# Semi-Supervised Multi-task Learning for Scene Recognition

Xiaoqiang Lu, Xuelong Li, *Fellow, IEEE*, and Lichao Mou

***Abstract*—Scene recognition has been widely studied to understand visual information from the level of objects and their relationships. Towards scene recognition, many methods have been proposed. They however encounter difficulty to improve the accuracy, mainly due to two limitations: 1) lack of analysis of intrinsic relationships across different scales, say, the initial input and its down-sampled versions. 2) redundant features exist. This paper develops a semi-supervised learning mechanism to reduce the above two limitations. To address the first limitation, we propose a multi-task model to integrate scene images of different resolutions. For the second limitation, we build a model of sparse feature selection based manifold regularized (SFSMR) to select the optimal information and preserve the underlying manifold structure of data. SFSMR coordinates the advantages of sparse feature selection and manifold regulation. Finally, we link the multi-task model and SFSMR, and propose the semi-supervised learning method to significantly reduce the two limitations. Experimental results reports the improvements of the accuracy in scene recognition.**

***Index Terms*—Scene recognition, multi-task learning, sparse selection, manifold regularized.**

## I. Introduction

In computer vision, scene recognition has been standing as a hot research topic to understand visual scenes in the last decade. For example, we can recognize the context of an image as a scene (forest, highway, and living room, *etc.*). Some applications of scene understanding are of great value, such as object recognition/detection [1], [2], [3], [4], content based image retrieval(CBIR) [5], [6], [7], [8], [9], and video frame quickly location *etc.*. Scene recognition is of great value for reducing the semantic gap between human beings and computers on scene understanding. However, it is a challenging task in recognizing the semantic category of a given image due to the content variations in complex scenes. Many scene recognition methods have been proposed to learn a mapping between a set of low-level features and meaningful semantic categories. However, how to reduce the large semantic gap is still a challenging task in scene recognition.

There have been many efforts to decrease the so-called semantic gap between low features and human semantics. Many semantic modeling methods try to build an intermediate semantic representation to bridge over the semantic gap [10], [11], [12], [13], [14]. The semantic modeling methods can be roughly divided into three categories: object-based method, bag-of-visual-features method and attribute based method. Object-based methods [15], [16], [17], [18], [19], [20] defined a set of objects of the images as semantic representation to reduce the semantic gap. Generally, object-based method can be divided into two steps. First, different regions in a given image can be segmented and classifiers are exploited to label different regions which can be regarded as an object. Second, the global scene can be analyzed and recognized using the information of objects. The bag-of-visual-features (BoF) methods, which stem from the technology of bag-of-words (BoW), have achieved considerable success in the field of text analysis. The BoF-based methods constructed a set of visual words generated from the image on an evenly-spaced grid. In this case, the image can be represented by many visual words. Therefore, the BoF-based methods can obtain an intermediate representation by interpreting images at a high semantic level. For BoF-based methods, however, the process of describing an image with a histogram of visual words is sub-optimal. Moreover, the coarse quantization operated by using a pre-defined visual vocabulary will reduce the power of local descriptors and disregards features spatial layout [21]. To solve this problem, Lazebnik *et al.* [22] proposed Spatial Pyramid Matching on the basis of BoW. Additionally, some probability models are exploited to scene recognition such as Probabilistic Latent Semantic Analysis (pLSA) model [23] and Latent Dirichlet Allocation (LDA) [24] model. These models can represent image as an assembly of several topic semantics which can be seen as objects. In this case, a scene containing several objects can be considered to be constituted by a set of topic probabilistic models.

Recently, attribute based methods have been exploited to represent an image with a set of meaningful visual attributes in scene recognition due to their interesting properties [25], [26], [27], [28]. For example, Torresani *et al.* [25] divided an image into 2659-dimensional vector, which can be represented by a visual attribute. Parikh *et al.* [26] learned both discriminative and nameable visual attributes from a set of images in a semi-supervised manner.

Although most scene recognition methods can obtain competitive state-of-the-art recognition accuracy, there exist two limitations for these methods to prohibit the improvement of accuracy in scene recognition. This is mainly due to two reasons: 1) few methods can exploit the initial input and its different resolution version to capture the relationship across different scales. It is pointed out that utilizing multi-resolution images data is beneficial for scene semantics understanding. However, most scene recognition methods only extract low-level features from single resolution image which cannot

Manuscript received XX XX, XXXX.

X. Lu, X. Li and L. Mou are with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China (xuelong_li@opt.ac.cn).

well represent completely the whole scene. we argue that the differences existing in the features of local patch and the global spatial structure keeps the same for scene recognition. Hence, it is necessary to exploit the relationship among different resolutions to improve recognition accuracy. 2) some redundant features may disturb the accuracy of scene recognition. One critical issue of extracting features is whether the features are beneficial to recognize the semantic category of an image. Some unfavorable features used in scene recognition methods may disturb the accuracy of recognition. Many researchers have proved that appropriate features can significantly improve the performance on challenging scene recognition tasks. In addition, the computational complexity of scene recognition methods has been important over the past years. This is because many large scale datasets are emerging, such as *MIT* Indoor [29], a dataset of 67 indoor scenes categories, or *SUN* dataset [30]. Hence, it is possible to leverage the optimal feature data to classifier, which can boost the computational efficiency.

In this paper, we develop a mechanism to overcome the aforementioned limitations. To address the first limitations, we propose a multi-task model to share the information contained in different resolutions images. This is because multi-task model can improve the performance of multiple related tasks by exploiting the intrinsic relationships among multi-resolution images. For the second limitation, we combine sparse feature selection and manifold regularized learning, called *sparse feature selection based manifold regularized method* (SFSMR), to select the optimal information while preserving the underlying manifold structure of data. Finally, by integrating the multi-task model and SFSMR, the proposed method can be designed to relieve the aforementioned limitations and further improve the accuracy of scene recognition. Compared with the existing methods, SFSMR has the following three main advantages. First, SFSMR takes multi-resolution images as multiple related tasks and utilize the common knowledge of multiple tasks. This is because multi-resolution images generated from the same scene have the same global spatial structures and the differently local features. In this case, different tasks are simultaneously learned in a joint framework to improve the performance in scene recognition. Second, by using the proposed method, the underlying manifold structure of each feature data is preserved and the optimal features can be chosen, resulting in a more faithful result. Third, by using the $l_{2,1}$ norm term and the trace norm term, the correlation of different features at multi-resolutions can be exploited and the information from different tasks can be transferred among multiple tasks.

The rest of this paper is organized as follows: Section II briefly discusses the related work of semi-supervised manifold learning. The proposed method for scene recognition is shown in section III. In section IV, the experimental results on different datasets are reported to verify the effectiveness of the proposed method. Section V concludes.

## II. RELATED WORK

Denote $[x_1, x_2, \ldots, x_p, \ldots, x_n]$ as the $n$ training samples from $c$ classes and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n] \in \{0,1\}^{n\times c}$ as the corresponding label matrix. Let $\mathbf{F} = [\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_n]^T \in \mathbb{R}^{n\times c}$, where $F_i$ is the predicted label vector of $x_i$. The prior semi-supervised learning work, including local and global consistency (LGC) [31] was briefly reviewed in this subsection. LGC is a graph based classification method, whose objection function is shown as following:

$$g_l(F) = \sum_{i,j=1}^{n} S_{ij} \| \frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{jj}}} \|_F^2 + \lambda \sum_{i=1}^{n} \|F_i - Y_i\|^2, \quad (1)$$

where $\lambda$ is a parameter and $D$ is a diagonal matrix with its diagonal element $D_{ii} = \sum_j S_{ij}$. Eq. (1) can be transformed as:

$$g_l(F) = Tr(F^T M F) + Tr(F - Y)^T U (F - Y), \quad (2)$$

where $M \in \mathbb{R}^{m\times m}$ is a graph Laplacian matrix and $U \in \mathbb{R}^{m\times m}$ is a diagonal matrix. LGC can effectively exploit label information and manifold structure of the labeled data form both labeled and unlabeled data [31].

## III. THE PROPOSED METHOD

In this section, we will show a multi-task model which exploits the information correlation in different resolutions images, as shown in Fig. 1. And we combine a sparse feature selection and manifold regularized learning, called sparse feature selection based manifold regularized method (SFSMR), to select the optimal information and preserve the underlying manifold structure of data.

### A. Multi-task Learning Model

Multi-task framework has shown its success in various computer vision and pattern recognition problems [32], [33], [34], [35], [36]. This section will discuss how the multi-task model is employed to share the information among different resolution images of same scene.

Recently, most scene recognition methods only extract low-level features from a single resolution image, which cannot represent completely the whole scene. Hence, how to simultaneously exploit the information of different resolutions images is crucially important for the scene recognition problem.

Researchers have witnessed the success of multi-task learning in image processing and pattern recognition [37], [38]. The success of multi-task framework may shed light on the problem of scene recognition. The main idea of multi-task learning is that there are a set of models to be learned and these models are affected by some common factors. These models or tasks are related, i.e., they are not independent. The shared information can be transferred from one task to other related tasks [39]. When the size of training set is small, the multi-task learning method can significantly improve the ability of generalization [40]. Motivated by the success of multi-task learning, a multi-task model is proposed to share the information and structure of different resolutions image in scene recognition. It is assumed that there are $t$ tasks ($t$ resolutions) including all $c$ scene categories. The $l$-th task $X_l$ includes $m$ training data $X_l = [x_1, \ldots, x_m]$ with ground-truth labels $Y_l = [y_1, \ldots, y_m]$ from $c$ scene categories. That is,

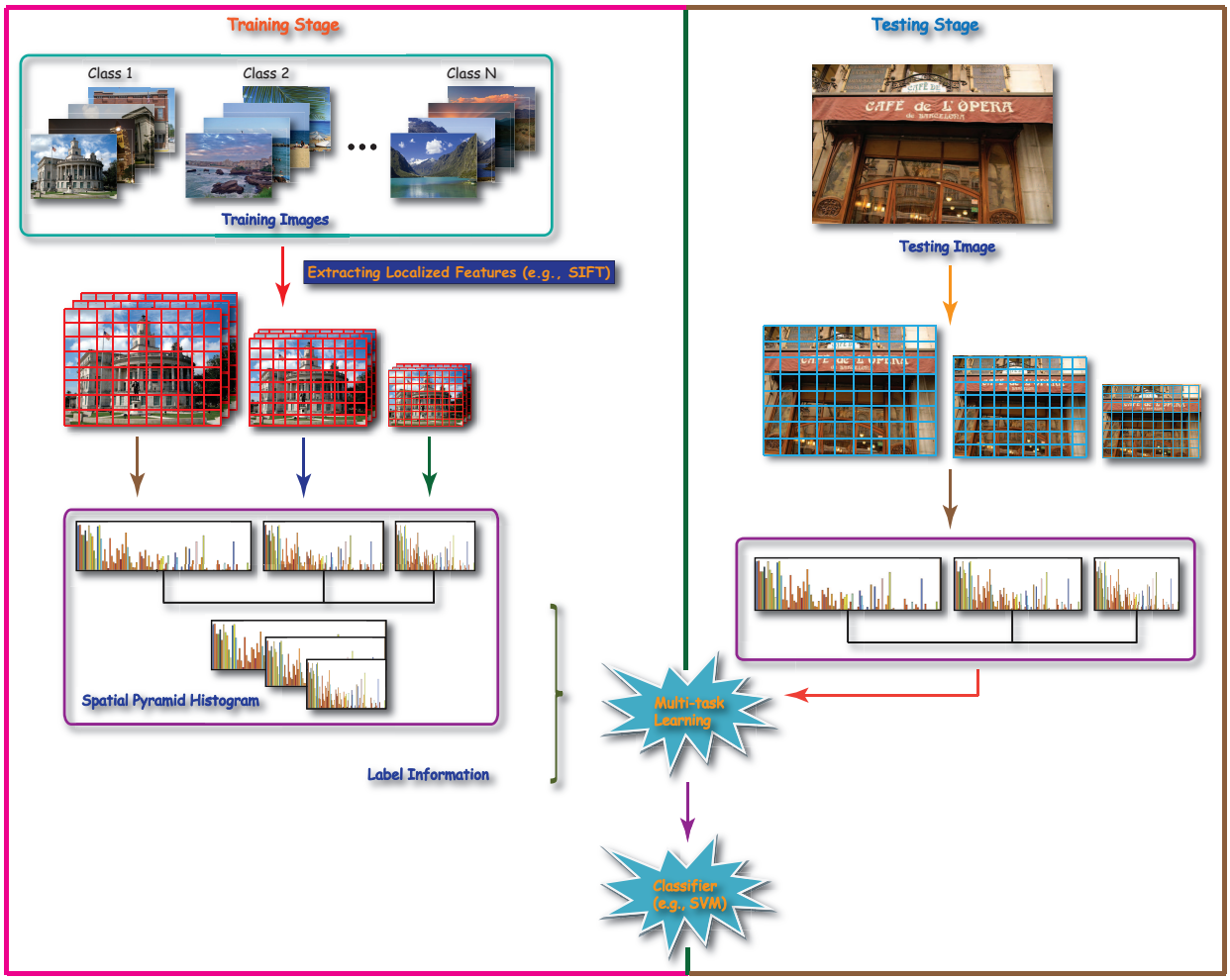


Fig 1: The framework overview.

$X_l = [x_1, \ldots, x_m]$ is composed of $m$ different scene images at $l$-th resolution and $Y_l = [y_1, \ldots, y_m]$ is the corresponding label matrix. We construct different resolution images and extract the corresponding features. And these features at each resolution are regarded as one task. In this case, different tasks are simultaneously learned in a joint framework to improve the performance in scene recognition. Additionally, it is necessary to consider the manifold structure of the training data in multi-task model, which can result in a more faithful result. Hence, the proposed multi-task model can be written as following:

$$\min_{f_l} \sum_{l=1}^{t} \sum_{i=1}^{m} loss(f_l(x_l^i), F_l^i) + g_1(F_l^i), \tag{3}$$

where $loss(f_l(x_l^i), F_l^i)$ is a loss function that measures the consistence between true and predicted labels, the definition of $g_1(F_l^i)$ can be found in Eq.(1), $F_l^i$ is the predicted label of ground-truth labels $y_l^i$, $f_l$ is the $l$-th row of mapping $f$. To share the common knowledge across multi-task and evaluate the informativeness of all features jointly for each task (each resolution), the mapping $f$ in Eq.(3) can be restricted to be sparse. For each task, the objective function in Eq.(3) can be redefined as:

$$\begin{aligned} \min \sum_{l=1}^{t} \sum_{i=1}^{m} \big(loss(f_l(x_l^i), F_l^i) + g_1(F_l^i) + \alpha g_2(f_l)\big) \\ + \beta \Omega(f), \end{aligned} \tag{4}$$

where $g_2(f_l)$ is the $l_{2,1}$-norm regularization function and $\Omega(f)$ is the regularized term, $\alpha$, and $\beta$ are the parameters. Eq.(4) can be expanded as following:

$$\begin{aligned} \min_{W_l, b_l, F_l} \sum_{l=1}^{t} (\|W_l^T X_l + b_l \mathbf{1}_i^T - F_l\|_F^2 \\ + Tr(F_l - Y_l)(F_l - Y_l)^T \\ + Tr(F_l M_l F_l^T) + \alpha \|W_l\|_{2,1}) + \beta \Omega(f), \end{aligned} \tag{5}$$

where $b_l$ is the bias term, $\mathbf{1}_l$ is a vector whose elements are all 1 and $W_l$ is the matrix. Given a matrix $A \in \mathbb{R}^{m \times d}$, we denote the $i$-th row of $A$ by $A_{i\cdot}$. The $l_{2,1}$-norm of $A$ is defined as

$$\|A\|_{2,1} = \sum_{i=1}^{m} \|A_{i\cdot}\|_2, \tag{6}$$

When the $l_{2,1}$ norm is minimized, some rows of matrix $A$ will shrink to 0.

In previous work [41], the common component of multiple tasks can be shared by the low-rank matrix of $W = \{W_l\}_{l=1}^{t}$. It is well known that we can exploit correlations between features and improve predictive accuracy by restricting $W$ to be a low rank matrix. Moreover, When $W$ is low-rank, distances are equivalently computed in a low-dimensional subspace, which allowing for efficient storage and retrieval [42]. In this paper, the objection function of the proposed method is presented as following:

$$\begin{aligned}\min_{W_l,b_l,F_l} \sum_{l=1}^{t} (&\|W_l^T X_l + b_l \mathbf{1}_i^T - F_l\|_F^2 \\ &+ Tr(F_l - Y_l)(F_l - Y_l)^T \\ &+ Tr(F_l M_l F_l^T) + \alpha \|W_l\|_{2,1}) + \beta \|W\|_*,\end{aligned} \tag{7}$$

where $\|W\|_*$ is the low rank term and the minimization of the rank of matrix is non-convex. In this case, we replace the low rank term $\|W\|_*$ with the trace norm term, which is the convex hull of the rank of $W$. Given a matrix A, we denote the trace norm as following:

$$\|A\|_* = Tr\left(AA^T\right)^{\frac{1}{2}}, \tag{8}$$

where $Tr$ is the trace operator. Therefore our objective function in Eq.(7) can be formulated as follows:

$$\begin{aligned}\min_{W_l,b_l,F_l} \sum_{l=1}^{t} (&\|W_l^T X_l + b_l \mathbf{1}_i^T - F_l\|_F^2 \\ &+ Tr(F_l - Y_l)(F_l - Y_l)^T \\ &+ Tr(F_l M_l F_l^T) + \alpha \|W_l\|_{2,1}) + \beta Tr(WW^T).\end{aligned} \tag{9}$$

It can be seen from Eq.(9) that the proposed objective function has two main advantages compared with other scene recognition methods. First, based on multi-task model, the relationships exploited in different resolutions images are beneficial for scene recognition. Second, by using the the $l_{2,1}$ norm term and the trace norm term $\|W\|_*$, our method can reduce the redundant features and obtain better interpretability of the features by sharing the common knowledge across different resolution images.

### *B. Optimization*

In this section, we give an iterative method to optimize the object function. The detailed procedure of the proposed method is described in Algorithm 1. By considering Eq.(8) and Eq.(9), we rewrite the objective function as follow:

$$\begin{aligned}\min_{W_l,b_l,F_l} \sum_{l=1}^{t} (&\|W_l^T X_l + b_l \mathbf{1}_i^T - F_l\|_F^2 \\ &+ \alpha Tr(W_l^T D_l W_l) + Tr(F_l - Y_l)(F_l - Y_l)^T \\ &+ Tr(F_l M_l F_l^T)) + \frac{\beta}{2} Tr(W^T (WW^T)^{-\frac{1}{2}} W),\end{aligned} \tag{10}$$

where $D_l$ is a diagonal matrix which is defined as:

$$D_l = \begin{bmatrix} \frac{1}{2\|w_l^1\|_2} & & \\ & \ddots & \\ & & \frac{1}{2\|w_l^d\|_2} \end{bmatrix} \tag{11}$$

By setting the derivative of Eq.(10) w.r.t. $b_l$ to 0, we get

$$b_l = \frac{1}{m} F_l \mathbf{1}_l - \frac{1}{m} W_l^T X_l \mathbf{1}_l. \tag{12}$$

Substituting Eq.(12) to Eq.(10), we can get

$$\begin{aligned}\min_{W_l,b_l,F_l} \sum_{l=1}^{t} (&\|W_l^T X_l + (\frac{1}{m} F_l \mathbf{1}_l - \frac{1}{m} W_l^T X_l \mathbf{1}_l)\mathbf{1}_i^T - F_l\|_F^2 \\ &+ \alpha Tr(W_l^T D_l W_l) + Tr(F_l - Y_l)(F_l - Y_l)^T \\ &+ Tr(F_l M_l F_l^T)) + \frac{\beta}{2} Tr(W^T (WW^T)^{-\frac{1}{2}} W), \\ \min_{W_l,b_l,F_l} \sum_{l=1}^{t} &\|W_l^T X_l (I_l - \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T) - Y_l (I_l - \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T)\|_F^2 \\ &+ \alpha Tr(W_l^T D_l W_l) + Tr(F_l - Y_l)(F_l - Y_l)^T \\ &+ Tr(F_l M_l F_l^T)) + \frac{\beta}{2} Tr(W^T (WW^T)^{-\frac{1}{2}} W),\end{aligned} \tag{13}$$

where $I$ is an identity matrix. Denote $H_l = I_l - \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T$ as a centering matrix, Eq.(13) can be rewritten as:

$$\begin{aligned}\min_{W_l,b_l,F_l} \sum_{l=1}^{t} &\|W_l^T X_l H_l - F_l H_l)\|_F^2 + \alpha Tr(W_l^T D_l W_l) \\ &+ Tr(F_l - Y_l)(F_l - Y_l)^T + Tr(F_l M_l F_l^T)) \\ &+ \frac{\beta}{2} Tr(W^T (WW^T)^{-\frac{1}{2}} W).\end{aligned} \tag{14}$$

By setting the derivative of Eq.(14) w.r.t. $W_l$ to 0, we can obtain

$$W_l = (X_l H_l H_l^T X_l^T + \alpha D_l + \beta \widetilde{D})^{-1} X_l H_l F_l^T = A_l F_l^T, \tag{15}$$

where $A_l = (X_l H_l H_l^T X_l^T + \alpha D_l + \beta \widetilde{D})^{-1}$. When considering $W_l$ in Eq.(15) and $b_l$ in Eq.(12),Eq.(16) can obtain:

$$\begin{aligned}&W_l^T X_l + b_l \mathbf{1}^T \\ =&(AF_l^T)^T X_l + \frac{1}{m} F_l \mathbf{1}_l \mathbf{1}_l^T - \frac{1}{m} F_l A_l^T X_l \mathbf{1}_l \mathbf{1}_l^T \\ =&F_l A_l^T X_l (I_l - \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T) + \frac{1}{m} F_l \mathbf{1}_l \mathbf{1}_l^T \\ =&F_l (A_l^T X_l H_l + \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T) \\ =&F_l B_l,\end{aligned} \tag{16}$$

where $B_l = A_l^T X_l H_l + \frac{1}{m}\mathbf{1}_l\mathbf{1}_l^T$. When $W$ and $b$ are added into Eq.(10), we have

$$\begin{aligned}F^* =&Tr(F_l B_l - F_l)(F_l B_l - F_l)^T \\ &+ \alpha Tr(A_l F_l^T)^T (A_l F_l^T) \\ &+ Tr(F_l - Y_l)(F_l - Y_l)^T + Tr(F_l M_l F_l^T).\end{aligned} \tag{17}$$

By setting the derivative of Eq.(17) as 0, we can get the prediction label $F$:

$$F_l = Y_l (I_l + M_l + (B_l - I_l)^T (B_l - I_l) + \alpha A_l^T A_l)^{-1}. \tag{18}$$

Once the prediction label is obtained, the matrix $W$ in Eq.(15) can be obtained. The detailed procedure of the proposed method can be listed in Algorithm 1.

**Algorithm 1** Sparse Feature Selection based Manifold Regularized Method

**Require:**
Input data $X_l \in \Re^{d\times m}\,(1 \le l \le t)$ and labels $Y_l \in \Re^{c\times m}$;
Laplace matrix $M_l$ (Computed by $X_l$);
Regularization parameters $\alpha$ and $\beta$.

**Ensure:**
Feature selection matrix $W_l\,|_{l=1}^{t}$.

1: **Initialize** Set $r = 0$, initialize $W_l\,|_{l=1}^{t}$ randomly, $W^0 = [W_1, ..., W_t]$;
2: **Repeat** until convergence:
3: $l = 1$;
4: **Repeat** until convergence:
5: Compute $D_l^r$, $\tilde{D}^r$, $H_l$, $A_l$, $B_l$, $F_l$;
6: Update $W_l^r$:
$W_l^r = A_l F_l^T$;
7: Update $b_l$:
$b_l = (1/m)\left(F_l 1_l - W_l^T X_l I_l\right)$;
8: $l = l + 1$;
9: Checking the convergence conditions:
$l > t$
10: $W^{r+1} = [W_{1,...,}W_t]$;
11: $r = r + 1$;
12: Checking the convergence conditions;
13: **Return** $W_l$ and $b_l$ for $1 \le l \le t$.

## IV. EXPERIMENTS

In this section, we will introduce the datasets and conduct several experiments to evaluate the performance of the proposed method. The experimental results demonstrate the efficacy of the proposed method.

### A. Datasets

The proposed method will be evaluated over four commonly used datasets, which are described as follows:

1. The first dataset is 8-category scenes dataset, namely OT dataset, which is proposed by Oliva and Torralba [2] in MIT. This dataset includes 2688 images, divided into 8 categories.

2. The second dataset is 15-category scenes, namely LS dataset, which are proposed by Lazbnik *et al*. [22]. This dataset includes 4688 images, divided into 15 categories of which 2688 images are from Olivas 8-scene dataset: coast (360 images), forest(328 images), mountain (274 images), open country (410 images), highway (260 images), inside city (308 images), tall building (365 images), street (292 images), bedroom (216 images), kitchen (210 images), livingroom (289 images), PARoffice (215 images), Calsuburb (241 images), industrial (311 images) and store (315 images).

3. The third dataset is 8-category sports events, namely LF dataset, which is proposed by Li-Jia and Fei-Fei [43]. LF dataset includes 1597 images and contains 8 categories: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images) and rock climbing (194 images).

4. The fourth data is SUN dataset containing a large variety of environmental scenes, places and the objects. Sun dataset is a larger and more challenging scene recognition dataset which provides a comprehensive collection of annotated images. The SUN dataset contains a full variety of 899 scene categories. In this paper, the experiment is conducted on the well-known SUN-397 dataset provided by Xiao *et al*. [30], which sampled from the whole SUN dataset for scene recognition.

### B. Implementation

We perform extensive experiments on four popular, benchmark datasets by cross-validation. Fig. 2 shows several samples from the aforementioned datasets. The prevalent training/testing configurations are followed in the literature. For each image in the aforementioned datasets, we generate the corresponding different resolution images by downsampling in our experiments. The different resolution images can be generated by downsampling the input image by a factor of 2 and 4, respectively. As shown in Fig. 3, the ratio of different resolutions is 1:4:16. In this case, each group is composed of three different resolution images from the same scene. In the aforementioned datasets. In each group, other two resolution images in each group can be generated using downsampling the high resolution image by a factor of 2 and 4, respectively. The appearance representation is based on SIFT descriptors extracted over 448 bins from 99 patches. And all experiments involving spatial pyramids relied on three pyramid levels aim to consider the spatial structure of images. For 15 scene categories dataset, each category of scenes is divided into two separate sets of images, $N$ groups for training and the rest groups for testing (Here, $N$=100). In training dataset, each group contains three different resolution images. Similarly, for 8 sports event categories dataset, 70 randomly selected groups are exploited for training and 60 are used for testing. For each category in SUN dataset, 50 groups are exploited to train the classifier and the other 50 groups are utilized to evaluate the performance of the feature. In this paper, multi-class classification is done with support vector machine (SVM) classifier using the one-versus-all rule. The rule can be defined that a classifier is learned to separate each class from the rest, and a test group is assigned the label of the classifier with the highest response. The LIBLINEAR is used to train a linear SVM and switched from one-versus-one to one-versus-all multi-class classification.

In the following, the experiments results on three datasets, including LS, OT and LF datasets, can be conducted under *multi-resolutions* and *single resolution* conditions. In this paper, there exist two free parameters: $\alpha$ and $\beta$. It is worth mentioning that the parameters ($\alpha$ and $\beta$) have few influence on the results from many experiments and we define the parameters $\alpha = 0.1$ and $\beta = 0.1$ in our experiments. We demonstrate and validate our method to verify the effectiveness of the proposed method under *multi-resolutions* and *single resolution* conditions. Fig. 4, Fig. 6 and Fig. 8 show the accuracy of recognition on different dataset under *multi-resolutions* and *single resolution* conditions. It can be shown from Fig. 4, Fig. 6 and Fig. 8 that the accuracy of recognition under *multi-resolutions* condition is high compared with the accuracy of recognition under *single resolution* conditions.

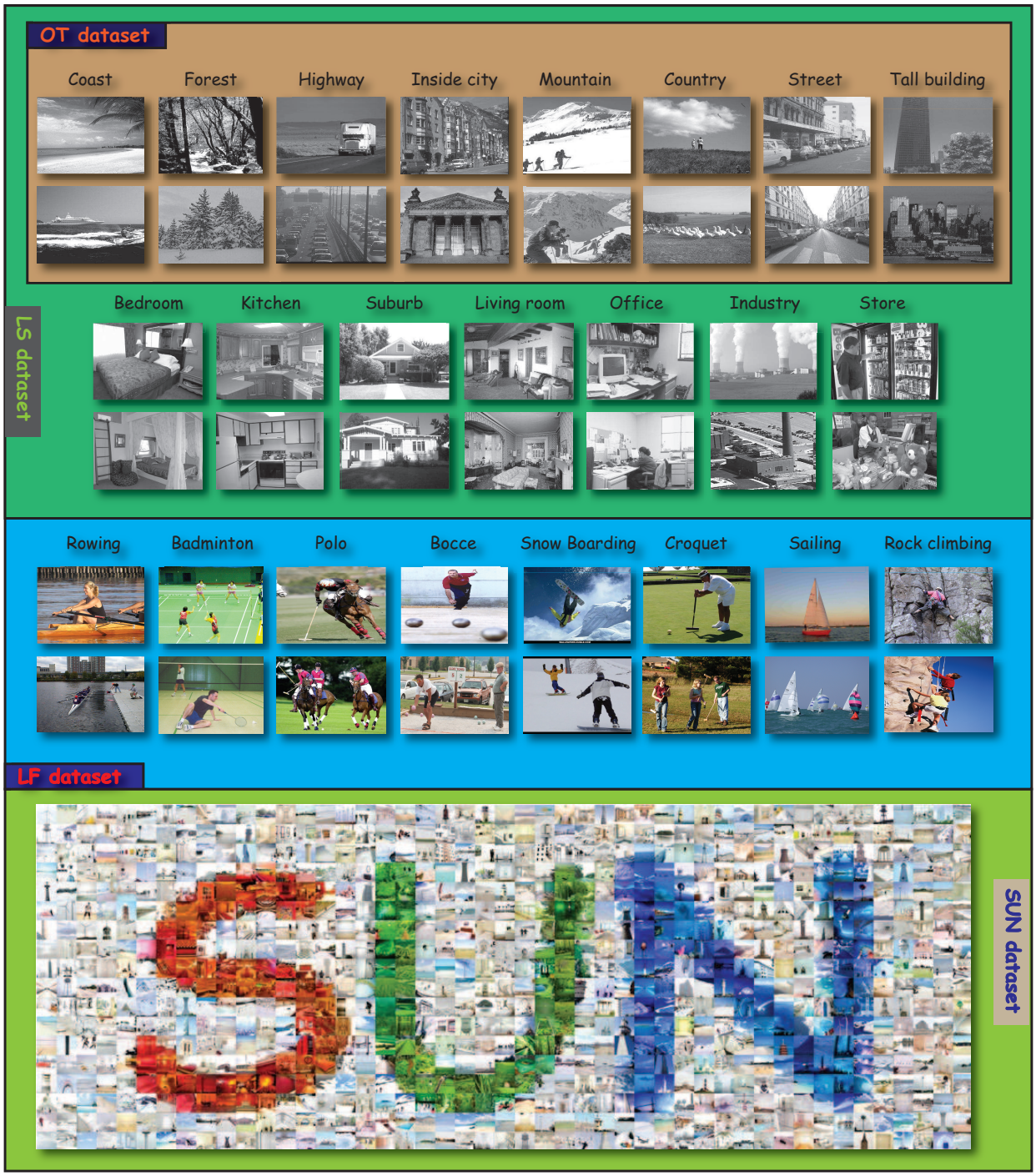


Fig 2: Sample images from OT, LF, LS datasets and SUN dataset.

Fig. 5, Fig. 7 and Fig. 9 show different confusion matrix of the annotation results over different dataset under the optimal condition (feature number is 3000).

### C. Comparing to the State-of-the-Art

In this section, we compare the proposed method with the state-of-the-art on three standard datasets respectively. First, the performance of some methods can be observed on the 15 scene categories dataset (LS dataset). The outstanding spatial pyramid matching was provided by Lazebnik *et al.* [22]. In [22], image was partitioned into the fine sub-regions and the histograms of local features within each sub-region are computed. In this case, the recognition accuracy of 81.4% can be obtained in [22]. Dixit *et al.* [44] presented a general formulation of Bayesian adaptation, which targeted class adaptation and exploited both the generative and discriminative strategies for the task of image classification. Kwittet *et al.* [45] proposed a new architecture, denoted spatial pyramid matching on the semantic manifold (SPMSM) for scene recognition and made a correct recognition rate of 85.4%


Fig 3: Illustration of multi-resolution image.

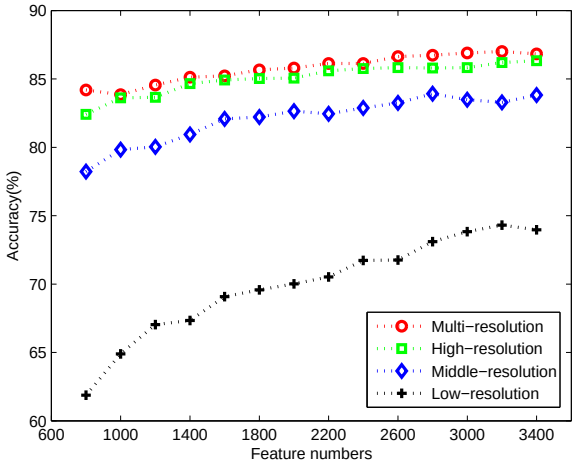


Fig 4: Accuracy with different feature numbers for the LS dataset.

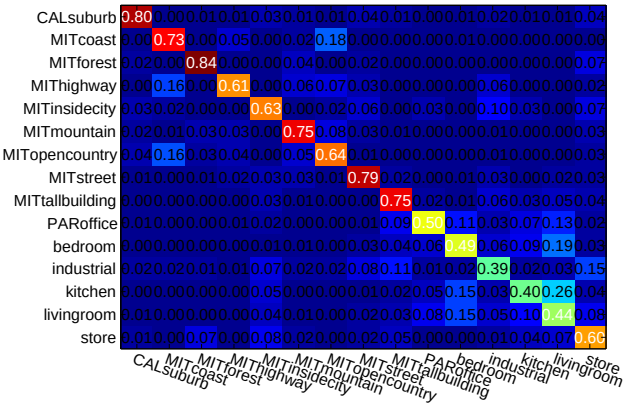


Fig 5: Confusion table for the LS dataset.

on the 15 scene categories dataset. SPMSM established a connection between the semantic simplex and a Riemmanian manifold, and achieved similarity measure for the manifold structure of the semantic space. Dixit *et al*. made a recognition accuracy of 82.3% on the 15 scene categories dataset by SPMSM. To verify the performance of the proposed method, the obtained features are adopted respectively as the learned features to run the SVM classifier for the task of scene recognition. As shown in Table I, the average performance of the proposed method over the LS dataset is significantly improved to 3.5% compared with other methods. Similarly, it can be verified by comparing with the state-of-the-art on the 8 sports event categories dataset(OT dataset). Socher *et al*. [14] introduced a recursive neural network architecture which can successfully merge image segments or natural language words based on deep learned semantic transformations of their original features. By using the architecture, they show that their system is capable of recognition these event categories at 87.8% accuracy. The work of Wu and Rehg [46] shown that the Histogram intersection Kernel (HIK) is either more effective than the Euclidean distance in supervised learning tasks with histogram features or is used in an unsupervised manner to significantly improve the generation of visual codebooks. The HIK method has consistently higher accuracy than k-means codebooks by 2-4%, and can achieve the 84.3% accuracy on the 8 sports event categories dataset. Kwittet *et al*. [45] also have tested their method on these sports event dataset and made a recognition accuracy of 83.0%. As shown in TTable I, Compared with the other two methods, our proposed method achieves the highest average value of recognition rate on LF-8 dataset(=94.2It can be seen in Table I that the proposed method outperforms other state-of-the-art method. In general, the comparable performance between the proposed method and other methods is due to the proposed method's capability to better learn and interpret the effective feature. Finally, in order to further evaluate the performance of the feature learned using the proposed method on the large dataset, our method is compared with other methods on the SUN dataset. The performance of the proposed method outperforms other state-of-the-art methods, as shown in Table I.

## V. CONCLUSIONS

In this paper, a new scheme for scene recognition utilizing multiple resolution information is reported. we combine a sparse feature selection method and a manifold regularized learning framework, called sparse feature selection based manifold regularized method (SFSMR), to select the optimal

Table I: Comparison to the state-of-the-art on dataset.

| Methods | LS-15 | OT-8 | LF-8 | SUN |
|---|---|---|---|---|
| Lazebnik *et al.* [22] | 81.4% | – | – | – |
| Nakayam *et al.* [47] | 86.1% | – | – | – |
| Bosch *et al.* [12] | 83.7% | 87.8% | – | – |
| Dixitet *et al.* [44] | 82.3% | 84.7% | – | – |
| Wu *et al.* [46] | 83.9% | 86.2% | 84.3% | – |
| Kwittet *et al.* [45] | 85.4% | – | 83.0% | 28.9% |
| **Our approach** | **87.3%** | **95.6%** | **94.2%** | **30.5%** |

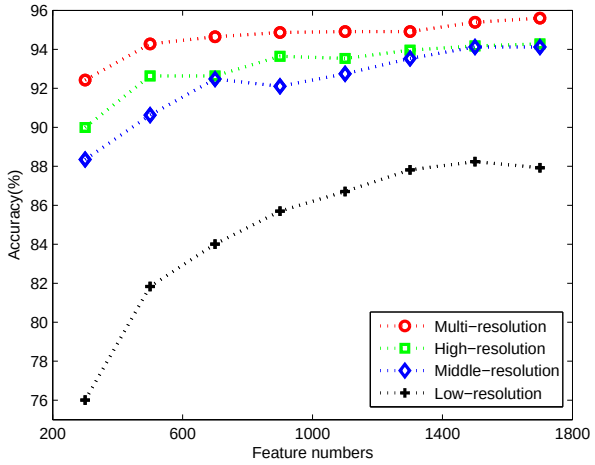


Fig 6: Accuracy with different feature numbers for the OT dataset.

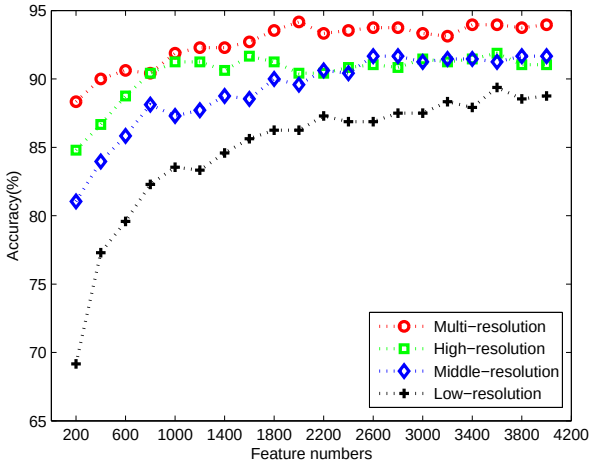


Fig 8: Accuracy with different feature numbers for the LF dataset.

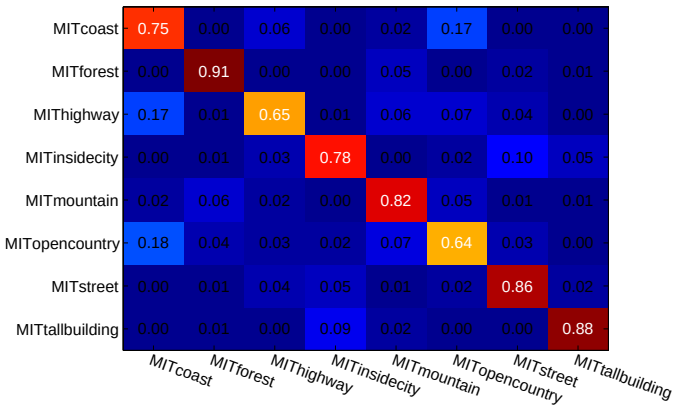


Fig 7: Confusion table for the OT dataset.

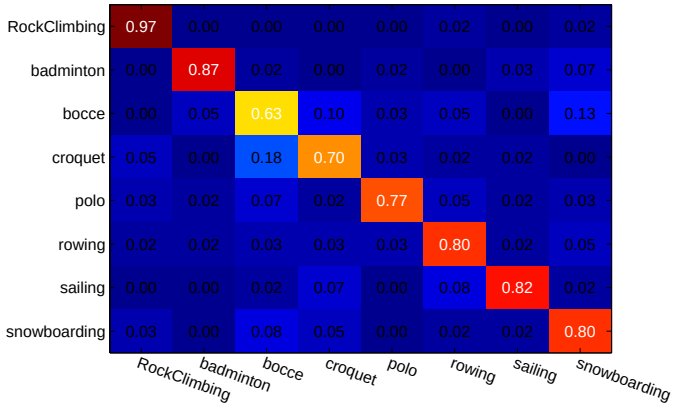


Fig 9: Confusion table for the LF dataset.

information and preserve the underlying manifold structure of data. Finally, by integrating the multi-task model and SFSMR, the proposed semi-supervised learning method can further improve the accuracy of scene recognition.

## Acknowledgement

The authors would like to thank P. Cao for helpful discussions on this manuscript.

## References

[1] Y. Huang, K. Huang, D. Tao, T. Tan, and X. Li, "Enhanced biologically inspired model for object recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 6, pp. 1668–1680, 2011.

[2] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[3] G. Schneider, H. Wersing, B. Sendhoff, and E. Körner, "Evolutionary optimization of a hierarchical object recognition model," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 3, pp. 426–437, 2005.

[4] Y. Lin and B. Bhanu, "Object detection via feature synthesis using mdl-based genetic programming," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 3, pp. 538–547, 2005.

[5] L. Zhang, L. Wang, and W. Lin, "Generalized biased discriminant analysis for content-based image retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 1, pp. 282–290, 2012.

[6] D. Tao, L. Jin, Z. Yang, and X. Li, "Rank preserving sparse learning for kinect based scene classification," *IEEE T. Cybernetics*, vol. 43, no. 5, pp. 1406–1417, 2013.

[7] S. Zhang, J. Huang, H. Li, and D. N. Metaxas, "Automatic image annotation and retrieval using group sparsity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 42, no. 3, pp. 838–849, 2012.

[8] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 702–712, 2006.

[9] A. Bishnu, B. B. Bhattacharya, M. K. Kundu, C. A. Murthy, and T. Acharya, "Euler vector for search and retrieval of gray-tone images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 4, pp. 801–812, 2005.

[10] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR (2)*, 2005, pp. 524–531.

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars,

"A thousand words in a scene," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1575–1589, 2007.

[12] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.

[13] J. Wu and J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *ICCV*, 2009, pp. 630–637.

[14] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *ICML*, 2011, pp. 129–136.

[15] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular visual scene understanding: Understanding multi-object traffic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 882–897, 2013.

[16] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *CVPR*, 2012, pp. 702–709.

[17] R. Eckhorn, "Neural mechanisms of scene segmentation: recordings from the visual cortex suggest basic circuits for linking field models," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 464–479, 1999.

[18] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *CVPR*, 2009, pp. 1972–1979.

[19] X. Yu, C. Fermüller, C. L. Teo, Y. Yang, and Y. Aloimonos, "Active scene recognition with vision and language," in *ICCV*, 2011, pp. 810–817.

[20] X. Yu, C. Fermuller, C. Teo, Y. Yang, and Y. Aloimonos, "Active scene recognition with vision and language," in *ICCV*, 2011, pp. 810–817.

[21] O. Boiman, E. Shechtman, and M. Irani, "Ain defense of nearest-neighbor based image classification," in *CVPR*, 2008.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.

[23] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999, pp. 289–296.

[24] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] M. S. L. Torresani and A. Fitzgibbon, "Efficient object category recognition using classiness," in *ECCV*, 2010.

[26] D. Parikh and K. Grauman, "Interactively building a discriminative vocabulary of nameable attributes," in *CVPR*, 2011.

[27] S. Wang, J. Joo, Y. Wang, and S. C. Zhu, "Weakly supervised learning for attribute localization in outdoor scenes," in *CVPR*, 2013, pp. 3111–3118.

[28] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in *CVPR*, 2012, pp. 2751–2758.

[29] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*, 2009, pp. 413–420.

[30] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010, pp. 3485–3492.

[31] D.Zhou, O.Bousquet, T. Lal, J.Weston, and B. Scholkopf, "Learning with local and global consistency," in *NIPS*, 2004, pp. 321–328.

[32] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 24, no. 12, pp. 2013–2026, 2013.

[33] K. Ni, J. W. Paisley, L. Carin, and D. B. Dunson, "Multi-task learning for analyzing and sorting large databases of sequential data," *IEEE Transactions on Signal Processing*, vol. 56, no. 8-2, pp. 3918–3931, 2008.

[34] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 5, pp. 623–636, 2011.

[35] S. Parameswaran and K. Q. Weinberger, "Large margin multi-task metric learning," in *NIPS*, 2010, pp. 1867–1875.

[36] X.Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *CVPR*, 2010, pp. 3493–3500.

[37] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An efficient projection for $l1$,infinity regularization," in *ICML*, 2009.

[38] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *CVPR*, 2012.

[39] R. Caruana, "Multi-task learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[40] Q. Liu, X. Liao, H. Li, J. Stack, and L. Carin, "Semisupervised multitask learning," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 1074–1086, 2009.

[41] Z. Ma, Y.Yan, F. Nie, J. Uijlings, and N.Sebe, "Exploiting the entire features space with sparsity for automatic image annotation," in *in Proc. ACM Multimedia*, 2011.

[42] D. Lim, B. McFee, and G. Lanckriet, "Robust structural metric learning," in *ICML*, 2013, pp. 615–623.

[43] J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *ICCV*, 2007.

[44] M. Dixit, N. Rasiwasia, and N. Vasconcelos, "Adapted gaussian models for image classification," in *CVPR*, 2011, pp. 937–943.

[45] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *ECCV*, 2012, pp. 359–372.

[46] J. Wu and J. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *ICCV*, 2009.

[47] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Global gaussian approach for scene categorization using information geometry," in *CVPR*, 2010, pp. 2336–2343.