MULTIMODAL SEMI-SUPERVISED IMAGE CLASSIFICATION BY COMBINING TAG REFINEMENT, GRAPH-BASED LEARNING AND SUPPORT VECTOR REGRESSION

Wenxuan Xie, Zhiwu Lu, Yuxin Peng[*] *and Jianguo Xiao*

Institute of Computer Science and Technology, Peking University, Beijing 100871, China

ABSTRACT

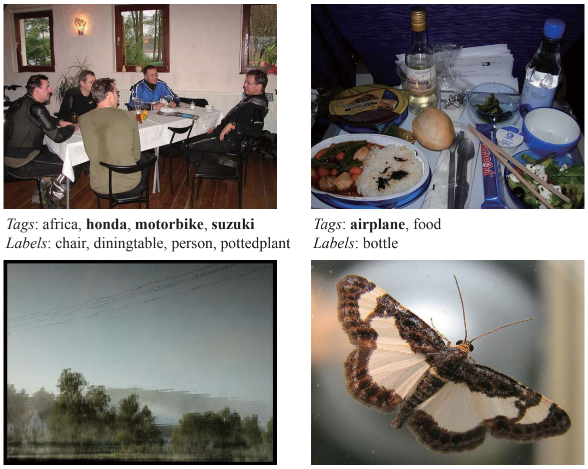
We investigate an image classification task where the training images come along with tags, but only a subset being labeled, and the goal is to predict the class label of test images without tags. This task is crucial for image search engine on photo sharing websites. In previous work, it is handled by first learning a multiple kernel learning classifier using both image content and tags to score unlabeled training images, and then building up a least-squares regression (LSR) model on visual features to predict the label of test images. However, there exist three important issues in the task: 1) Image tags on photo sharing websites tend to be inaccurate and incomplete, and thus refining them is beneficial; 2) Supervised learning with a limited number of labeled samples may be unreliable to some extent, while a graph-based semi-supervised approach can be adopted by also considering similarities of unlabeled data; 3) LSR is established upon centered visual kernel columns and breaks the symmetry of kernel matrix, whereas support vector regression can readily use the original visual kernel and thus leverage its full power. To handle the task more effectively, we propose to combine tag refinement, graph-based learning and support vector regression together. Experimental results on the PASCAL VOC'07 and MIR Flickr datasets show the superior performance of the proposed approach.

Index Terms— Tag refinement, Graph-based semi-supervised learning, Support vector regression

1. INTRODUCTION

Image classification, whose goal is to determine whether an image belongs to a certain category or not, has been studied for decades. In the literature, different types of categories have been considered, e.g., scenes [1] or objects [2]. To tackle an image classification problem, a supervised learning framework can be used, where a binary classifier is first learned from manually labeled training images and then used to predict the class label of test images. The learned classifier can be enhanced by increasing the quantity and diversity of manually labeled training images. However, manually labeling images is a time-consuming task. In practice, we often have to handle a challenging image classification problem by using only a limited number of labeled samples. In the literature, semi-supervised learning [3] has been proposed to exploit the large number of unlabeled samples and thus helps to handle the scarcity of labeled samples to some extent.

In this paper, we investigate a multimodal semi-supervised image classification problem originally raised in [4]. In this problem, the training images have associated tags (e.g., from Flickr), and only a few of the training samples have class labels. The goal is to predict the class label of test images without tags. This is a crucial problem for image search engine on photo sharing websites, since a newly uploaded images and also a considerable part of the existing images have no tags. To solve this problem, a two-step method was proposed in [4]. In the first step, a multiple kernel learning (MKL) [5] classifier is learned by utilizing both image content and tags, which is then used to score unlabeled training images. In the second step, a least-squares regression (LSR) model is learned on the training set by using centered classification scores and centered visual kernel columns, which is then used to predict the class label of test images.



Tags: africa, **honda**, **motorbike**, **suzuki**
Labels: chair, diningtable, person, pottedplant

Tags: **airplane**, food
Labels: bottle

Tags: **water**, reflection, **river**, morning
Labels: plant life, sky, structures, tree

Tags: **night**, naturesfinest, usa, insect
Labels: animals

Fig. 1. Example images from PASCAL VOC'07 (top row) and MIR Flickr (bottom row) datasets with their associated tags and class labels. Tags in **bold** are inaccurate ones.

However, *three* important issues need to be concerned to deal with this problem. *Firstly*, image tags on photo sharing websites (e.g., Flickr) tend to be inaccurate and incomplete, i.e., they may not directly relate to the image content and typically only a few relevant tags are associated with each image. Some example images are shown in Fig. 1. Since the original tags are imperfect, it is a suboptimal choice to directly use them. Hence, we propose to refine tags of training images as the first step.

Secondly, as the number of labeled samples on the training set is limited, an MKL classifier learned by using only these samples may be unreliable to some extent. With this in mind, we propose to use a graph-based semi-supervised learning method which can fully leverage the unlabeled samples from the training set. Empirical results show that the graph-based method can score unlabeled training images more effectively.

Thirdly, the LSR model used in [4] is based on centered visual kernel columns, which breaks the symmetry of the visual kernel ma-

[*] Corresponding author.

978-1-4799-2341-0/13/$31.00 ©2013 IEEE

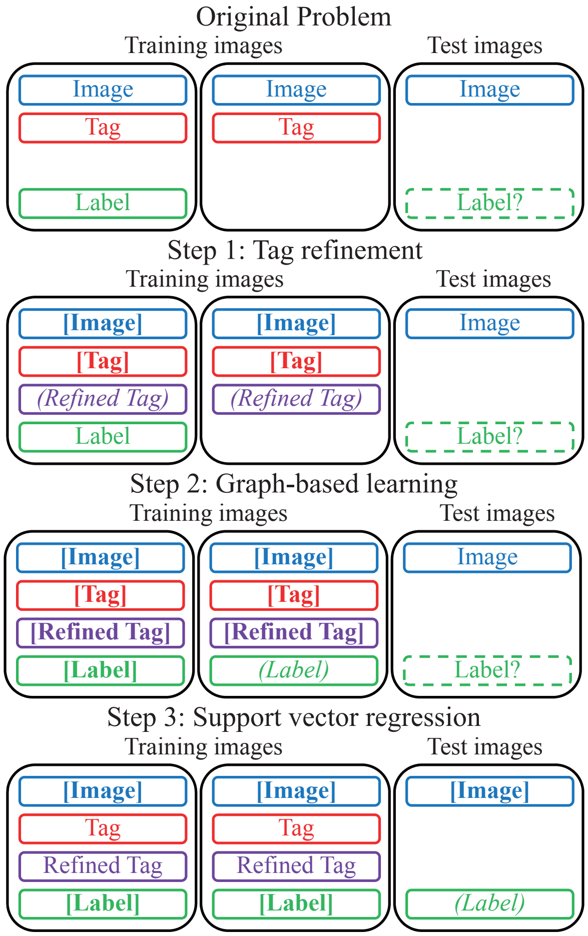


Fig. 2. Illustration of handling multimodal semi-supervised image classification by combining tag refinement, graph-based learning and support vector regression. Inputs and outputs of the corresponding step are denoted by **bold** words in square brackets and *italic* words in parentheses, respectively.

trix. Moreover, the singular value decomposition (SVD) step in the LSR model is time-consuming. Instead of LSR, we propose to use support vector regression (SVR) to predict the class label of test images, since SVR can readily leverage the original visual kernel and make full use of image features in the reproducing kernel Hilbert space (RKHS).

To summarize, we propose to tackle the aforementioned multimodal semi-supervised image classification problem by combining tag refinement, graph-based learning and support vector regression together. The schematic overview of the proposed approach is shown in Fig. 2. The goal is to predict the class label of test images without tags. First of all, we perform tag refinement for the training samples by using the original images and tags. Then, as the next step, class labels of all training samples are predicted by using graph-based semi-supervised learning, where the similarities of the training samples are determined by all the feature representations (i.e., original images, original tags, and refined tags). Finally, the class label of test images are predicted by an SVR model, which can be readily built up by using image features and class labels of training samples.

We conduct experiments on two publicly available datasets to evaluate the proposed approach. Experimental results show not only that tag refinement is beneficial for the multimodal semi-supervised image classification task, but also that the proposed approach performs significantly better than existing methods.

The rest of the paper is organized as follows. In Section 2, we discuss the proposed method in detail. In Section 3 and Section 4, we present experimental setup and results, respectively. Finally, Section 5 draws the conclusions.

2. OUR METHOD

We propose to handle the aforementioned problem by tackling the *three* important problems mentioned in Section 1. We first present the formalized definition of the problem, and then introduce the three components of our method in detail. Finally, the time complexity of the proposed algorithm is analyzed.

2.1. Problem definition

We denote by $I_{tr} = \{x_1, x_2, \ldots, x_{n_1}\}$ the training image set and $I_{te} = \{x_{n_1+1}, x_{n_1+2}, \ldots, x_{n_1+n_2}\}$ the test image set, respectively. Note that $n = n_1 + n_2$ is the total number of samples. Training images come along with tags, which is represented by a binary matrix $T_{tr} \in \{0,1\}^{n_1 \times m}$ indicating tag presence and m is the total number of unique tags. Moreover, only a few of the training images are assigned with class labels from c categories, and the label matrix is denoted as $Y_{tr} \in \{-1,0,1\}^{n_1 \times c}$, whose element $Y_{tr}(i,j)$ indicates the label of image x_i, i.e., $Y_{tr}(i,j) = 1/-1$ if x_i is labeled as a positive/negative sample of category j, and $Y_{tr}(i,j) = 0$ if x_i is unlabeled. The goal is to predict the class label of test images without tags, i.e., a $n_2 \times c$ matrix Y_{te}.

2.2. Tag refinement

As shown in Fig. 1, image tags tend to be inaccurate and incomplete. Instead of directly using the original imperfect tags, we propose to refine them with the help of image content. Although there have already been many approaches to tag refinement in the literature [6, 7], we adopt the local and global consistency method [8], given that our focus in this paper is proposing a significantly better solution to the aforementioned image classification problem.

We denote the visual kernel of training samples by K_{tr}^v, and the normalized Laplacian $L_{tr}^v = I - D^{-1/2} K_{tr}^v D^{-1/2}$, where D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th column of K_{tr}^v. The objective function for tag refinement is shown as follows.

$$\min_{T_{tr*}} \alpha_1 \operatorname{tr}(T_{tr*}^\top L_{tr}^v T_{tr*}) + (1-\alpha_1)\|T_{tr*} - T_{tr}\|_F^2 \qquad (1)$$

The analytical solution of Eq. 1 is given by $T_{tr*} = (1-\alpha_1)(\alpha_1 L_{tr}^v + (1-\alpha_1)I)^{-1} T_{tr}$.

2.3. Graph-based semi-supervised learning

After obtaining refined image tags, we may have a better similarity measure of training samples. The next issue is to infer the class label of unlabeled training images. We propose to use a graph-based semi-supervised learning method to tackle this issue by fully leveraging unlabeled samples. To be consistent with the aforementioned tag refinement procedure, we similarly adopt the local and global consistency method [8]. We denote by K_{tr} the kernel and L_{tr} the corresponding normalized Laplacian, and thus the objective function for scoring unlabeled training images is shown as follows.

$$\min_{Y_{tr*}} \alpha_2 \operatorname{tr}(Y_{tr*}^\top L_{tr} Y_{tr*}) + (1-\alpha_2)\|Y_{tr*} - Y_{tr}\|_F^2 \tag{2}$$

The analytical solution is given similarly as that of Eq. 1, i.e., $Y_{tr*} = (1-\alpha_2)(\alpha_2 L_{tr} + (1-\alpha_2)I)^{-1} Y_{tr}$.

It should be noted that most of the elements in Y_{tr*} have small absolute values (i.e., close to 0), which may lead to inferior final results. To imitate the decision values outputted by an SVM classifier, we use a simple algorithm to normalize Y_{tr*} as follows. Note that we define Y_{tr*}^1 as the subset of Y_{tr*} where the corresponding original labels in Y_{tr} equals 1 (i.e., positive), and we can similarly define Y_{tr*}^{-1} and Y_{tr*}^0.

$$\begin{aligned} &Y_{tr*}^1 \longleftarrow 1,\ Y_{tr*}^{-1} \longleftarrow -1 \\ &Y_{tr*}^0 \longleftarrow Y_{tr*}^0 - \frac{1}{2}(\max(Y_{tr*}^0) + \min(Y_{tr*}^0)) \\ &Y_{tr*}^0 \longleftarrow Y_{tr*}^0 / \max(Y_{tr*}^0) \end{aligned} \tag{3}$$

2.4. Support vector regression

After obtaining scores of all training samples, the class label of test images can be inferred by learning a classification or regression model. Since the predicted scores of training samples are continuous (i.e., not quantized to 0 or 1), a regression model is preferred. In [4], SVD is performed on the centered kernel matrix for K_{tr}^v (i.e., each column of K_{tr}^v is normalized to 0 mean), and the regression coefficients can be computed by multiplying the pseudo-inverse matrix of K_{tr}^v (which can be easily obtained after performing SVD) by the centered scores of training samples.

However, the symmetry of the visual kernel matrix K_{tr}^v is broken in [4], and the SVD step is time-consuming. In order to directly leverage K_{tr}^v and make the learning algorithm more efficient, we propose to use SVR as the regression model. Similar to the SVM classifier, SVR can be kernelized to fully leverage image features in the RKHS along with the continuous predicted scores of all training samples. The scores predicted by SVR are the final results of the aforementioned problem.

2.5. Complexity analysis

Recall that we denote by n the sample size. Since training sample size and test sample size have the same orders of magnitude, we do not explicitly distinguish between them. The method proposed in [4] consists of an MKL classifier and an LSR model. The most time-consuming step is the SVD of the centered visual kernel matrix, where the time complexity is $O(n^3)$.

As a comparison, the proposed method is made up of tag refinement, graph-based learning and SVR. The most time-consuming step is the inversion of an $n \times n$ matrix when computing the analytical solution to a semi-supervised problem, where the time complexity is also $O(n^3)$. However, we can adopt the iterative step suggested in [8] to accelerate the semi-supervised learning algorithm, and thus the complexity of our method can be reduced to $O(n^d)$ with $d \in (2, 3)$, which is determined by the SVR implementation. Therefore, our algorithm is more efficient than that proposed in [4].

3. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed approach, we conduct experiments on two publicly available datasets: PASCAL VOC'07 [2] and MIR Flickr [9], both of which have been used in [4]. In particular, there are 9,963 images from 20 categories with 804 tags on the PASCAL VOC'07 set and 25,000 images from 38 categories with 457 tags on the MIR Flickr set. Moreover, the PASCAL VOC'07 set is split into a training set of 5,011 images and a test set of 4,952 images, and the MIR Flickr set is split into a training set of 12,500 images and a test set of 12,500 images.

There are $M = 15$ different image representations and a tag representation on both datasets. We use the same visual kernel as that in [4]. Specifically, we average the distances between images based on these different representations, and use it to compute an RBF kernel, which is defined as

$$k^v(x_i, x_j) = \exp(-\lambda^{-1} d(x_i, x_j)) \tag{4}$$

where the scale factor λ is set to be the average pairwise distance, $\lambda = n^{-2} \sum_{i,j=1}^{n} d(x_i, x_j)$, and the pairwise distance is defined as $d(x_i, x_j) = \sum_{m=1}^{M} \lambda_m^{-1} d_m(x_i, x_j)$, where $\lambda_m = \max_{i,j} d_m(x_i, x_j)$. Following the settings in [4], we use L1 distance for the color histograms, L2 for GIST, and χ^2 for the visual word histograms. In addition, we compute the cosine similarity kernel for tag features.

There are two tunable parameters in our model, i.e., α_1 and α_2. We empirically set $\alpha_1 = 0.9$ and $\alpha_2 = 0.1$ for both datasets. Moreover, we use the LIBSVM implementation [10] for SVR, where the regularization parameter is set to be $C = 10$, following the SVM settings in [4].

4. EXPERIMENTAL RESULTS

In our experiments, we evaluate results using the average precision (AP) criterion for each class, and also using the mean AP (mAP) over all classes. To be in accordance with [4], we adopt the evaluation criterion in the PASCAL VOC challenge [2], which is given as

$$AP = \frac{1}{11} \sum_r P(r) \tag{5}$$

where $P(r)$ denotes the maximum precision over all recalls larger than $r \in \{0, 0.1, 0.2, \ldots, 1.0\}$. A larger value indicates a better performance. It should be noted that, all the AP scores are computed based on the ranked lists of all test samples. To demonstrate the effectiveness of our approach, we compare the following methods:

- MKL+LSR[4]: An MKL classifier learned on labeled training samples, followed by least-squares regression on the MKL scores for all training samples to obtain the visual classifier.
- GSSL+SVR(ours): A graph-based semi-supervised learning method based on a combined kernel K_{tr} by averagely fusing visual kernel and tag kernel, followed by SVR on the normalized decision values of all training samples to predict the scores of test samples.
- TR+GSSL+SVR(ours): Tag refinement by using the local and global consistency method [8], followed by a graph-based semi-supervised learning method based on a combined kernel K_{tr} by averagely fusing visual kernel, tag kernel and refined tag kernel. Finally, SVR is learned on the normalized decision values.

It should be noted that there is also a related paper [11] on multi-label image classification using the same datasets, where the authors assume the class label vectors (i.e., all class label assignments) of

Table 1. AP scores for all the classes of the two datasets using 50 positive and 50 negative labeled examples for each class.

PASCAL VOC'07	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable
MKL+LSR[4]	0.592	0.324	0.376	0.519	0.154	0.278	0.501	0.366	0.300	0.117	0.255
GSSL+SVR(ours)	0.641	0.439	**0.394**	**0.540**	**0.175**	**0.443**	0.471	0.425	0.317	0.271	0.226
TR+GSSL+SVR(ours)	**0.649**	**0.447**	0.388	0.525	0.143	0.422	**0.548**	**0.430**	**0.323**	**0.276**	**0.277**
	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor		Mean
MKL+LSR[4]	0.331	0.637	0.383	0.703	0.212	0.218	0.191	0.617	0.236		0.366
GSSL+SVR(ours)	0.342	0.643	0.396	**0.699**	**0.283**	0.308	0.189	0.678	0.188		0.403
TR+GSSL+SVR(ours)	**0.347**	**0.698**	**0.478**	0.690	0.262	**0.331**	**0.224**	**0.690**	**0.300**		**0.422**

MIR Flickr	animals	baby	baby∗	bird	bird∗	car	car∗	clouds	clouds∗	dog	dog∗	female	female∗
MKL+LSR[4]	0.310	0.075	0.161	0.124	0.163	0.229	0.305	0.612	0.537	0.182	0.212	**0.440**	0.313
GSSL+SVR(ours)	0.324	**0.171**	0.193	0.160	**0.191**	0.126	**0.444**	**0.649**	0.567	0.262	**0.280**	0.416	**0.396**
TR+GSSL+SVR(ours)	**0.355**	0.142	**0.227**	**0.173**	0.190	**0.282**	0.422	0.640	**0.577**	**0.269**	0.270	0.422	0.307
	flower	flower∗	food	indoor	lake	male	male∗	night	night∗	people	people∗	plant life	portrait
MKL+LSR[4]	0.373	0.424	0.333	0.514	0.159	0.366	0.255	0.471	0.368	0.629	0.554	0.613	0.474
GSSL+SVR(ours)	**0.443**	0.484	0.347	**0.578**	0.206	**0.400**	0.246	0.425	0.465	0.656	0.547	0.595	0.398
TR+GSSL+SVR(ours)	0.432	**0.529**	**0.419**	0.543	**0.215**	0.353	**0.257**	**0.542**	**0.468**	**0.673**	**0.578**	**0.630**	**0.488**
	portrait∗	river	river∗	sea	sea∗	sky	structures	sunset	transport	tree	tree∗	water	Mean
MKL+LSR[4]	0.429	**0.234**	0.047	0.437	0.255	0.693	0.655	0.543	0.321	**0.453**	0.231	0.452	0.367
GSSL+SVR(ours)	0.449	0.189	**0.094**	**0.457**	0.274	0.708	0.632	**0.584**	**0.335**	0.347	0.351	0.468	0.391
TR+GSSL+SVR(ours)	**0.498**	0.190	0.076	0.442	**0.284**	**0.712**	**0.669**	0.565	0.266	0.432	**0.393**	**0.489**	**0.406**

Table 2. Performance in mAP on the two datasets using 20/50/100 positive and 20/50/100 negative labeled examples for each class.

PASCAL VOC'07	20	50	100
MKL+LSR[4]	0.336	0.366	0.406
GSSL+SVR(ours)	0.369	0.403	0.445
TR+GSSL+SVR(ours)	**0.385**	**0.422**	**0.461**
MIR Flickr	20	50	100
MKL+LSR[4]	0.316	0.367	0.395
GSSL+SVR(ours)	0.343	0.391	0.421
TR+GSSL+SVR(ours)	**0.354**	**0.406**	**0.430**

some given samples are already known. However, following the problem settings in [4], we randomly choose positive and negative samples for only one class at a time. Most probably, different labeled samples are chosen for different classes, and thus the aforementioned problem does not belong to a multi-label classification problem. Due to different settings of input class labels, we do not make direct comparisons with the results reported in [11].

We list in Table 1 the per-class results using 50 positive and 50 negative labeled examples for each class. Due to the space limit, we report in Table 2 the mAP scores for both datasets with varying amounts of labeled data. The averaged performance over 10 random sampling of labeled training images is reported in the tables. We can observe that *GSSL+SVR* performs significantly better than *MKL+LSR*, and that *TR+GSSL+SVR* can induce a further improvement over *GSSL+SVR*. The improvement is mainly due to our solutions to the three issues mentioned in Section 1. Although MKL [5] is a powerful tool, it performs unsatisfactorily when using only a limited number of labeled samples. In contrast, graph-based semi-supervised learning is good at dealing with such problems. The LSR model is based on centered visual columns, while SVR can readily use the original visual kernel and thus leverage its full power. In addition, tag refinement plays a key role in further improving the results, since the original tags are inaccurate and incomplete.

5. CONCLUSION

In this paper, we propose to tackle the multimodal semi-supervised image classification problem by combining tag refinement, graph-based learning and support vector regression together. In the experiments, our method is shown to achieve significantly better results than existing methods. The improvement mainly lies in the graph-based semi-supervised learning method which is good at classification with a limited number of labeled samples. The SVR model also plays an important role, since it can fully leverage image features in the RKHS. Moreover, tag refinement can lead to a further improvement upon the two aforementioned components.

6. ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China under Grants 61073084 and 61202231, Beijing Natural Science Foundation of China under Grants 4122035 and 4132037, Ph.D. Programs Foundation of Ministry of Education of China under Grants 20120001110097 and 20120001120130, National Hi-Tech Research and Development Program (863 Program) of China under Grant 2012AA012503, and National Key Technology Research and Development Program of China under Grant 2012BAH07B01.

7. REFERENCES

[1] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. 2, pp. 2169–2178.

[2] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[3] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

[4] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 902–909.

[5] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, et al., "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[6] L. Chen, D. Xu, I.W. Tsang, and J. Luo, "Tag-based web photo retrieval improved by batch mode re-tagging," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3440–3446.

[7] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 461–470.

[8] D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, pp. 321–328, 2004.

[9] M.J. Huiskes and M.S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.

[10] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[11] Y. Luo, D. Tao, B. Geng, C. Xu, and S. Maybank, "Manifold regularized multi-task learning for semi-supervised multi-label image classification," *IEEE Transactions on Image Processing*, 2012.