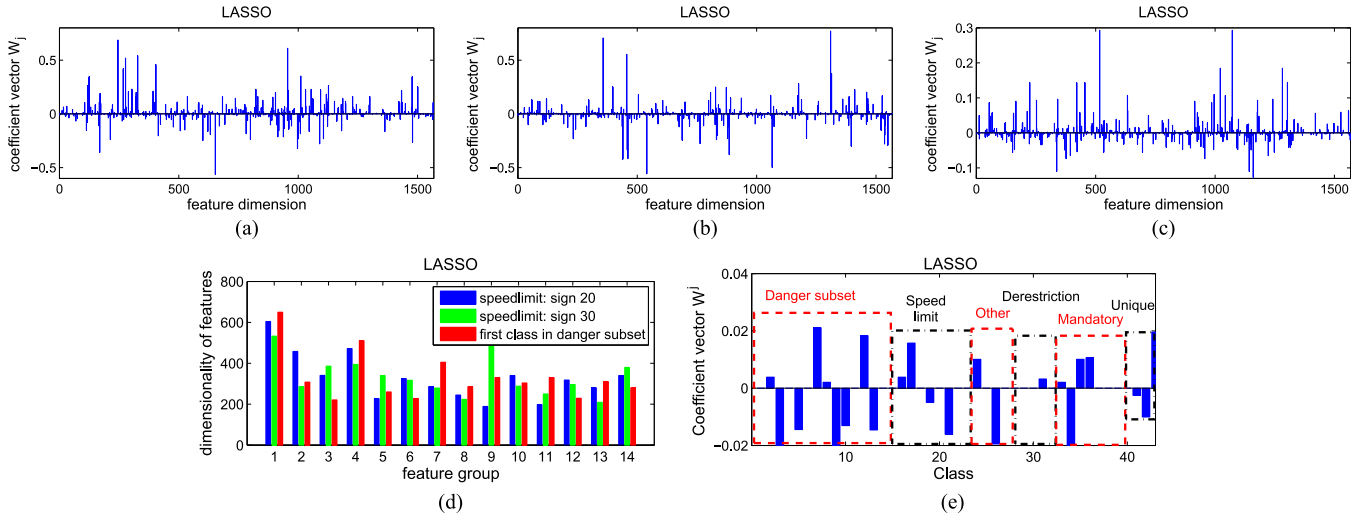


Fig. 10. Lasso results for feature selection: coefficient vectors of the HOG features for (a) signs 20; (b) signs 30 from the speed limit subset and (c) the first sign in the danger subset, which show the results for feature selection within modality; (d) results of feature selection between modalities, where different colors of histograms indicate the dimensionality of features in different modalities; and (e) the 322th row vector of the HOG feature coefficient, in which the groups for the six subsets are marked.

comparative method SRGE. We randomly select a portion of samples ($p\%$, $p \in \{40, 50, 60, 70, 80, 90\}$) in each category from the training set to train new models, and Fig. 11 shows the overall recognition accuracy results. It can be seen that when reducing training samples, the performance degradation of SRGE is much severe than our method, which demonstrates that our

TABLE V
INDIVIDUAL SUBSETS AND OVERALL RECOGNITION ACCURACY RESULTS (IN %) FOR GTSRB

Methods/Subsets	Speed limits	Other prohibitions	Derestriction	Mandatory	Danger	Unique	Overall
Committee of CNNs[19]	99.47	99.93	99.72	99.89	99.07	99.22	99.46
Multi-Scale CNN[18]	98.61	99.87	94.44	97.18	98.03	98.63	98.31
Human (best individual)[4]	98.32	99.87	98.89	100.00	99.21	100.00	99.22
Human (average)[4]	97.63	99.93	98.89	99.72	98.67	100.00	98.84
LDA[4]	95.37	96.80	85.83	97.18	93.73	98.63	95.68
SRLP[24]	96.04	99.32	99.67	95.89	95.99	99.90	96.75
Random Forests[4]	95.95	96.80	85.83	97.18	93.73	98.63	96.14
LSVM	95.87	95.92	88.31	96.39	94.22	98.15	95.74
SRGE[25]	97.68	99.24	97.23	98.54	97.31	99.51	98.19
M^2-tMTL	98.33	98.97	98.71	97.26	97.23	99.61	**98.27**

TABLE VI
DETAILS OF THE COMPUTATION COST OF CCNNS AND M^2-tMTL

Methods/Details	Hardware platform	Training time	Testing time(in multiplication operations)
Committee of CNNs[19]	1 Core i7-950(3.33GHz), 4 GTX 580 graphics cards	37 hours	∼ten millions
M^2-tMTL	1 Core i5-4590(3.3GHz), no graphics card	0.5 hours	∼a hundred thousand

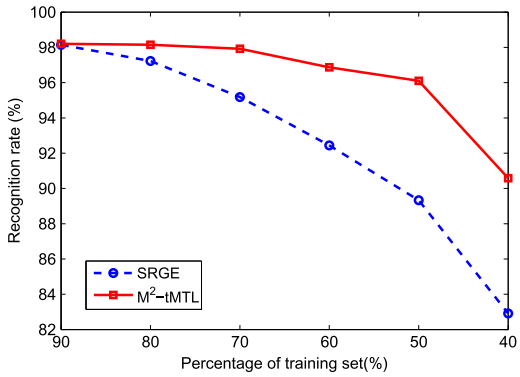


Fig. 11. Overall recognition accuracy for reducing training samples' situations.

method is particularly suitable for small training samples situation. The main reason is that the sparse representation-based method needs plenty of training samples for representing the subspace each class may embed in, while the advantage of our method lies in the ability to explore the relationship between the classification tasks for improving generation ability.

To explore the performance of M^2-tMTL in difficult situations, such as motion blurring, various lighting conditions and partial occlusion (see Fig. 5), we select samples from the entire test set to form three new subsets, and compare M^2-tMTL with two dimensionality reduction-based methods: LDA, SRGE, and linear SVM (LSVM). The results are listed in Table VII. SRGE and M^2-tMTL achieves considerably better performance than the other two methods in three situations. As sparse representation-based method is rather effective under partial occlusion [23], SRGE achieves the best performance in this situation. However, as sparse representation-based method degrades severely when signs are not aligned, our method works better than SRGE in other two situations.

D. Recognition Accuracy on BelgiumTSC and LISA Data Sets

According to the similarity between classes, the traffic signs of the BelgiumTSC and LISA Data Sets can be separated into subsets (see Figs. 6 and 7). These subsets can be used as the prior knowledge for M^2-tMTL to construct the tree structure. All classes of the two data sets are used to test the methods, and we compare our method with LDA, SRGE, and LSVM. The individual results for each subset and overall recognition accuracy of the two data sets are listed in Tables VIII and IX, respectively. It can be seen that our method achieves the best performance both on these two data sets.

TABLE VII
RECOGNITION ACCURACY RESULTS (IN %) FOR SUBSETS OF GTSRB WITH DIFFERENT RECOGNITION CHALLENGES

Difficult situations/Methods	LDA	LSVM	SRGE	M^2-tMTL
Motion blurring	95.32	95.05	96.47	96.75
Various lighting	82.50	81.69	92.56	93.19
Partial occlusion	77.13	79.93	93.23	92.65

TABLE VIII
INDIVIDUAL SUBSETS AND OVERALL RECOGNITION ACCURACY RESULTS (IN %) FOR BELGIUMTSC

	Danger	Unique	Circle	Rectangle	Limit	Overall
LDA	89.87	95.01	98.48	97.26	91.35	93.98
LSVM	91.45	95.21	97.66	98.14	92.05	94.48
SRGE	93.33	98.56	99.42	99.39	94.01	96.58
M^2-tMTL	94.18	97.96	99.50	99.18	95.13	96.90

TABLE IX
INDIVIDUAL SUBSETS AND OVERALL RECOGNITION ACCURACY RESULTS (IN %) FOR LISA DATA SET

	Speedlimits	Mandatory	Warning	Overall
LDA	67.55	59.16	62.48	62.59
LSVM	71.53	70.81	67.22	69.38
SRGE	79.25	81.43	77.29	79.08
M^2-tMTL	78.16	83.25	80.30	80.75

E. Discussion on Relative Difficulty of the Data Sets

From the recognition results presented above, it can be seen that many methods have achieved 98%+ recognition rates on GTSRB, the main reason is that each physical sign has

30 different tracking images in GTSRB, which makes the data set quite saturated. Furthermore, the portion of signs under difficult situations is relatively small. Although the number of samples for each class in BelgiumTSC is much smaller than that in GTSRB, algorithms still exhibit good performance on it. Observing the samples in BelgiumTSC, we can see that the sign sizes are relatively larger than that in GTSRB, so that the features extracted are not so noisy. In addition, the most frequently difficult situation is the change of view, which can be overcome by rescale the sample signs. Compared with the former two data sets, sample sizes in LISA data set are much smaller, and the difficult situation is most serious in the three data sets. Thus, LISA data set seems to be less saturated than the other two data sets, as it provides more informative of the relative merits for the given algorithms, new algorithms can be more easily compared in this data set.

VII. Conclusion

We proposed a multi-modal tree-structure embedded multi-task learning algorithm called M^2-tMTL for traffic sign recognition. For the recognition of traffic signs with similarity, the algorithm can select similar discriminative features that are shared between classes with the similar classes, because the tree-structure is used to model the hierarchical correlations between the classification tasks, and multi-task learning is adopted to learn the relations across tasks jointly. Furthermore, our algorithm leverages multi-modal features to give a better description of the signs, and only a small portion of them are needed for final classification by using structured sparsity-induced norm. Finally, ADMM is proved to be very effective for solving the highly non-smooth objective with guaranteed convergence. Experiments on three benchmark data sets show that the performance of our algorithm is better than or competitive comparative to other state-of-the-art algorithms with less computational and memory cost.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers for their invaluable suggestions, which have been incorporated to improve the quality of this paper dramatically.

References

[1] A. Mϕgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1484–1497, Dec. 2012.

[2] A. de la Escalera, J. M. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image Vis. Comput.*, vol. 21, no. 3, pp. 247–258, 2003.

[3] V. Kastrinaki, M. Zervakis, and K. Kalaitzakis, "A survey of video processing techniques for traffic applications," *Image Vis. Comput.*, vol. 21, no. 4, pp. 359–381, 2003.

[4] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Netw.*, vol. 32, pp. 323–332, 2012.

[5] J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 436–443.

[6] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 606–613.

[7] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, 2008.

[8] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, 2012.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[10] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Structured sparsity through convex optimization," *Statist. Sci.*, vol. 27, no. 4, pp. 450–468, 2012.

[11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., B, Statist. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.

[12] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Mach. Learn.*, vol. 73, no. 3, pp. 243–272, 2008.

[13] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.

[14] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l 2, 1-norm minimization," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, 2009, pp. 339–348.

[15] 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.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 1, pp. 886–893.

[17] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2002, vol. 1, pp. I-900–I-903.

[18] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Proc. IEEE IJCNN*, 2011, pp. 2809–2813.

[19] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Proc. IEEE IJCNN*, 2011, pp. 1918–1921.

[20] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1991–2000, Oct. 2014.

[21] P. Paclík, J. Novovičová, P. Pudil, and P. Somol, "Road sign classification using Laplace kernel classifier," *Pattern Recognit. Lett.*, vol. 21, no. 13/14, pp. 1165–1173, 2000.

[22] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. Lopez-Ferreras, "Road-sign detection and recognition based on support vector machines," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 2, pp. 264–278, Jun. 2007.

[23] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[24] R. Timofte and L. Van Gool, "Sparse representation based projections," in *Proc. 22nd Brit. Mach. Vis. Conf.*, 2011, pp. 61.1–61.12.

[25] K. Lu, Z. Ding, and S. Ge, "Sparse-representation-based graph embedding for traffic sign recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1515–1524, Dec. 2012.

[26] A. de la Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," *IEEE Trans. Ind. Electron.*, vol. 44, no. 6, pp. 848–859, Dec. 1997.

[27] J. F. Khan, S. M. A. Bhuiyan, and R. R. Adhami, "Image segmentation and shape analysis for road-sign detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 83–96, Mar. 2011.

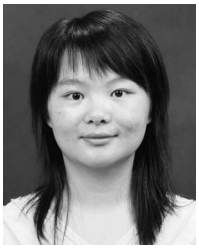
[28] P. Gil Jiménez, S. M. Bascón, H. G. Moreno, S. L. Arroyo, and F. L. Ferreras, "Traffic sign shape classification and localization based on the normalized FFT of the signature of blobs and 2D homographies," *Signal Process.*, vol. 88, no. 12, pp. 2943–2955, 2008.

[29] L. W. Tsai, J. W. Hsieh, C. H. Chuang, Y. J. Tseng, K. C. Fan, and C. C. Lee, "Road sign detection using eigen colour," *IET Comput. Vis.*, vol. 2, no. 3, p. 164, 2008.

[30] J. Liu and J. Ye, "Moreau-Yosida regularization for grouped tree structure learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1459–1467.

[31] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 543–550.

[32] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *arXiv preprint arXiv:1001.0736*. [Online]. Available: http://arxiv.org/abs/1001.0736
[33] R. Glowinski and A. Marroco, "Sur l'approximation, parílíments finis d'ordre un, et la rísolution, par pínalisation-dualití d'une classe de problímes de Dirichlet non liníaires," *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathímatique et Analyse Numírique*, vol. 9, no. R2, pp. 41–76, 1975.
[34] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.
[35] Y. Nesterov, "Gradient methods for minimizing composite objective function. [Online]. Available: http://dial.uclouvain.be/pr/boreal/object/boreal:5122
[36] X. Chen, W. Pan, J. T. Kwok, and J. G. Carbonell, "Accelerated gradient method for multi-task sparse learning problem," in *Proc. IEEE Int. Conf. Data Mining*, 2009, pp. 746–751.
[37] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[38] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Programm.*, vol. 55, no. 1–3, pp. 293–318, 1992.
[39] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German traffic sign recognition benchmark: A multi-class classification competition," in *Proc. IEEE IJCNN*, 2011, pp. 1453–1460.
[40] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 633–647, 2014.
[41] F. Larsson and M. Felsberg, "Using Fourier descriptors and spatial models for traffic sign recognition," in *Image Analysis*. Berlin, Germany: Springer-Verlag, 2011, pp. 238–249.
[42] C. Grigorescu and N. Petkov, "Distance sets for shape filters and shape recognition," *IEEE Trans. Image Process.*, vol. 12, no. 10, pp. 1274–1286, Oct. 2003.
[43] R. Belaroussi, P. Foucher, J.-P. Tarel, B. Soheilian, P. Charbonnier, and N. Paparoditis, "Road sign detection in images: A case study," in *Proc. IEEE 20th ICPR*, 2010, pp. 484–488.

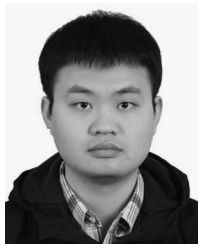


Xiao Lu received the B.E. degree from Hunan University, Changsha, China, in 2007; the M.S. degree from Southeast University, Nanjing, China, in 2010; and the Ph.D. degree from Hunan University in 2015, all in electrical engineering. During her Ph.D. studies, she was with the National Engineering Laboratory for Robot Visual Perception and Control Technology, Hunan University.

She is currently an Assistant Professor with the College of Engineering and Design, Hunan Normal University, Changsha. Her research interests include machine vision, pattern recognition, and machine learning.

Yaonan Wang received the B.S. degree in computer engineering from East China University of Science and Technology, Shanghai, China, in 1981 and the M.S. and Ph.D. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1994, respectively.

He was a Postdoctoral Researcher in electrical engineering with the National University of Defense Technology, Changsha, from 1995 to 1997 and with Alexander von Humboldt Stiftung in 1997. From 1998 to 2000, he was a Senior Humboldt Fellow in Germany, and from 2001 to 2004, he was a Visiting Professor with the University of Bremen, Bremen, Germany. Since 1995, he has been a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests are in artificial intelligence, robotic control, and computer vision for industrial applications.



Xuanyu Zhou received the B.S. degree in electrical engineering and the M.S. degree in computer science from Wuhan University, Wuhan, China, in 2007, in 2011, respectively, where he is currently a Ph.D. candidate in computer science with the School of Computer Science.

His research interests include data mining, pattern recognition, and machine learning.

Zhenjun Zhang received the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology, Wuhan, China, in 2011.

He is currently an Assistant Professor with the College of Electrical and Information Engineering, Hunan University, Changsha, China. His research interests include machine learning, data mining, computer vision, and the applications.

Zhigang Ling received the B.E. degree from Chang'an University, Xi'an, China, in 2000 and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, in 2003 and 2010, respectively, all in electrical engineering.

He is currently an Assistant Professor with the College of Electrical and Information Engineering, Hunan University, Changsha, China. His research interests include image processing, pattern recognition, and machine learning.