Heterogeneous Image Features Integration via Multi-Modal Semi-Supervised Learning Model

Xiao Cai[†], Feiping Nie[†], Weidong Cai[‡], Heng Huang[†*]
[†]University of Texas at Arlington, Arlington, Texas 76019, USA
[‡]School of Information Technologies, University of Sydney, NSW 2006, Australia
xiao.cai@mavs.uta.edu, feipingnie@gmail.com, tom.cai@sydney.edu.au, heng@uta.edu

Abstract

Automatic image categorization has become increasingly important with the development of Internet and the growth in the size of image databases. Although the image categorization can be formulated as a typical multi-class classification problem, two major challenges have been raised by the real-world images. On one hand, though using more labeled training data may improve the prediction performance, obtaining the image labels is a time consuming as well as biased process. On the other hand, more and more visual descriptors have been proposed to describe objects and scenes appearing in images and different features describe different aspects of the visual characteristics. Therefore, how to integrate heterogeneous visual features to do the semi-supervised learning is crucial for categorizing large-scale image data. In this paper, we propose a novel approach to integrate heterogeneous features by performing multi-modal semi-supervised classification on unlabeled as well as unsegmented images. Considering each type of feature as one modality, taking advantage of the large amount of unlabeled data information, our new adaptive multi-modal semi-supervised classification (AMMSS) algorithm learns a commonly shared class indicator matrix and the weights for different modalities (image features) simultaneously.

1. Introduction

With the proliferation of digital photography and online data sources, automatic image categorization becomes increasingly important. As a multi-class classification problem, image categorization has been widely studied in the computer vision community. The target categories usually come from the various real world applications in objective recognition and scene classification, *e.g.* defined by presence of a certain salient object, such as the stop sign or the staple [9] or defined with respect to scene types, such as forest, highway and inside city, *etc* [17]. It is a challenging task not only due to the fact that the number of labeled images is much smaller than that of the unlabeled images in the real world but also because of image's variability, ambiguity, and wide range of illumination. As we know, in the traditional supervised learning paradigm, increasing the quantity and diversity of labeled images enhances the performance of the learned classifier. Nevertheless, labeling image is a time consuming as well as biased task. Although it is possible to label large amounts of images for research purposes, this is often unrealistic in practice. To solve the classification problem caused by the scarce or expensive labeled data, we resort to semi-supervised learning, which takes advantage of the combination of both labeled and unlabeled images.

The most popular way to do semi-supervised learning for image categorization is to use some low-level image descriptor. In order to overcome the image content representation issue, more and more visual descriptors have been proposed. Some focus on the local information, while others are holistic descriptors. If we integrate all the descriptors via a proper learning method, we could create a generally more accurate and more robust descriptor than any single one.

In this paper, we propose a novel semi-supervised learning approach to integrate heterogeneous features from both labeled and unlabeled as well as unsegmented images. Considering each type of feature as one modality, taking advantage of the large amount of unlabeled data information, our new adaptive multi-modal semi-supervised classification (AMMSS) algorithm propagates the class labels from labeled images to unlabeled images based on the integrated multi-modal feature similarity and learn the weights for different modalities (image features) simultaneously. We applied our AMMSS method to integrate multiple popularly used image features, which describe the image content from different perspectives, and evaluated the performance by

*Corresponding author. This project was partially supported by U.S. NSF IIS-1117965, IIS-1302675, IIS- 1344152, and ARC grant.

four benchmark datasets. Compared with the existing semi-supervised scene and object categorization methods, our approach always achieves superior performances in terms of both macro and micro classification accuracy.

2. Related Work

As the most popularly used semi-supervised learning models, the graph based semi-supervised methods define a graph where the nodes encompass labeled as well as unlabeled data, and edges (may be weighted) reflect the similarity of data points [26, 27, 28, 7]. Nevertheless, they take advantage of the affinity matrix or graph Laplacian matrix extracted from one visual descriptor only.

Co-training semi-supervised learning model trains two or more separate classifiers. It can fuse the descriptors by learning a separate classifier using each visual feature and iteratively training examples for each classifier based on the output of the other classifier. Each classifier then classifies the unlabeled data and teaches the other classifier with the few unlabeled examples they feel most confident. Each classifier is retrained with the additional training examples given by the other classifier and the process repeats.But the drawback of the co-training is that it requires all the classifiers over the separate feature sets to be accurate. In other words, the performance of co-training is not robust to the outlier feature set whose performance is far below the average, since the outlier will provide erroneous information to other classifiers and deteriorate the overall classification result.

How to properly integrate heterogeneous features is becoming an emerging topic nowadays. As a multi-kernel learning algorithm, the heterogeneous feature machine (HFM) [3] was recently proposed based on logistic regression loss function and group LASSO regularization to ***supervised*** fuse the multiple types of features for visual classifications. Constructing a shared common cluster indicator with non-negative constraint via non-negative matrix factorization, multi-modality spectral clustering (MMSC) [2, 5] ***unsupervised*** merges the different features. Although many supervised and unsupervised methods have been proposed to integrate the multi-modal features, there is no semi-supervised learning model to integrate heterogeneous image visual features. The structured sparse-inducing norms were also used for feature integration in different applications [21, 1, 20, 19].

In this paper, we will tackle this problem by a novel graph based semi-supervised learning model to adaptively fuse different visual features for **semi-supervised** image categorizations. To begin with, let's first summarize the notation that will be used in this paper. Matrices are written as uppercase letters and vectors are written as boldface lowercase letters. m_{ij} is the entry located at i-th row and j-th column of matrix M.

2.1. Basic Framework of Graph Based Semi-Supervised Learning

Assume we have n images $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, where each image is abstracted as a data point $\mathbf{x}_i \in \mathbb{R}^p$. Each data point $\mathbf{x}_i$ belongs to one of K classes $\mathcal{C} = \{c_1, \cdots, c_K\}$ represented by $\mathbf{y}_i \in \{0,1\}^K$, such that $\mathbf{y}_i(k) = 1$ if $\mathbf{x}_i$ is classified into k-th class, and 0 otherwise. Without loss of generality, we assume the first $l \ll n$ data are already labeled, which are denoted as $T = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{l}$. Our task is to learn a function $f : \chi \to \{0,1\}^K$ from T that is able to classify the given unlabeled data $\mathbf{x}_i (l+1 \le i \le n)$ into one and only one class in $\mathcal{C}$. For simplicity, we use u to denote the number of unlabeled data point. that is, $l + u = n$ and split the label matrix $Y = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T$, $\mathbf{y}_i \in \mathbb{R}^K$ into 2 blocks: $Y = \begin{bmatrix} Y_l \\ Y_u \end{bmatrix}$.

Given the dataset X, all the image data including the labeled and unlabeled ones are abstracted as the vertices on $\mathscr{K} - NN$ graph. To be specific, We connect $\mathbf{x}_i$, $\mathbf{x}_j$ if one of them is among the other's $\mathscr{K}$-nearest neighbor by Euclidean distance and define the corresponding weight on the edge as the following,

$$w_{ij} = \begin{cases} \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}), & if\ \mathbf{x}_i\ and\ \mathbf{x}_j\ are\ connected \\ 0, & otherwise \end{cases} \tag{1}$$

where σ is the bandwidth parameter. Therefore, $W = \{w_{i,j}\}$ is an $(l+u) \times (l+u)$ symmetric undirected matrix with non-negative edge weight. Let $d_{ii} = \sum_{j=1}^{l+u} w_{ij}$ and D be the diagonal matrix by substituting d_{ii}, $i = 1, 2, ...(l+u)$ on the diagonal. The normalized graph Laplacian matrix L is defined as

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{2}$$

2.2. Label Propogation for Single Modality

According to graph theory, if the edge weight between two vertices on affinity matrix is large, then the class labels of these two instances should be similar. Based on the above assumption, denote $G \in \mathbb{R}^{n \times K}$ as the class label matrix, for each feature modality, we use the following way to propagate the class label information from labeled data to unlabeled data,

$$\min_G G^T L G \quad s.t. \quad \mathbf{g}_i = \mathbf{y}_i, \quad \forall i = 1, 2, ..., l, \tag{3}$$

where L is the normalized Laplacian matrix defined in Eq. (2).
Eq. (3) can be rewritten as the following,

$$\min_{G_u} Tr(\begin{bmatrix} Y_l \\ G_u \end{bmatrix}^T \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \begin{bmatrix} Y_l \\ G_u \end{bmatrix}), \tag{4}$$

since we know the labels Y_l for the first l instances, which has the following unique solution,

$$G_u = -L_{uu}^{-1} L_{ul} Y_l \tag{5}$$

2.3. Label Propogation by Adaptive Multi-Modalities Semi-Supervised Learning

In order to properly and naturally integrate heterogeneous image features to do semi-supervised learning, we need a co-regularization term to learn a consensus class label matrix and let the differences between that consensus label matrix and the class label matrix of each feature modality as small as possible. With the addition of weight factor for each feature modality, we adaptively learn the weight for each feature modality, assigning the more discriminative modality with higher weight. We summarize the proposed AMMSS method as the following objective function,

$$\begin{aligned} &\min_{G, G^{(v)}, \alpha^{(v)}} \sum_{v=1}^{V} (\alpha^{(v)})^r Tr(G^{(v)^T} L^{(v)} G^{(v)}) \\ &+\lambda \sum_{v=1}^{V} Tr((G - G^{(v)})^T (G - G^{(v)})) \\ &s.t. \quad \mathbf{g}_i = \mathbf{y}_i, \quad \forall i = 1, 2, ...l, \quad \sum_{v=1}^{V} \alpha^{(v)} = 1, \quad \alpha^{(v)} \geq 0, \end{aligned} \tag{6}$$

where V is the number of image visual features, $\alpha^{(v)}$ is the non-negative normalized weight factor for the v-th modality, $L^{(v)}$ and $G^{(v)}$ are the normalized Laplacian matrix and class label matrix for the v-th feature modality respectively. G is the shared consensus class label matrix that we are interested. We use the scalar r to control the distribution of different weights for different feature modalities and λ is the regularization parameter to balance the 1st term and the 2nd term. We want to solve for G, $G^{(v)}$ and $\alpha^{(v)}$ simultaneously via the proposed Eq. (6).

3. Optimization Algorithms

3.1. The Optimization Algorithm of AMMSS

We decompose Eq. (6) as the following three subproblems and solve them alternatively and iteratively.

The first step is fixing G and $G^{(v)}$, solving $\alpha^{(v)}$. Then, the objective function becomes

$$\min_{\alpha^{(v)}} \sum_{v=1}^{V} (\alpha^{(v)})^r Tr(G^{(v)^T} L^{(v)} G^{(v)}), \quad s.t. \sum_{v=1}^{V} \alpha^{(v)} = 1, \ \alpha^{(v)} \geq 0 \tag{7}$$

Let $p^{(v)} = Tr(G^{(v)^T} L^{(v)} G^{(v)})$, then the Eq. (7) can be rewritten as

$$\sum_{v=1}^{V} (\alpha^{(v)})^r p^{(v)}, \quad s.t. \quad \sum_{v=1}^{V} \alpha^{(v)} = 1, \quad \alpha^{(v)} \geq 0 \tag{8}$$

Algorithm 1 The algorithm of AMMSS

Input:
1. Affinity matrices $\{W^{(1)}, \cdots, W^{(V)}\} \in \mathbb{R}^{n \times n}$
2. The labels for the first l images, $Y_l = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_l]^T$, $\mathbf{y}_i \in \mathbb{B}^{K \times 1}, \quad \forall i = 1, 2, ..., l$.
3. The parameters r and λ.

Output:
1. The predicted labels for the unlabeled images $y_i, \quad \forall i = l+1, l+2, ..., n$.
2. The weight scalar $\alpha^{(v)}, \quad \forall v = 1, 2, ..., V$ for each modality.

Initialization:
1. Set $t = 0$
2. Initialize the weight for each modality, $\alpha_t^{(v)} = \frac{1}{V}, \quad \forall v = 1, 2, ..., V$
3. Initialize the common class label matrix, $G_t = \begin{bmatrix} G_{l_t} \\ G_{u_t} \end{bmatrix} = \begin{bmatrix} Y_l \\ Y_u \end{bmatrix}$ where $Y_u \in \mathbb{R}^{u \times K}$ is a random matrix and each entry $u_{i,j} \in \{0, 1\}$.
4. Calculate the normalized Laplacian matrices for each feature modality, $L_t^{(v)} = I - (D_t^{(v)})^{-\frac{1}{2}} W_t^{(v)} (D_t^{(v)})^{-\frac{1}{2}}$

Procedure:
repeat
1. Calculate $\tilde{L}_t^{(v)} = (\alpha_t^{(v)})^r L_t^{(v)}$
2. Calculate the class indicator matrix for each modality $G_t^{(v)} = \lambda(\tilde{L}_t^{(v)} + \lambda I)^{-1} G_t$
3. Calculate $H_t = \sum_{v=1}^{V} \left(I - \lambda(\tilde{L}_t^{(v)} + \lambda I)^{-1}\right)$ and split the H_t by Eq. (16).
4. Calculate $p_t^{(v)} = Tr(G_t^{(v)^T} L_t^{(v)} G_t^{(v)})$
5. Update the weight for each modality as $\alpha_{t+1}^{(v)} = (rp_t^{(v)})^{\frac{1}{1-r}} / \sum_{v=1}^{V} (rp_t^{(v)})^{\frac{1}{1-r}}$
6. Update $G_{u_{t+1}} = -H_{uu_t}^{-1} H_{ul_t} Y_l$. And update $G_{t+1} = \begin{bmatrix} Y_l \\ G_{u_t} \end{bmatrix}$
7. Update $t = t + 1$
until Converges
Assign the single class label for the unlabeled images by Eq. (20).

Thus, the Lagrange function of Eq. (8) is

$$\sum_{v=1}^{V} (\alpha^{(v)})^r p^{(v)} - \beta(\sum_{v=1}^{V} \alpha^{(v)} - 1) \tag{9}$$

where β is the Lagrange multiplier. In order to get the optimal solution of the above subproblem, set the derivative of Eq. (9) with respect to $\alpha^{(v)}$ to zero. We have

$$\alpha^{(v)} = \left(\frac{\beta}{rp^{(v)}}\right)^{\frac{1}{r-1}} \tag{10}$$

Substitute the resultant $\alpha^{(v)}$ in Eq. (10) into the constraint $\sum_v \alpha^{(v)} = 1$, we get

$$a^{(v)} = (rp^{(v)})^{\frac{1}{1-r}} / \sum_{v=1}^{V} (rp^{(v)})^{\frac{1}{1-r}} \tag{11}$$

The second step is fixing $\alpha^{(v)}$ and G, solving $G^{(v)}$. We

change the variable and let $\tilde{L}^{(v)} = (\alpha^{(v)})^r L^{(v)}$ then the objective function becomes

$$\begin{aligned} &\min_{G,G^{(v)}} \sum_v Tr({G^{(v)}}^T \tilde{L}^{(v)} G^{(v)}) + \lambda \sum_v Tr((G-G^{(v)})^T (G-G^{(v)})) \\ &s.t. \quad \mathbf{g}_i = \mathbf{y}_i, \quad \forall i = 1,2,...,l \end{aligned} \tag{12}$$

Set the derivative of Eq. (12) with respect to $G^{(v)}$ to zero. We have

$$G^{(v)} = \lambda(\tilde{L}^{(v)} + \lambda I)^{-1} G \tag{13}$$

The third step is fixing $\alpha^{(v)}$ and $G^{(v)}$, solving G. Substitute the resultant $G^{(v)}$ in Eq. (13) into the Eq. (12), we get (The proof is in Appendix)

$$\begin{aligned} &\sum_v Tr({G^{(v)}}^T \tilde{L}^{(v)} G^{(v)}) + \lambda \sum_v Tr((G-G^{(v)})^T (G-G^{(v)})) \\ &= \lambda Tr(G^T (\sum_v (I - \lambda(\tilde{L}^{(v)} + \lambda I)^{-1}) G) \end{aligned} \tag{14}$$

Let $H = \sum_v \left(I - \lambda(\tilde{L}^{(v)} + \lambda I)^{-1} \right)$. Therefore, Eq. (12) is equivalent to the following optimization problem,

$$\begin{aligned} &\min_G Tr(G^T H G) \\ &s.t. \quad \mathbf{g}_i = \mathbf{y}_i, i = 1,2,...,l \end{aligned} \tag{15}$$

To compute class label matrix for the unlabeled image explicitly in terms of matrix operations, we split the matrix H into 4 blocks by the l-th row and l-th column:

$$H = \begin{bmatrix} H_{ll} & H_{lu} \\ H_{ul} & H_{uu} \end{bmatrix} \tag{16}$$

Therefore,

$$\begin{aligned} &Tr(G^T H G) \\ &= Tr\left(\begin{bmatrix} G_l \\ G_u \end{bmatrix}^T \begin{bmatrix} H_{ll} & H_{lu} \\ H_{ul} & H_{uu} \end{bmatrix} \begin{bmatrix} G_l \\ G_u \end{bmatrix} \right) \\ &= Tr\left(\begin{bmatrix} Y_l \\ G_u \end{bmatrix}^T \begin{bmatrix} H_{ll} & H_{lu} \\ H_{ul} & H_{uu} \end{bmatrix} \begin{bmatrix} Y_l \\ G_u \end{bmatrix} \right) \\ &= Tr(Y_l^T H_{ll} Y_l + G_u^T H_{ul} Y_l + Y_l^T H_{lu} G_u + G_u^T H_{uu} G_u) \\ &= Tr(Y_l^T H_{ll} Y_l + G_u^T H_{ul} Y_l + G_u^T H_{ul} Y_l + G_u^T H_{uu} G_u) \end{aligned} \tag{17}$$

Thus optimization problem in Eq. (15) is equivalent to the subsequent problem,

$$\min_{G_u} [2Tr(G_u^T H_{ul} Y_l) + Tr(G_u^T H_{uu} G_u)] \tag{18}$$

Setting the derivative of Eq. (18) to zero with respect to G_u, we get

$$G_u = -H_{uu}^{-1} H_{ul} Y_l \tag{19}$$

By the above three steps, we alternatively update $\alpha^{(v)}$, $G^{(v)}$ and G and repeat them iteratively untill the objective function converges. At last, we resort to the following decision function to assign the single class label to the unlabeled images,

$$y_i = \arg\max_j G_{ij}, \ \forall i = l+1, l+2, ..., n. \forall j = 1,2,...,K. \tag{20}$$

We summarize the algorithm in Alg. 1.

3.2. Convergence of The Algorithm

We will prove the convergence of the proposed Alg. 1 as following: We divide the original problem Eq. (1) into three subproblems and each of them is convex problem. Since the original problem is not a joint convex problem, by solving the subproblems alternatively, Alg. 1 will converge to the local solution and we use $1/V$ as the initial weight for each modality. Later in our experiment we will demonstrate the fast convergence of our algorithm.

3.3. Discussion of The Parameter r

In AMMSS, we use one parameter r to control the distribution of weight factors for different feature modalities. From Eq. (11), we can see that when $r \to \infty$, we will get equal weight factors. And when $r \to 1$, we will assign 1 to the weight factor of the modality whose $p^{(v)}$ value is the smallest and assign 0 to the weights of other modalities. Using such kind of strategy, on one hand, we avoid the trivial solution to the weight distribution of the different modalities, that is, the solution when $r \to 1$. On the other hand, surprisingly, we can take advantage of only one parameter r to control the whole weights, reducing the parameters of the model greatly.

4. Experimental Results

Since our AMMSS is a kind of graph based semi-supervised learning Algorithm, we will compare the performance of our AMMSS and related graph based state-of-art semi-supervised methods on five benchmark datasets: Caltech-101 [14], Microsoft Research Cambridge Volume 1 (MSRC-v1) [22], Handwritten numerals (HW) [10] and Animal with Attributes(AwA) [12]. The image classification performance is evaluated in terms of average macro and micro classification accuracy.

4.1. Dataset Descriptions

Caltech-101 Images The Caltech101 image dataset contains 8677 images of objects, each with approximately 0.1 mega pixel resolution, belonging to 101 categories. We follow [8] to choose 7 and 20 classes dataset respectively from 101 classes. The 7 classes include Faces, Motorbikes, Dolla-Bill, Garfield, Snoopy, Stop-Sign, Windsor-Chair and have 441 images in total. The 20 classes include Faces, Leopards, Motorbikes, Binocular, Brain, Camera, Car-Side, Dollar-Bill, Ferry, Garfield, Hedgehog, Pagoda,

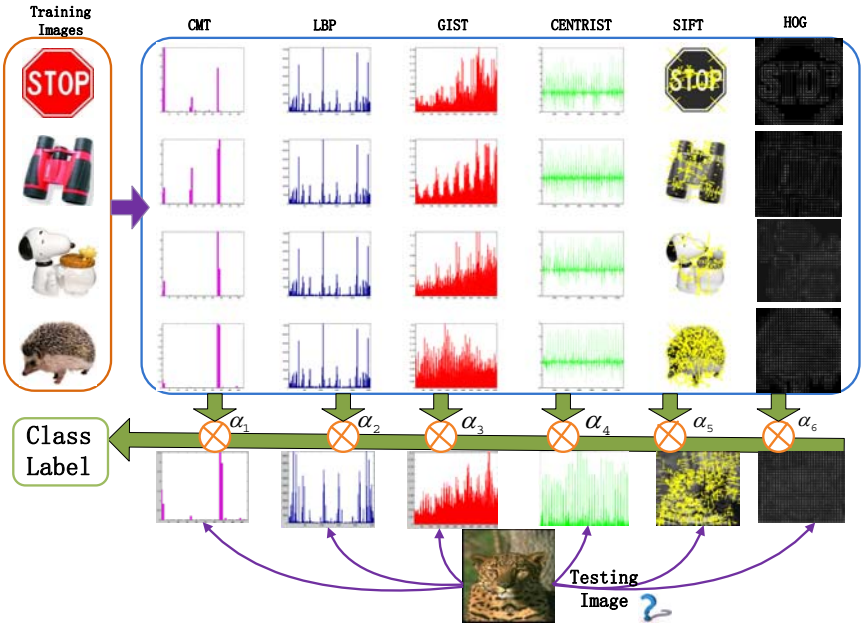


Figure 1. The demonstration of different visual descriptors from Caltech 101 dataset. The final class label of the testing image is decided by the weighted six different feature modalities, where the weight for different feature modality is learned by the training images.

Rhino, Snoopy, Stapler, Stop-Sign, Water-Lilly, Windsor-Chair, Wrench, Yin-Yang and have 1230 images all together.

MSRC-v1 Images We follow Lee and Grauman's approach [13] to refine the dataset, getting 7 classes composed of tree, building, airplane, cow, face, car, bicycle, and each refined class has 30 images. Compared to the Caltech101 dataset, MSRC-v1 has more clutter and variability in the objects appearances.

Since there is no published image descriptors for Caltech-101 and MSRC-v1 datasets, we extract the following six popular visual features for each image: On one hand, we extract three holistic visual features for each image, *i.e.* 45 dimension color moment (CMT) [24]; 512 dimension GIST feature [17]; 1302 dimension CENTRIST feature [23]. On the other hand, we collect three local descriptor as well, *i.e.* 256 dimension local binary pattern (LBP) [16]; 576 dimension HOG feature and famous 128 dimension DoG-SIFT descriptor [15].

Handwritten numerals (HW) Handwritten numerals dataset consists of 2000 data point for 0 to 9 ten digit classes. (Each class has 200 data points.) We use the published six visual features [10] extracted from each image. Specifically, the six visual features are 76 dimension Fourier coefficients of the character shapes (FOU), 216 dimension profile correlations (FAC), 64 dimension Karhunen-love coefficients (KAR), 240 dimension pixel averages in 2×3 windows (PIX), 47 dimension Zernike moment (ZER) and 6 dimension morphological (MOR) features.

Animal with attributes (AWA) Animal with attributes data set is the largest data set, which is also an image data set consisting of 6 feature 50 classes. We randomly sample 50 images for each class and get 2500 images in total. We utilize all the published features, that is, 2688 dimension Color Histogram (CQ) features , 2000 dimension Local Self-Similarity (LSS) features , 252 dimension PyramidHOG (PHOG) features, 2000 dimension SIFT features, 2000 dimension colorSIFT (RGSIFT) feature and 2000 dimension SURF features.

4.2. Experimental Setup

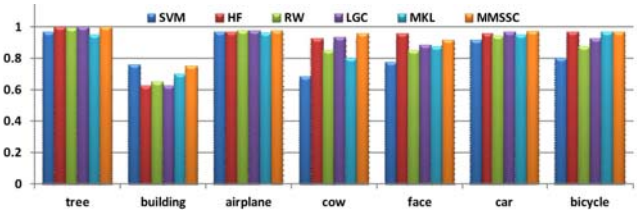
We use the Gaussian Kernel in Eq. (1) with 7-nearest neighbor to get the affinity matrices for different visual features. We utilize self-tuning method [25] to calculate the bandwidth parameter σ. In order to solve the inequality length problem of the DoG-SIFT feature, we utilize the pyramid match kernel [11] to build the similarity matrix, using the LIBPMK toolkit. Thus, given an image, we have multiple similarity (affinity) matrices calculated from different modalities. In our experiment for each dataset to mimic the "real" situation in semi-supervised learning case ($l \ll u$), we randomly choose 20% data for training and use the rest for testing. We repeat the above procedure 10 times and report the average result. r is the parameter to control the distribution of the weights for different feature modalities, which we will discuss in detail later. We search the logarithm of the parameter r, that is, $log_{10}r$ in the range from 0.1 to 2 with incremental step 0.2 and search the regularization parameter λ in the range from 0 to 1 with incremental step 0.1 to get the best parameters r^* as well as λ^* based on the 2-fold cross validation inside the training data only.

4.3. Classification Results Comparison

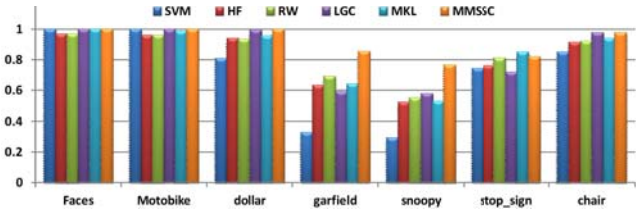
First of all, in order to test the feature integration power of our method, we compare classification performance using all the feature modalities with that using only one feature modality. From Table 1 to Table 3, we can draw the conclusion that the performance of our proposed AMMSS can beat the best of single modality, which tackles the problem of Eq. (3).

We also compare our methods with some graph based state-of-the-art semi-supervised learning methods: (a) the harmonic function (HF) approach [28], (b) learning with local and global consistency approach (LGC) [26] and (c) the random walk approach (RW) [27]. For each of the above three methods, we use the kernel addition (KA), that is, the simple average of equal weighted Laplacian matrices or the graph Laplacian of the concatenated features of all modalities (FC) as the input for HF, LGC as well as RW. Moreover, for sake of completeness, we also compare the results of support vector machine with the precomputed kernel Eq. (1) implemented by LIBSVM [4]. Since Multiple Kernel Learning (MKL) approaches [18] can also realize feature integration if we consider one feature modality as one kernel, we report its classification result as well. Moreover, since our method can learn the weight for each feature modality adaptively, we compare the results of out model using equal weight (MMSS). We adopt the optimal parameter settings for the above methods empirically. As for performance evaluation, we utilize the widely-used performance metrics, average macro classification accuracy as well as average micro classification accuracy for each class. Average macro classification accuracy is shown in Table 4 and micro accuracy for all the datasets are shown in Fig. 3. We can see that our method always achieves consistently better results than the other state-of-art methods in terms of average macro classification accuracy and choosing different weight for different features can even boost the performance of multi modality semi-supervised learning results. As for average micro classification accuracy, the results of AMMSS are the best for most classes. The confusion matrices of MSRCV1, Caltech101-7 and handwritten numbers are shown in Fig. 2.

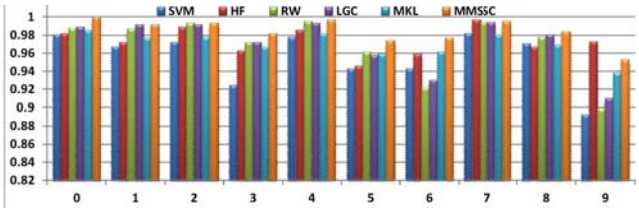
Moreover, since our method can learn the weight for each feature modality after convergence, we add the generalization ability of the objective function Eq. (6). Fig. 4 shows the learned weight by our Alg. 1 on five benchmark datasets. From it, we can observe that DoG-SIFT has the most discriminate power in Caltech$101-7$ dataset, CENTRIST has the highest weight for Caltech$101-20$ dataset while for MSRCV1 dataset, GIST is the best feature modality among the six which is consistent with single modality's performance shown in Table 1. Instead of treating each feature modality equally, our method can do weighting each feature modality and classification simultaneously.



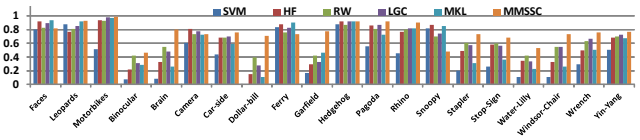
(a) MSRCV-1



(b) Caltech101-7



(c) handwritten number



(d) Caltech101-20

Figure 3. The Micro accuracy on two datasets (a) MSRCV1 (b) Caltech101-7. (c) Handwritten number (d) Caltech101-20

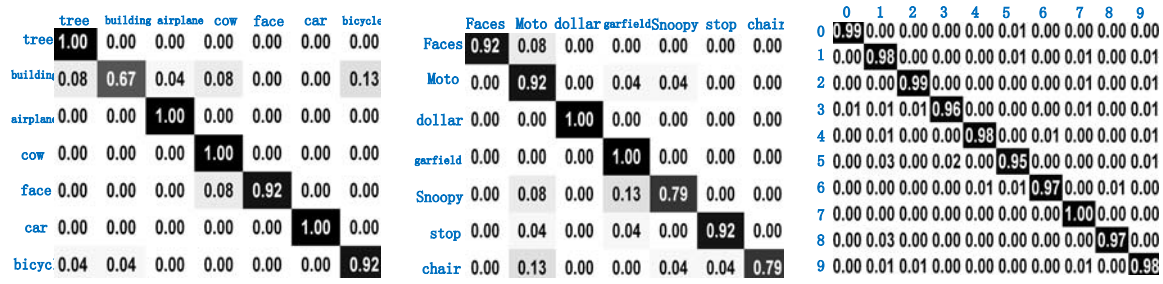
Table 1. The average macro classification accuracy compared with single view on Caltech101-7, Caltech101-20 and MSRCV1 datasets.

Methods	Caltech7	Caltech20	MSRCV1
CTM [24]	0.45	0.27	0.30
LBP [16]	0.66	0.39	0.71
GIST [17]	0.80	0.51	0.79
CENTRIST [23]	0.79	0.70	0.77
DoG-SIFT [15]	0.81	0.30	0.51
HOG [6]	0.89	0.27	0.69
AMMSS	**0.91**	**0.74**	**0.94**

Table 2. The average macro classification accuracy compared with single view on handwritten numbers dataset.

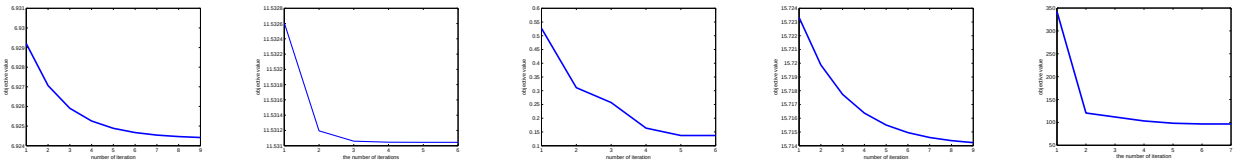
Data	FOU	FAC	KAR	PIX	ZER	MOR	AMMSS
HW	0.92	0.82	0.93	0.46	0.94	0.82	**0.98**

At last, we test the convergency speed of our AMMSS algorithm, which is shown in Fig. 5. From it, we can observe that our AMMSS algorithm converges very fast on all



(a) avg. accuracy **94**% (b) avg. accuracy **91**% (c) avg. accuracy **98**%

Figure 2. Calculated confusion matrix by AMMSS method (a) MSRCV1 (b) Caltech101-7 (c) Handwritten numerals.



(a) Caltech101 – 7 (b) Caltech101 – 20 (c) MSRCV1 (d) Handwritten (e) AwA

Figure 5. The convergency of five datasets (a) Caltech101-7 (b) Caltech101-20 (c) MSRCV1 (d) Handwritten Numbers (e) AwA

Table 3. The average macro classification accuracy compared with single view on animal with attribute dataset.

Data	CQ	LSS	PHOG	RGISIFT	SIFT	SURF	AMMSS
AWA	0.057	0.062	0.050	0.054	0.065	0.072	**0.095**

Table 4. The average macro classification accuracy compared with baseline methods on all datasets.

Methods	Caltech7	Caltech20	MSRCV1	HW	AWA
SVM [4]	0.85	0.59	0.86	0.95	0.076
MKL [18]	0.89	0.68	0.89	0.96	0.079
HF(KA) [28]	0.84	0.70	0.92	0.97	0.079
HF(FC) [28]	0.82	0.68	0.89	0.96	0.077
RW(KA) [27]	0.89	0.72	0.88	0.97	0.080
RW(FC) [27]	0.86	0.69	0.87	0.96	0.079
LGC(KA) [7]	0.87	0.72	0.90	0.97	0.081
LGC(FC) [7]	0.89	0.71	0.88	0.96	0.079
MMSS	0.89	0.72	0.92	0.97	0.086
AMMSS	**0.91**	**0.74**	**0.94**	**0.98**	**0.095**

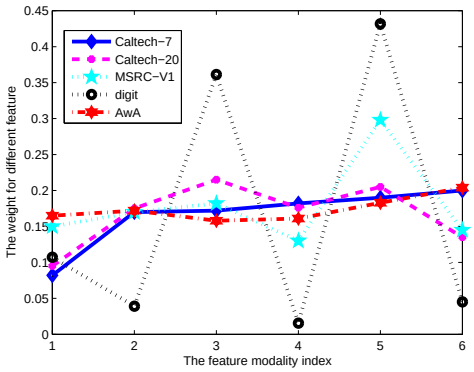


Figure 4. The learned weight factor for different modalities on five dataset. The feature index on x-axis from 1 to 6 stands for CMT, LBP, GIST, HOG, CENTRIST and DOG-SIFT respectively for Caltech-7, Caltech-20 and MSRCV1 datasets. And the index on x-axis from 1 to 6 stands for FOU, FAC, KAR, PIX, ZER, MOR respectively for Handwritten numbers dataset. The index on x-axis from 1 to 6 stands for CQ, LSS, PHOG, RGSIFT, SIFT, SURF respectively for AwA dataset.

the datasets and usually the number of iteration is less than 10.

5. Conclusion

In this paper, we proposed a novel adaptive multi-modality semi-supervised learning method (AMMSS) to jointly learn the weight for each feature modality as well as the common class labels for the unlabeled data. Utilizing our algorithm, decomposing the original problem into three convex subproblems, we can solve the proposed model iteratively with the proof of convergence. We use a single parameter r to control the weights for different feature modalities, avoiding the trouble of tuning lots of parameters. Our method has been evaluated on five benchmark datasets and achieves the best performance with comparison to five state-of-art methods in terms of macro and micro classification accuracy. In the future work, we will adjust the proposed method to the additional text or attribute feature modality evaluated on the recently widely studied large image dataset with associated tags or attribute. With the help of more general text or attribute representations other than visual fea-

tures only, we hope to explore more powerful shared common label matrix and improve the state-of-art performance on the corresponding benchmark datasets.

Appendix A

Proof of Eq. (14):

$$
\begin{aligned}
&\sum_v Tr((G^{(v)})^T\tilde{L}^{(v)}G^{(v)}) + \lambda\sum_v Tr(G-G^{(v)})^T(G-G^{(v)}) \\
&= \sum_v Tr(G^T\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}\tilde{L}^{(v)}(\tilde{L}^{(v)}+\lambda I)^{-1}G) \\
&+\lambda\sum_v Tr(G^T(I-\lambda(\tilde{L}^{(v)}+\lambda I)^{-1})^2G) \\
&= Tr(G^T(\sum_v(\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}\tilde{L}^{(v)}(\tilde{L}^{(v)}+\lambda I)^{-1} \\
&+\lambda(I-\lambda(\tilde{L}^{(v)}+\lambda I)^{-1})^2)G) \\
&= Tr(G^T(\sum_v(\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}\tilde{L}^{(v)}(\tilde{L}^{(v)}+\lambda I)^{-1} \\
&\quad +\lambda(I-2\lambda(\tilde{L}^{(v)}+\lambda I)^{-1}+\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-2})G) \\
&= Tr(G^T(\sum_v(\lambda I-2\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}+\lambda(\tilde{L}^{(v)}+\lambda I)^{-1}) \\
&\quad +\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}(\tilde{L}^{(v)}(\tilde{L}^{(v)}+\lambda I)^{-1})G) \\
&= Tr(G^T(\sum_v(\lambda I-2\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1}+\lambda^2(\tilde{L}^{(v)}+\lambda I)^{-1})G) \\
&= \lambda Tr(G^T(\sum_v(I-\lambda(\tilde{L}^{(v)}+\lambda I)^{-1})G)\square
\end{aligned}
$$

References

[1] X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. *International Joint Conference on Artificial Intelligence*, pages 2598–2604, 2013.

[2] X. Cai, F. Nie, H. Huang, and F. Kamangar. Heterogeneous image feature integration via multi-modal spectral clustering. In *CVPR*, pages 1977–1984, 2011.

[3] L. Cao, J. Luo, F. Liang, and T. Huang. Heterogeneous feature machines for visual recognition. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1095–1102. IEEE, 2010.

[4] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):27, 2011.

[5] H. Chen, X. Cai, D. Zhu, F. Nie, T. Liu, and H. Huang. Group-wise consistent parcellation of gyri via adaptive multi-view spectral clustering of fiber shapes. *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 271–279, 2012.

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

[7] C. H. Q. Ding, R. Jin, T. Li, and H. D. Simon. A learning framework using green's function and kernel regularization with application to recommender system. In *KDD*, pages 260–269, 2007.

[8] D. Dueck and B. J. Frey. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, pages 1–8, 2007.

[9] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, 2004.

[10] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[11] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[12] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[13] Y. J. Lee and K. Grauman. Foreground focus: Unsupervised learning from partially matching images. *International Journal of Computer Vision*, 85(2):143–166, 2009.

[14] F.-F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[17] 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.

[18] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

[19] H. Wang, F. Nie, and H. Huang. Multi-view clustering and feature learning via structured sparsity. *International Conference on Machine Learning*, pages 352–360, 2013.

[20] H. Wang, F. Nie, H. Huang, and C. Ding. Heterogeneous visual features fusion via sparse multimodal machine. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3097–3102, 2013.

[21] H. Wang, F. Nie, H. Huang, 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.

[22] J. M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, pages 756–763, 2005.

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

[24] H. Yu, M. Li, H. Zhang, and J. Feng. Color texture moments for content-based image retrieval. In *ICIP (3)*, pages 929–932, 2002.

[25] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.

[26] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2003.

[27] D. Zhou and B. Schölkopf. Learning from labeled and unlabeled data using random walks. In *DAGM-Symposium*, pages 237–244, 2004.

[28] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.