Heterogeneous Visual Features Fusion via Sparse Multimodal Machine

Hua Wang[†], Feiping Nie[‡], Heng Huang[‡]*, Chris Ding[‡]
[†]Department of Electrical Engineering and Computer Science
Colorado School of Mines, Golden, Colorado 80401, USA
[‡]Department of Computer Science and Engineering
University of Texas at Arlington, Arlington, Texas 76019, USA
huawangcs@gmail.com, feipingnie@gmail.com, heng@uta.edu, chqding@uta.edu

Abstract

To better understand, search, and classify image and video information, many visual feature descriptors have been proposed to describe elementary visual characteristics, such as the shape, the color, the texture, etc. *How to integrate these heterogeneous visual features and identify the important ones from them for specific vision tasks has become an increasingly critical problem. In this paper, We propose a novel Sparse Multimodal Learning (SMML) approach to integrate such heterogeneous features by using the joint structured sparsity regularizations to learn the feature importance of for the vision tasks from both group-wise and individual point of views. A new optimization algorithm is also introduced to solve the non-smooth objective with rigorously proved global convergence. We applied our SMML method to five broadly used object categorization and scene understanding image data sets for both single-label and multi-label image classification tasks. For each data set we integrate six different types of popularly used image features. Compared to existing scene and object categorization methods using either single modality or multi-modalities of features, our approach always achieves better performances measured.*

1. Introduction

Scene categorization and visual recognition problem analyzes and classifies the images into semantically meaningful categories. It is without doubts a difficult task in computer vision research field, because any scene/object category can be characterized by a high degree of diversity and potential ambiguities. To enhance the visual understanding, computer vision researchers have proposed many feature representation methods to describe the visual objects in different types of images [12, 18, 3]. However, it is usually not clear what the best feature descriptor to solve a given application problem is. Because different features describe different aspects of the visual characteristics, one descriptor can be better than others under certain circumstances.

How to integrate heterogeneous features is becoming a challenging as well as attractive problem nowadays. Considering different feature representations give rise to different kernel functions, the Multiple Kernel Learning (MKL) approaches [19, 9, 13] have been recently studied and employed to integrate heterogenous features or data and select multi-modal features. Particularly, [5] surveyed several MKL methods, as well as their variants using boosting method, for computer vision tasks and applied them in object classification. However, such models train a weight for each type of features, *i.e.*, when multiple types of features are combined together, all features from the same type are weighted equally. This crucial drawback often causes the low performance.

To address the integration of the heterogeneous image features for visual recognition tasks, in this paper, we propose a novel Sparse Multimodal Learning (SMML) method by utilizing a new mixed structured sparsity norms. In our method, we concatenate all features of one image together as its feature vector and learn the weight of each feature in the classification decision functions. The main difficulty of integrating heterogeneous image features is to simultaneously consider the feature group property (*e.g.* GIST is good at detecting natural scenes and the GIST features should have large weights for classes related to outdoor natural scenes) and individual feature property, *i.e.*, although a type of features are not useful to categorize certain specific classes, a small number of individual features from the same type can still be discriminative for these classes.

We propose to utilize two structured sparsity regularizers to capture both group and individual properties of different modalities (types) of features. Meanwhile the sparse weight matrix provides a natural feature selection results. Our new

*Corresponding author. This work was partially supported by NSF IIS-1117965

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.398

objective, employing the hinge loss and the two non-smooth regularizers, is highly non-smooth and difficult to solve in general. Thus, we derive a new efficient algorithm to solve it with rigorously proved global convergence. We applied our new sparse multimodal learning method to five broadly used object categorization and scene understanding image data sets for both single-label and multi-label image classification tasks. For each data set we integrate six different types of popularly used image features. In all experimental results, our new method always achieves a better object and scene categorization performance than traditional classification methods utilizing each single type descriptor and the widely used MKL based feature integration methods.

2. Sparse Multimodal Learning

In recent research, sparse regularizations have been widely investigated and applied into different computer vision and machine learning studies [1, 16, 11]. The sparse representations are typically achieved by imposing non-smooth norms as regularizers in the optimization problems. Because the structured sparsity regularizers can capture the structures existing in data points and features, they are useful in discovering the underlying patterns. We will use a new joint structured sparsity regularizers to explore both group-wise and individual importance of each feature for different classes.

2.1. Joint Structured Sparsity Regularizations

In the supervised learning setting, we are given n training images $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, where $\mathbf{x}_i = \left[\left(\mathbf{x}_i^1\right)^T, \cdots, \left(\mathbf{x}_i^k\right)^T\right]^T \in \Re^d$ is the input vector including all features from a total of k modalities and each modality j has d_j features ($d = \sum_{j=1}^k d_j$). $\mathbf{y}_i \in \Re^c$ is the class label vector of data point $\mathbf{x}_i$, where c is the number of classes. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_c] \in \Re^{c \times n}$. In this paper, we write matrices as boldface uppercase letters and vectors as boldface lowercase letters. For matrix $\mathbf{W} = (w_{ij})$, its i-th row and j-th column are denoted by $\mathbf{w}^i$ and $\mathbf{w}_j$ respectively.

Different to MKL, we directly learn a $d \times c$ feature weight matrix $\mathbf{W} = [\mathbf{w}_1^1, \cdots, \mathbf{w}_c^1; \cdots, \cdots, \cdots; \mathbf{w}_1^k, \cdots, \mathbf{w}_c^k] \in \Re^{d \times c}$, where $\mathbf{w}_p^q \in \Re^{d_q}$ indicates the weights of all features from the q-th modality in the classification decision function of the p-th class. Typically we can use a convex loss function $\mathcal{L}(\mathbf{X}, \mathbf{W})$ to measure the loss incurred by $\mathbf{W}$ on the training images. We choose to use the hinge loss, because the hinge loss based SVM has shown state-of-the-art performance in classifications.

Compared to MKL approaches that learn weight matrices in unsupervised way, our method will learn all weights of features using supervised learning model. Because the heterogeneous features come from different visual descriptors, we cannot only use the loss function that equivalents to assign the uniform weights to all features. The regularizer $\mathcal{R}$ has to be added to impose the interrelationships of modalities and features as:

$$\min_{\mathbf{W}} \sum_{i=1}^{c} \sum_{j=1}^{n} \left(1 - y_{ji}(\mathbf{w}_i^T \mathbf{x}_j + b_i)\right)_+ + \gamma \mathcal{R}, \tag{1}$$

where $\gamma > 0$ is a trade-off parameter and the function $(a)_+$ is defined as $(a)_+ = \max(0, a)$. For brevity, we denote $f_i(\mathbf{w}_i, b_i) = \sum_{j=1}^{n} \left(1 - y_{ji}\left(\mathbf{w}_i^T \mathbf{x}_j + b_i\right)\right)_+$ in the sequel.

In heterogeneous features fusion, from multimodal viewpoint, the features of a specific modality can be more or less discriminative for specific classes. For instance, the color features substantially increases the detection of stop signs while they are almost irrelevant for finding cars in images. Thus, we propose a new group ℓ_1-norm (G_1-norm) as regularizer in Eq. (1), which is defined as $\|\mathbf{W}\|_{G_1} = \sum_{i=1}^{c} \sum_{j=1}^{k} ||\mathbf{w}_i^j||_2$ [17]. Then Eq. (1) becomes:

$$\min_{\mathbf{W}} \sum_{i=1}^{c} f_i(\mathbf{w}_i, b_i) + \gamma_1 \|\mathbf{W}\|_{G_1}. \tag{2}$$

Because the group ℓ_1-norm uses ℓ_2-norm within each modality and ℓ_1-norm between modalities, it enforces the sparsity between different modalities, *i.e.*, if one modality of features are not discriminative for certain tasks, the objective in Eq. (2) will assign zeros (in ideal case, usually they are very small values) to them for corresponding tasks; otherwise, their weights are large. This group ℓ_1-norm regularizer captures the global relationships between modalities.

However, in certain cases, even if most features in one modality are not discriminative for certain visual categories, a small number of features from the same modality can still be highly discriminative. From the multi-task learning point of view, such important features should be shared by all tasks. Thus, we add one more $\ell_{2,1}$-norm regularizer into Eq. (2) and get the final objective as:

$$\min_{\mathbf{W}} \sum_{i=1}^{c} f_i(\mathbf{w}_i, b_i) + 2\gamma_1 \|\mathbf{W}\|_{G_1} + 2\gamma_2 \|\mathbf{W}\|_{2,1}. \tag{3}$$

Because the $\ell_{2,1}$-norm regularizer imposes the sparsity between all features and non-sparsity between classes, the features that are discriminative for all classes will get large weights. Because of the imposed sparsity, the weights of most features are close to zeroes and only the features important to classification tasks have large weights. Thus, our SMML results automatically perform the feature selection procedure.

2.2. A New Efficient Optimization Algorithm

The objective of Eq. (3) is a highly non-smooth problem and cannot be easily solved in general. Thus, we derive a new efficient algorithm to solve this problem as summarized in Algorithm 1, whose convergence to the global optimum is guaranteed by the following theorem.

Theorem 1 *In Algorithm 1, the value of the objective in Eq.* (3) *is monotonically decreased in each iteration.*

Proof: In each iteration t, according to the Step 3 in the Algorithm 1, we have:

$$\begin{aligned} \mathbf{W}_{t+1} = \min_{\mathbf{W}} \sum_{i=1}^{c} f_i(\mathbf{w}_i, b_i) \\ + \gamma_1 \sum_{i=1}^{c} \mathbf{D}_t^i \left\| \mathbf{w}_i \right\|_2^2 + \gamma_2 \, \mathbf{tr} \left(\mathbf{W}^T \tilde{\mathbf{D}}_t \mathbf{W} \right), \end{aligned} \tag{4}$$

by which we can derive

$$\begin{aligned} & \sum_{i=1}^{c} f_i((\mathbf{w}_{t+1})_i, (b_{t+1})_i) \\ & + \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left\| (\mathbf{w}_{t+1})_i^j \right\|_2^2}{2 \left\| (\mathbf{w}_t)_i^j \right\|_2} + \gamma_2 \sum_{i=1}^{d} \frac{\left\| \mathbf{w}_{t+1}^i \right\|_2^2}{2 \| \mathbf{w}_t^i \|_2} \\ \leq & \sum_{i=1}^{c} f_i((\mathbf{w}_t)_i, (b_t)_i) \\ & + \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left\| (\mathbf{w}_t)_i^j \right\|_2^2}{2 \left\| (\mathbf{w}_t)_i^j \right\|_2} + \gamma_2 \sum_{i=1}^{d} \frac{\left\| \mathbf{w}_t^i \right\|_2^2}{2 \| \mathbf{w}_t^i \|_2}. \end{aligned} \tag{5}$$

Because it can verified that for function $g(x) = x - \frac{x^2}{2\alpha}$, given any $x \neq \alpha \in \Re^n$, $g(x) \leq g(\alpha)$ holds, we can derive:

$$\begin{aligned} & \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \left\| (\mathbf{w}_{t+1})_i^j \right\|_2 - \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left\| (\mathbf{w}_{t+1})_i^j \right\|_2^2}{2 \left\| (\mathbf{w}_t)_i^j \right\|_2} \\ \leq & \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \left\| (\mathbf{w}_t)_i^j \right\|_2 - \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \frac{\left\| (\mathbf{w}_t)_i^j \right\|_2^2}{2 \left\| (\mathbf{w}_t)_i^j \right\|_2}; \end{aligned} \tag{6}$$

$$\begin{aligned} \gamma_2 \sum_{i=1}^{d} \left\| \mathbf{w}_{t+1}^i \right\|_2 - \gamma_2 \sum_{i=1}^{d} \frac{\left\| \mathbf{w}_{t+1}^i \right\|_2^2}{2 \left\| \mathbf{w}_t^i \right\|_2} \leq \\ \gamma_2 \sum_{i=1}^{d} \left\| \mathbf{w}_t^i \right\|_2 - \gamma_2 \sum_{i=1}^{d} \frac{\left\| \mathbf{w}_t^i \right\|_2^2}{2 \left\| \mathbf{w}_t^i \right\|_2}. \end{aligned} \tag{7}$$

Adding Eqs. (5)-(7) in both sides, we have

$$\begin{aligned} & \sum_{i=1}^{c} f_i((\mathbf{w}_{t+1})_i, (b_{t+1})_i) + \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \left\| (\mathbf{w}_{t+1})_i^j \right\|_2 + \gamma_2 \sum_{i=1}^{d} \left\| \mathbf{w}_{t+1}^i \right\|_2 \\ \leq & \sum_{i=1}^{c} f_i((\mathbf{w}_t)_i, (b_t)_i) + \gamma_1 \sum_{i=1}^{c} \sum_{j=1}^{k} \left\| (\mathbf{w}_t)_i^j \right\|_2 + \gamma_2 \sum_{i=1}^{d} \left\| \mathbf{w}_t^i \right\|_2. \end{aligned} \tag{8}$$

Therefore, the algorithm decreases the objective value in each iteration. □

Because the problem (3) is a convex problem, Algorithm 1 will converge to the global optimum.

```
Input: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$,
       $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_c] \in \Re^{n \times c}$.
1. Let $t = 1$. Initialize $\mathbf{W}_t \in \Re^{d \times c}$.
while Not converges do
    2. Calculate the block diagonal matrices $\mathbf{D}_t^i (1 \leq i \leq c)$, where the $j$-th diagonal block of $\mathbf{D}_t^i$ is $\frac{1}{2 \| (\mathbf{w}_t)_i^j \|_2} \mathbf{I}_j$. Calculate the diagonal matrix $\tilde{\mathbf{D}}_t$, where the $i$-th diagonal element is $\frac{1}{2 \| \mathbf{w}_t^i \|_2}$.
    3. For each $\mathbf{w}_i (1 \leq i \leq c)$, calculate $(\mathbf{w}_{t+1})_i = \mathbf{D}^{-\frac{1}{2}} (\tilde{\mathbf{w}}_t)_i$, where $\tilde{\mathbf{w}}_i = \arg \min_{\tilde{\mathbf{w}}_i} f_i(\tilde{\mathbf{w}}_i, b_i; \tilde{\mathbf{X}}) + \tilde{\mathbf{w}}_i^T \tilde{\mathbf{w}}_i$, $\tilde{\mathbf{X}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{X}$ and $\mathbf{D} = \gamma_1 \mathbf{D}^i + \gamma_2 \tilde{\mathbf{D}}$.
    4. $t = t + 1$.
end
Output: $\mathbf{W}_t \in \Re^{d \times c}$.
```

Algorithm 1: An efficient iterative algorithm to solve the optimization problem in Eq. (3).

3. Experimental Results

In this section, we experimentally evaluate the proposed Sparse Multi-Modal Learning (SMML) approach in both single-label image classification tasks and multi-label image classification tasks.

3.1. Evaluation in Single-Label Image Classification

We first evaluate the proposed approach in single-label image classification, in which each image belongs to one and only one class. We experiment with the following three benchmark single-label multi-modal image data sets, which are broadly used computer vision studies.

NUS-WIDE-Object data set[1] contains 30,000 images and 31 classes. Following [4], we select to use a subset of 26 classes in our experiments.

Animal data set[2] contains 30457 images for 50 animals (classes).

MSRC-v1 data set[3] contains 240 images with 9 classes. Following [2], we refine the data set to get 7 classes including tree, building, airplane, cow, face, car, bicycle, and each refined class has 30 images.

All the three data sets are described by a set of 6 different image descriptors, which are listed in Table 1.

Experimental setups. We classify the images in the above three data sets using the proposed methods by integrating the six types of image features of each of them. We compare the proposed method against several most recent multiple

[1] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

[2] http://attributes.kyb.tuebingen.mpg.de/

[3] http://research.microsoft.com/en-us/projects/objectclassrecognition/

Table 1. Image feature descriptors of image data sets used in experiments.

Type	NUS-WIDE-Object		Animal		MSRC-v1 / MSRC-v2 / TRECVID 2005	
	Features	Dimension	Features	Dimension	Features	Dimension
1	Color moments	255	Self-Similarity	2000	Color moment	48
2	Color histogram	64	Color histogram	2688	LBP	256
3	Color correlogram	144	PyramidHOG	252	HOG	100
4	Wavelet texture	128	SIFT	2000	SIFT	1230
5	Edge distribution	73	colorSIFT	2000	GIST	512
6	Visual words	500	SURF	2000	Centrist	1320

Table 2. Classification accuracies (mean ± std) of the compared methods in three single-label image classification tasks.

Methods	NUS-WIDE	Animal	MSRC-v1
SVM (Type 1)	0.152±0.018	0.542±0.016	0.777±0.019
SVM (Type 2)	0.149±0.020	0.551±0.019	0.768±0.018
SVM (Type 3)	0.146±0.016	0.569±0.021	0.781±0.022
SVM (Type 4)	0.150±0.018	0.541±0.023	0.784±0.026
SVM (Type 5)	0.141±0.017	0.566±0.021	0.773±0.023
SVM (Type 6)	0.149±0.018	0.554±0.200	0.789±0.021
SVM (all by concatenation)	0.138±0.020	0.547±0.019	0.793±0.025
SVM ℓ_∞ MKL method	0.211±0.023	0.603±0.017	0.820±0.023
SVM ℓ_1 MKL method	0.207±0.020	0.599±0.019	0.813±0.019
SVM ℓ_2 MKL method	0.202±0.021	0.593±0.018	0.789±0.022
LSSVM ℓ_∞ MKL method	0.200±0.018	0.588±0.025	0.778±0.025
LSSVM ℓ_1 MKL method	0.195±0.022	0.586±0.023	0.808±0.027
LSSVM ℓ_2 MKL method	0.187±0.021	0.578±0.019	0.796±0.018
GP method	0.181±0.020	0.569±0.022	0.794±0.015
LPboost-β	0.220±0.015	0.612±0.011	0.815±0.010
LPboost-B	0.219±0.012	0.610±0.014	0.813±0.013
Our method (G_1-norm only)	0.222±0.013	0.615±0.013	0.818±0.012
Our method ($\ell_{2,1}$-norm only)	0.223±0.011	0.618±0.014	0.815±0.013
Our method	**0.245 ± 0.013**	**0.641 ± 0.012**	**0.834 ± 0.052**

kernel learning (MKL) methods that are able to make use of multiple types of data: (1) SVM ℓ_∞ MKL method [13], (2) SVM ℓ_1 MKL [9], (3) SVM ℓ_2 MKL method [8], (4) least square (LSSVM) ℓ_∞ MKL method [19], (5) LSSVM ℓ_1 MKL method [14] and (6) LSSVM ℓ_2 MKL method [20]. Besides, we also compare our method to three most recent multi-model image classification methods published in computer vision community, including Gaussian process (GP) method [7], LPBoost-β method [5] and LPBoost-B method [5], which have demonstrated state-of-the-art object categorization performance. In addition, we also report the classification performances by SVM on each individual type of features and a straightforward concatenation of all six types of features as baselines.

We implement three versions of the proposed method. First, we set γ_2 in Eq. (3) as 0, which only uses the G_1-norm as regularization thereby only takes into account the structure over modalities. We denote it as "Our method (G_1-norm only)". Second, we set γ_1 in Eq. (3) as 0 to only use $\ell_{2,1}$-norm regularization, which thereby select feature shared across tasks yet modality structure is not considered. We denote this degenerate version of the proposed method as "Our method ($\ell_{2,1}$-norm only)". Finally, the full version of the proposed method by Eq. (3) is implemented and denoted as "Our method".

We conduct standard 5-fold cross-validation and report the average results. For each of the 5 trials, within the training data, an internal 5-fold cross-validation is performed to fine tune the parameters. The parameters of our method (γ_1 and γ_2 in Eq. (3)) are optimized in the range of $\left\{10^{-5}, 10^{-4}, \ldots, 10^{4}, 10^{5}\right\}$. For SVM method and MKL methods, one Gaussian kernel is constructed for each type of features (*i.e.*, $\mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$), where the parameter γ is fine tuned in the same range used as our method. We implement the compared MKL methods using the codes published by [20]. Following [20], in LSSVM ℓ_∞ and ℓ_2 methods, the regularization parameter λ is estimated jointly as the kernel coefficient of an identity matrix; in LSSVM ℓ_1 method, λ is set to 1; in all other SVM approaches, the C parameter of the box constraint is fine tuned in the same range as γ. For LPBoost-β and LPBoost-B methods, we use the codes published by the authors[4]. We use LIBSVM[5] software package to implement SVM in all our experiments.

Experimental results. Because we are concerned with single-label image classification, we employ the most widely used classification accuracy to assess the classification performance. The results of all compared methods in the three image classification tasks are reported in Table 2.

A first glance at the results shows that our methods generally outperform all other compared methods, which demonstrate the effectiveness of our methods in single-label image classification.

In addition, the methods using multiple data sources are significantly better than SVM using one single type of data. This confirms the usefulness of data integration in image classification.

Moreover, the results that our methods are always better than the MKL methods and boosting enhanced MKL methods is consistent with the previous theoretical analysis in that, although both of them take advantage of the information from multiple different sources, our method not only assigns proper weight to each type of features, but also rewards the relevances of the individual features inside a give feature type. In contrast, the MKL methods and boosting enhanced MKL methods only address the former while not being able to take into account the latter.

[4] http://www.vision.ee.ethz.ch/~pgehler/projects/iccv09/

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Finally, the performances of the full version of the proposed method is consistently better than those of its two degenerate versions, which demonstrate that both task-specific modality selection and across-task feature selection are necessary in image categorization, no one less.

3.2. Evaluation in Multi-Label Image Classification

Now we evaluate the proposed method in multi-label image classification, in which each image can be associated with more than one class label. We evaluate the proposed approach on the following two benchmark multi-label image data for image annotation tasks.

TRECVID 2005[6] data set contains 61901 images and labeled with 39 concepts (labels). As in most previous works [15], we randomly sample the data such that each concept has at least 100 images.

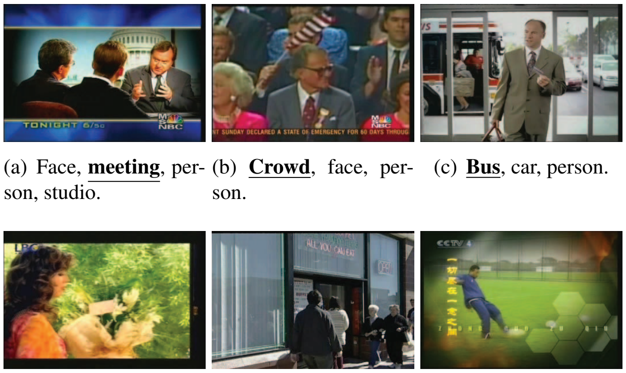
MSRC-v2[7] data set is an extension of MSRC-v1 data set, which has 591 images annotated by 22 classes.

Same as the MSRC-v1 data set, we extract six types of image features for these two data sets as detailed in the last column of Table 1 following [2].

Experimental setups. We still implement the three versions of our method and compare them against the MKL methods used in the above experiments with the same settings. For our method and MKL methods, we conduct classification for every class individually. For each class, we consider it as a binary classification task by using one-*vs.*-others strategy. For the classification on each individual data type and the simple mixture of all types of features, instead of using SVM as in the previous subsection for single-label data, we use multi-label k-Nearest Neighbor (Mk-NN) [21] method to classify the images, which is a broadly used multi-label classification method. In addition, we also implement two most recent multi-label classification methods including multi-label correlated Green's function (MCGF) [15] method and Multi-Label Least Square (MLLS) [6] method. However, these two methods are designed for data with single type of features, therefore we use the concatenation of all types of features as their input. We implement the two multi-label classification methods using the codes published by the authors.

The conventional classification performance metrics in statistical learning, *precision* and *F1 score*, are used to evaluate the compared methods. For every class, the precision and F1 score are computed following the standard definition for a binary classification problem. To address the multi-label scenario, following [10], macro average and micro average of precision and F1 score are computed to assess the overall performance across multiple labels.

Experimental results. The classification results by standard 5-fold cross-validation on TRECVID 2005 data set and MSRC-v2 data set are reported in Table 3. From the results, we can see that our method still performs the best on the both data sets. Besides, MKL methods using multiple types of features are generally better than the methods using single type of features. Although Mk-NN, MCGF and MLLS methods are designed for multi-label data, they can only work with one type of features. When using the feature concatenation as input, they assume all types of features as homogenous with no distinction, which, however, is not true in both our experiments as well as most real world applications. Although our method and MKL methods do not purposely address the multi-label settings, we still achieve better performance due to properly making use of the available information from various types of features. By further checking the detailed labeling results, we notice that many images can only be correctly annotated by our method, some of which are shown in Figure 1. The bolded and underlined labels can only be correctly predict by our method, but not the other compared methods. All these observations, again, confirm the effectiveness of the proposed approach in feature integration for multi-label image classification tasks.



(a) Face, **meeting**, person, studio.

(b) **Crowd**, face, person.

(c) **Bus**, car, person.

(d) **Outdoor**, face, person, vegetation.

(e) Building, **crowd**, person.

(f) Outdoor, person, **Sports**, vegetation.

Figure 1. Sample images from TRECVID 2005 data set, whose labels can only be correctly and completely predicted by the proposed method, not other compared methods. The bolded and underlined labels can only be correctly predict by our method.

4. Conclusions

We proposed a novel Sparse Multimodal Learning method to integrate different types of visual features for scene and object classifications. Instead of learning one parameter for all features from one modality as in multiple kernel learning, our method learned the parameters for each feature on different classes via the joint structured sparsity regularizations. Our new combined convex regularizations consider the importance of both feature modality and individual feature. The natural property of sparse regularization automatically identifies the important visual features for different visual recognition tasks. We derived an effi-

[6] http://www-nlpir.nist.gov/projects/trecvid

[7] http://research.microsoft.com/en-us/projects/objectclassrecognition/

Table 3. Multi-label classification performances (mean ± std) of the compared methods.

Data set	TRECVID 2005 data set				MSRC-v2 data set			
Methods	Macro average		Micro average		Macro average		Micro average	
	Precision	F1	Precision	F1	Precision	F1	Precision	F1
Mk-NN (Color moment)	0.415 ± 0.021	0.385 ± 0.018	0.408 ± 0.019	0.419 ± 0.023	0.391 ± 0.016	0.372 ± 0.017	0.380 ± 0.019	0.416 ± 0.021
Mk-NN (DoG-SIFT)	0.421 ± 0.022	0.392 ± 0.017	0.415 ± 0.020	0.429 ± 0.025	0.396 ± 0.018	0.375 ± 0.017	0.390 ± 0.016	0.420 ± 0.020
Mk-NN (LBP)	0.406 ± 0.016	0.379 ± 0.018	0.396 ± 0.020	0.409 ± 0.023	0.386 ± 0.019	0.366 ± 0.016	0.374 ± 0.017	0.408 ± 0.020
Mk-NN (HOG)	0.418 ± 0.021	0.389 ± 0.019	0.411 ± 0.022	0.425 ± 0.024	0.394 ± 0.019	0.374 ± 0.017	0.385 ± 0.020	0.417 ± 0.021
Mk-NN (GIST)	0.411 ± 0.022	0.383 ± 0.018	0.402 ± 0.020	0.418 ± 0.023	0.390 ± 0.018	0.368 ± 0.016	0.378 ± 0.021	0.413 ± 0.022
Mk-NN (CENTRIST)	0.425 ± 0.017	0.394 ± 0.019	0.419 ± 0.021	0.433 ± 0.026	0.399 ± 0.020	0.378 ± 0.017	0.393 ± 0.022	0.424 ± 0.024
Mk-NN (all by concatenation)	0.427 ± 0.024	0.398 ± 0.020	0.421 ± 0.023	0.437 ± 0.025	0.401 ± 0.019	0.383 ± 0.019	0.400 ± 0.020	0.428 ± 0.023
MCGF (all by concatenation)	0.431 ± 0.018	0.401 ± 0.019	0.422 ± 0.021	0.442 ± 0.024	0.404 ± 0.020	0.385 ± 0.017	0.405 ± 0.021	0.430 ± 0.022
MLLS (all by concatenation)	0.434 ± 0.025	0.405 ± 0.024	0.427 ± 0.026	0.446 ± 0.027	0.409 ± 0.018	0.390 ± 0.019	0.408 ± 0.022	0.434 ± 0.025
SVM ℓ_∞ MKL method	0.477 ± 0.026	0.429 ± 0.024	0.463 ± 0.025	0.486 ± 0.028	0.420 ± 0.021	0.395 ± 0.019	0.416 ± 0.017	0.443 ± 0.022
SVM ℓ_1 MKL method	0.470 ± 0.023	0.427 ± 0.020	0.458 ± 0.021	0.479 ± 0.025	0.415 ± 0.018	0.392 ± 0.016	0.410 ± 0.020	0.441 ± 0.023
SVM ℓ_2 MKL method	0.461 ± 0.020	0.412 ± 0.019	0.445 ± 0.018	0.463 ± 0.021	0.404 ± 0.017	0.381 ± 0.016	0.400 ± 0.019	0.429 ± 0.021
LSSVM ℓ_∞ MKL method	0.452 ± 0.022	0.404 ± 0.017	0.438 ± 0.018	0.458 ± 0.020	0.400 ± 0.018	0.374 ± 0.015	0.399 ± 0.020	0.426 ± 0.022
LSSVM ℓ_1 MKL method	0.466 ± 0.022	0.423 ± 0.020	0.451 ± 0.025	0.473 ± 0.027	0.412 ± 0.019	0.388 ± 0.018	0.408 ± 0.021	0.436 ± 0.024
LSSVM ℓ_2 MKL method	0.463 ± 0.019	0.417 ± 0.016	0.448 ± 0.019	0.469 ± 0.023	0.406 ± 0.018	0.384 ± 0.019	0.406 ± 0.021	0.433 ± 0.025
GP method	0.460 ± 0.019	0.414 ± 0.020	0.442 ± 0.021	0.466 ± 0.220	0.403 ± 0.020	0.389 ± 0.019	0.409 ± 0.022	0.437 ± 0.021
LPboost-β	0.471 ± 0.017	0.420 ± 0.019	0.459 ± 0.016	0.480 ± 0.219	0.428 ± 0.016	0.402 ± 0.015	0.412 ± 0.016	0.442 ± 0.015
LPboost-B	0.469 ± 0.020	0.417 ± 0.018	0.456 ± 0.020	0.479 ± 0.016	0.427 ± 0.018	0.400 ± 0.016	0.410 ± 0.018	0.440 ± 0.019
Our method (G_1-norm only)	0.472 ± 0.007	0.423 ± 0.016	0.462 ± 0.011	0.481 ± 0.014	0.431 ± 0.011	0.402 ± 0.012	0.415 ± 0.012	0.445 ± 0.009
Our method ($\ell_{2,1}$-norm only)	0.478 ± 0.010	0.421 ± 0.012	0.467 ± 0.018	0.484 ± 0.012	0.429 ± 0.012	0.405 ± 0.014	0.413 ± 0.014	0.443 ± 0.010
Our method	**0.509 ± 0.013**	**0.461 ± 0.019**	**0.503 ± 0.015**	**0.511 ± 0.016**	**0.451 ± 0.022**	**0.420 ± 0.023**	**0.439 ± 0.025**	**0.468 ± 0.026**

cient optimization algorithm to solve our non-smooth objective and provided a rigorous proof on its global convergence. Extensive experiments have been performed on both single-label and multi-label image categorization tasks, our approach outperforms other related methods in all benchmark data sets.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2007.

[2] X. Cai, F. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR (1)*, pages 886–893, 2005.

[4] S. Gao, L. Chia, and I. Tsang. Multi-layer group sparse codingłfor concurrent image classification and annotation. In *CVPR*, 2011.

[5] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009.

[6] S. Ji, L. Tang, S. Yu, and J. Ye. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–29, 2010.

[7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *International journal of computer vision*, 88(2):169–188, 2010.

[8] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg. Non-sparse multiple kernel learning. In *NIPS*.

[9] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

[10] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[11] F. Nie, H. Wang, H. Huang, and C. Ding. Unsupervised and semi-supervised learning via l_1-norm graph. In *ICCV*, pages 2268–2273, 2011.

[12] A. Oliva and A. B. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[14] J. Suykens, T. Van Gestel, and J. De Brabanter. *Least squares support vector machines*. World Scientific Pub Co Inc, 2002.

[15] H. Wang, H. Huang, and C. Ding. Image annotation using multi-label correlated Green's function. In *ICCV*, 2009.

[16] H. Wang, F. Nie, H. Huang, S. L. Risacher, C. Ding, A. J. Saykin, L. Shen, and ADNI. A new sparse multi-task regression and feature selection method to identify brain imaging predictors for memory performance. *ICCV 2011: IEEE Conference on Computer Vision*, pages 557–562, 2011.

[17] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012.

[18] J. Wu and J. M. Rehg. Where am i: Place instance and category recognition using spatial pact. In *CVPR*, 2008.

[19] J. Ye, S. Ji, and J. Chen. Multi-class discriminant kernel learning via convex programming. *JMLR*, 9:719–758, 2008.

[20] S. Yu, T. Falck, A. Daemen, L. Tranchevent, J. Suykens, B. De Moor, and Y. Moreau. L 2-norm multiple kernel learning and its application to biomedical data fusion. *BMC bioinformatics*, 11(1):309, 2010.

[21] M. Zhang and Z. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.