Shared Feature Extraction for Semi-supervised Image Classification*

Yong Luo[†] , Dacheng Tao[‡] , Bo Geng[†], Chao Xu[†], and Stephen Maybank[§]
[†]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China
[‡]Faculty of Engineering and Information Technology, University of Technology, Sydney, Sydney, Australia
[§]Department of Computer Science and Information Systems, Birkbeck College, Malet Street, London, UK
{luoyong, gengbo, xuchao}@cis.pku.edu.cn, Dacheng.Tao@uts.edu.au, sjmaybank@dcs.bbk.ac.uk

ABSTRACT

Multi-task learning (MTL) plays an important role in image analysis applications, e.g. image classification, face recognition and image annotation. That is because MTL can estimate the latent shared subspace to represent the common features given a set of images from different tasks. However, the geometry of the data probability distribution is always supported on an intrinsic image sub-manifold that is embedded in a high dimensional Euclidean space. Therefore, it is improper to directly apply MTL to multiclass image classification. In this paper, we propose a manifold regularized MTL (MRMTL) algorithm to discover the latent shared subspace by treating the high-dimensional image space as a sub-manifold embedded in an ambient space. We conduct experiments on the PASCAL VOC'07 dataset with 20 classes and the MIR dataset with 38 classes by comparing MRMTL with conventional MTL and several representative image classification algorithms. The results suggest that MRMTL can properly extract the common features for image representation and thus improve the generalization performance of the image classification models.

Categories and Subject Descriptors

I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement—*feature representation*; I.5.2 [**Pattern Recognition**]: Design Methodology—*pattern analysis*

General Terms

Algorithm, Experimentation, Theory

Keywords

Image classification, manifold regularization, multi-task learning, semi-supervised

1. INTRODUCTION

In real image analysis applications, e.g. image classification and face recognition, labeling is time consuming while large amounts of unlabeled images are available. Semi-supervised learning (SSL) can improve the generalization ability of supervised learning by leverage of the unlabeled samples under some circumstances [4, 3]. Particularly, in semi-supervised image classification, we usually have to predict multiple labels and a typical method is to divide it into multiple binary classification tasks [16]. The traditional method is to learn each task separately. However, learning all of these tasks simultaneously can be advantageous by utilizing the multi-task learning (MTL) framework [1].

MTL is an approach to learn a task together with other related tasks at the same time, using a shared representation. This can often lead to a better model for the main task, because it allows the learner to use the commonality among the tasks. MTL has been widely used in various image analysis applications. In [5], it was applied to locating doors and recognizing door types from image pixel level features. Torralba et al. [13] applied MTL to object detection by utilizing the common features shared by different object classes. Since then, a lot of consequent results have been obtained for face recognition and image annotation.

MTL makes sense in these applications because we usually have a large number of classes but a limited number of the labeled training images from each class. By the use of MTL, images from related tasks can be combined together to jointly discover the shared features. These features are useful for each particular task. Besides, it has proven that the number of training samples required to train each task decreases linearly with the increasing number of tasks [2]. Therefore, given a large number of image classification tasks, it is feasible to obtain satisfied classification performance by using a very small number of labeled samples from each task.

However, it is not advisable to directly apply MTL to the high dimensional Euclidean space because images represented by a specific feature space, lie on a very low-dimensional image sub-manifold embedded in the ambient space. In this situation, a small amount of labeled samples cannot represent the true underlying data distribution. Thus the predictive function estimated with a risk minimization principle only using the labeled samples is not optimal. To solve this problem for MTL based image classification, we propose to minimize the joint empirical risks (JER) in MTL and approximate the sub-manifold together to obtain a robust model, which can smooth the predictive functions along the sub-manifold. This is motivated by the manifold regularization (MR) [3] framework. MR is a data-dependent regularization that exploits the geometry of the probability distribution. This geometry can be used to approximate the

*Area chair: Bernard Merialdo

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

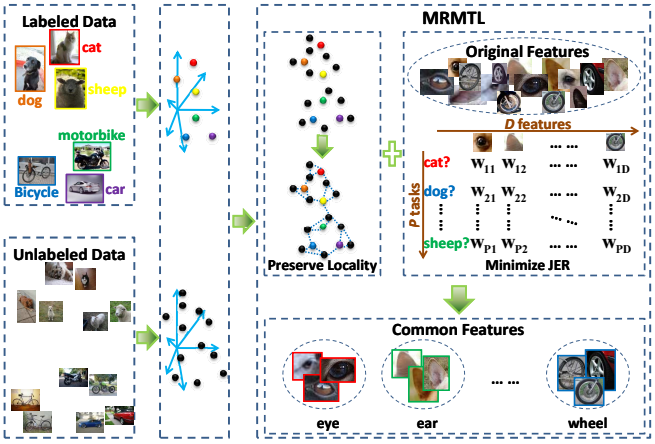

Figure 1: The framework of our proposed Manifold Regularized Multi-Task Learning algorithm.

true underlying image sub-manifold. MR has widely applied to many areas, e.g., image classification [7], image search ranking [8, 9], query suggestion [17], and video annotation [15, 14].

It is worthy emphasizing that the key in learning with MR is constructing a data adjacency graph to encode the low-dimensional sub-manifold. In single task learning, the amount of data may be insufficient for estimating the true underlying structure. The problem is lessened in MTL because the data used to construct the graph are from multiple tasks.

Therefore, we propose manifold regularized MTL (MRMTL) to discover the latent shared subspace of different tasks. Given several visual representations for image, we aim to learn a corresponding mapping by using MRMTL for each of them. We first calculate the graph Laplacian of the whole data set and conduct MRMTL to obtain the optimal transformation. Subsequently, original features are mapped to the low-dimensional shared subspace across labels by this transformation. Finally, we calculate kernels in the original features and cross-label features respectively and linearly combine the two kernels. We conduct extensive experiments on the PASCAL VOC'07 dataset with 20 classes [6] and the MIR dataset with 38 classes [11] by comparing MRMTL with conventional MTL [1] and several representative semi-supervised image classification algorithms [4, 10]. The experimental results demonstrate the effectiveness of the proposed MRMTL.

2. MANIFOLD REGULARIZED MULTI-TASK LEARNING

MRMTL is a particular implementation of SSL. Fig.1 shows the diagram of MRMTL for image classification: Given both labeled and unlabeled images, we first extract visual features for image representation. A data adjacency graph is then constructed by the use of all images from different tasks to generate a manifold regularization term. Multiple image classification tasks, e.g. "whether it is a dog or not?", are created and we can then find an optimal feature mapping by minimizing JER of these predictors with MR.

2.1 Notations

We use N, M, D and P to denote the number of training images, the number of different kinds of visual features for image representation, the data dimensionality, and the number of labels, respectively. The subscripts l, u and lu signify *labeled*, *unlabeled*, and *labeled + unlabeled*. The data matrix and the label indicator matrix for the mth visual feature are denoted as $X^m = [\mathbf{x}_1^m, \ldots, \mathbf{x}_N^m] \in \mathbb{R}^{D\times N}$ and $Y \in \mathbb{R}^{P\times N}$, where $\mathbf{x}_n^m \in \mathbb{R}^D$ is the nth instance, and $Y_{pn} = 1$ if the pth label is assigned to the nth instance, and -1 otherwise. The superscript m is omitted in our formulation because we handle each kind of feature in the same way.

2.2 Multi-Task Learning with MR

Given a set of N_l labeled samples $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N_l}, y_n \in \{+1, -1\}$ drawn from a probability distribution $\mathcal{P}$ and a set of N_u unlabeled samples $\{\mathbf{x}_n\}_{n=N_l+1}^{N_l+N_u}$ generated according to the marginal distribution $\mathcal{P}_X$ of $\mathcal{P}$, the manifold regularization framework is to estimate an unknown function f by minimizing

$$\underset{f\in\mathcal{H}_K}{\operatorname{argmin}} \frac{1}{N_l}\sum_{n=1}^{N_l} L(\mathbf{x}_n, y_n, f) + \gamma_A\|f\|_K^2 + \gamma_I\|f\|_I^2, \qquad (1)$$

where $\mathcal{H}_K$ is an associated Reproducing Kernel Hilbert Space (RKHS) and L is a prescribed loss function. $\|f\|_K^2$ penalizes the classifier complexities in the ambient space. $\|f\|_I^2$ is a smoothness penalty corresponding to the probability distribution. Both γ_A and γ_I are trade-off parameters. Following [3], $\|f\|_I^2$ can be approximated by $\frac{1}{(N_l+N_u)^2}\mathbf{f}^T\mathcal{L}\mathbf{f}$, where $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_{N_l+N_u})]^T$, $\frac{1}{(N_l+N_u)^2}$ is the normalizing coefficient and $\mathcal{L}$ is the graph Laplacian given by $\mathcal{L} = \mathcal{D} - \mathcal{W}$. Here, $\mathcal{W}_{ij}$ is the edge weight, e.g., the heat kernel weight $\mathcal{W}_{ij} = e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/4t}$ in the data adjacency graph constructed by using k nearest neighbors and the diagonal matrix $\mathcal{D}$ is given by $\mathcal{D}_{ii} = \sum_{j=1}^{N_l+N_u}\mathcal{W}_{ij}$.

The construction of the graph Laplacian needs large amounts of data. Now we learn multiple tasks at the same time and have enough data to construct a graph to approximate the true underlying data manifold. Thus, the proposed manifold regularized multi-task learning (MRMTL) can be written as:

$$\underset{\{f_p\}\in\mathcal{H}_K}{\operatorname{argmin}} \sum_{p=1}^{P}\left(\frac{1}{N_{lp}}\sum_{n=1}^{N_{lp}} L(\mathbf{x}_n^p, y_n^p, f_p) + \gamma_A\|f_p\|_K^2 + \gamma_I\|f_p\|_I^2\right), \qquad (2)$$

where $y_n^p = Y_{pn}$.This formulation differs from the traditional MTL framework in the added MR term, which is helpful for learning with image features as illustrated in section 1.

In MTL, the predictive function for the pth task(label) is,

$$f_p(\mathbf{x}) = \mathbf{w}_p^T\mathbf{x} + \mathbf{v}_p^T\Theta\mathbf{x}, \qquad (3)$$

where $\mathbf{w} \in \mathbb{R}^D$ and $\mathbf{v} \in \mathbb{R}^r$ are the weight vectors, $\Theta \in \mathbb{R}^{r\times D}$ is the linear transformation parameterizing the latent shared low-dimensional subspace, and r is the dimensionality of the latent shared subspace. The transformation Θ is common for all labels, and it has the orthogonality $\Theta\Theta^T = I$. We simplify the general formulation in (2) by assuming that the input data for each task are identical and we propose to learn an optimal feature map Θ from the data in the original feature space by minimizing the following regularized empirical risk:

$$\underset{\Theta}{\operatorname{argmin}} \sum_{p=1}^{P}\left(\frac{1}{N_l}\sum_{n=1}^{N_l} L(f_p(\mathbf{x}_n), y_n^p) + \alpha\|\mathbf{w}_p\|^2 + \beta\|\mathbf{w}_p + \Theta^T\mathbf{v}_p\|^2 + \frac{\gamma_I}{(N_l+N_u)^2}\mathbf{f}_p^T\mathcal{L}\mathbf{f}_p\right), \qquad (4)$$

where $\mathbf{f}_p = [f_p(\mathbf{x}_1), \ldots, f_p(\mathbf{x}_{N_l+N_u})]^T$, $\|\mathbf{w}_p\|^2$ is the regularizer used in MTL and $\|\mathbf{w}_p + \Theta^T\mathbf{v}_p\|^2$ controls the complexity

of the models. Both α and β are trade-off parameters. Let $\mathbf{u}_p = \mathbf{w}_p + \Theta^T \mathbf{v}_p$, This problem can be reformulated equivalently as

$$\min_{\Theta} \sum_{p=1}^{P} \left(\frac{1}{N_l} \sum_{n=1}^{N_l} L(\mathbf{u}_p^T \mathbf{x}_n, y_n^p) + \alpha \|\mathbf{u}_p - \Theta^T \mathbf{v}_p\|^2 + \beta \|\mathbf{u}_p\|^2 + \frac{\gamma_I}{(N_l+N_u)^2} \mathbf{f}_p^T \mathcal{L} \mathbf{f}_p \right), \quad s.t. \quad \Theta\Theta^T = I. \tag{5}$$

where $\mathbf{f}_p = [\mathbf{u}_p^T \mathbf{x}_1, \ldots, \mathbf{u}_p^T \mathbf{x}_{N_l+N_u}]^T$. We consider the least squares loss function,

$$L(\mathbf{u}_p^T \mathbf{x}_n, y_n^p) = (\mathbf{u}_p^T \mathbf{x}_n - y_n^p)^2.$$

Therefore, we can simplify (5) as:

$$\min_{\Theta} \frac{1}{N_l} \|X_l^T U - Y^T\|_F^2 + \alpha \|U - \Theta^T V\|_F^2 + \beta \|U\|_F^2 + \frac{\gamma_I}{(N_l+N_u)^2} tr\big((X_{lu}^T U)^T \mathcal{L} (X_{lu}^T U)\big) \quad s.t. \ \Theta\Theta^T = I. \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_P]$, and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_P]$.

This is a convex optimization problem and we can get a closed form solution. Fixing (Θ, U), we obtain the optimal V by taking the derivative of the expression in (6) with respect to V and setting it to be zero.

$$V^* = \Theta U.$$

Similarly, substituting $V^* = \Theta U$ into (6) and fixing Θ, we obtain the optimal U,

$$U^* = \frac{1}{N_l} \big(M - \alpha \Theta^T \Theta\big)^{-1} X_l Y^T, \tag{7}$$

where M is defined as:

$$M = \frac{1}{N_l} X_l X_l^T + (\alpha + \beta) I + \frac{\gamma_I}{(N_l+N_u)^2} X_{lu} \mathcal{L} X_{lu}^T. \tag{8}$$

Finally, we can substitute the expression for (U^*, V^*) into (6) and obtain the optimal Θ. It is direct that this can be done by solving the following trace maximization problem:

$$\max_{\Theta} \ tr\Big(\big(\Theta S_1 \Theta^T\big)^{-1} \Theta S_2 \Theta^T\Big) \quad s.t. \ \Theta\Theta^T = I, \tag{9}$$

where S_1 and S_2 are defined as:

$$S_1 = I - \alpha M^{-1} \tag{10}$$

$$S_2 = M^{-1} X_l Y^T Y X_l^T M^{-1}, \tag{11}$$

The obtained Θ^* can be used to induce a set of cross-label features, i.e., $\Theta^* \mathbf{x}$. We use these features to construct a visual kernel K_{v-comn} and combine it with the original kernel K_v, i.e.,

$$K_{v-new} = \lambda K_v + (1-\lambda) K_{v-comn}, \tag{12}$$

where $\lambda \in [0, 1]$ is the combination parameter. The new visual kernel K_{v-new} contains both the shared information of different image classification tasks extracted by MRMTL and also the discriminative information in the original features. This kernel is further used for semi-supervised image classification.

3. EXPERIMENTS

We evaluate our method on two data sets, the PASCAL VOC'07 [6] and the MIR Flickr [11], which have been used in [10]. There are around 10,000 images of 20 different object categories in the PASCAL VOC'07 set and 25,000 images of 38 categories in the MIR Flickr set. The 15 different kinds of visual representations as described in [10] are extracted. We measure the performance using the average precision (AP) criterion for each class and the mean AP (mAP) over all classes. The number of labeled training examples is 100 (50 positive and 50 negative) in our experiments.

3.1 Extracting Shared Subspace

This set of experiments evaluates the effectiveness of MRMTL. We compare it with two popular methods for extracting shared subspace. The experiments are performed by the use of features in the shared subspace for semi-supervised image classification and the 15 different image representations are used for evaluation individually. The experimental setup is summarized as follows:

- **MRMTL:** The regularization parameters α, β and γ_I are tuned from the candidate set $\{10^i | i = -4, -3, \ldots, 3, 4\}$. The parameters k and t used in computing the Laplacian matrix are tuned from $\{1, 2, \ldots, 10, 20, \ldots, 100\}$ and $\{10^i | i = 2, 3, \ldots, 9\}$ respectively. The performance of the proposed method is not sensitive to the dimensionality of the shared subspace r as long as it is not too small. Hence, it is fixed to $5 \times \lfloor (m-1)/5 \rfloor$, where m is the number of labels.
- **ML-LS:** The multi-label formulation is presented in [12]. The regularization parameters α and β are tuned from the candidate set $\{10^i | i = -4, -3, \ldots, 3, 4\}$.
- **ASO-SVM:** The alternating structural optimization (ASO) algorithm proposed in [1] with hinge loss. The regularization parameter is tuned on the set $\{10^i | i = -4, -3, \ldots, 2, 3\}$.

The experimental results on the two data sets are presented in Figure 2. We observe that the proposed MRMTL approach perform the best for 9 features in the VOC set and 7 features in the MIR set.

3.2 Kernel Combination

In this subsection we use the shared features extracted with MRMTL to calculate a kernel and combine it with the kernel obtained from the original features. We choose the best combination parameter for combining the two kernels with this set of experiments. For the 15 kinds of original features, we average the distances between images based on these descriptors and use it to compute an RBF kernel. For the common features, we average the similarities which are computed based on a linear kernel. We simply combine the two kernels with varying λ to create the new kernel and apply it to classification, the results are shown in Figure 3.

The results indicate that common features ($\lambda = 0$) are not as good as the original features ($\lambda = 1$) for classification tasks on both PASCAL VOC'07 and MIR Flickr. However, by properly combining common features and original features, we can obtain significant performance improvement. In particular, by setting $\lambda = 0.8$, we can obtain 13.7% improvement on the PASCAL VOC'07 dataset and 7.8% improvement on the MIR Flickr dataset compared to the use of only original features.

3.3 Semi-supervised Image Classification with the New Kernel

With the best combination parameter, we calculate the new kernel described in (12) and use it for semi-supervised image classification to further verify the effectiveness of MRMTL. This set of experiments are based on the same setting as that used in [10]. In which, some additional textual information is used. We specifically compare the following methods:

- **SVM**: visual classifier learned on labeled examples.
- **Co-training**: learn a visual classifier and a textual classifier on the labeled data set, and bootstrap training ex-

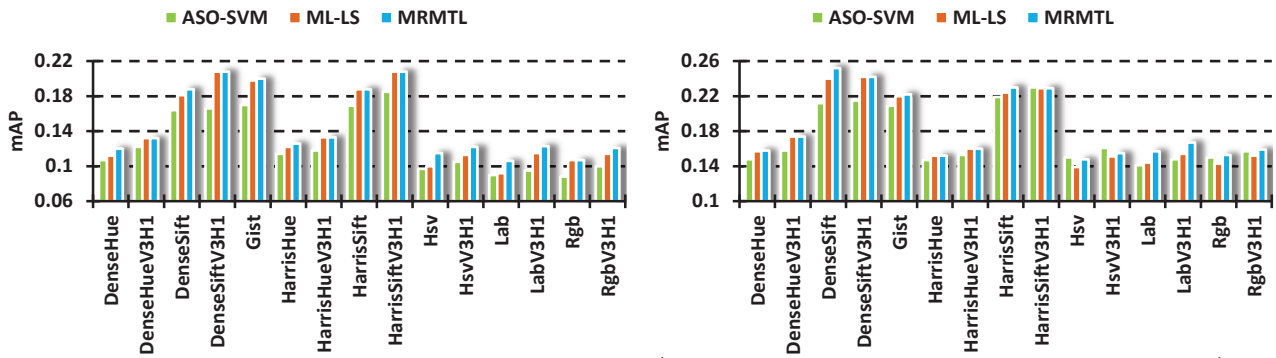

Figure 2: Performance in mAP on the two data sets (Left: PASCAL VOC'07; Right: MIR Flickr) for 15 kinds of features using 50 positive and 50 negative labeled examples for each class.

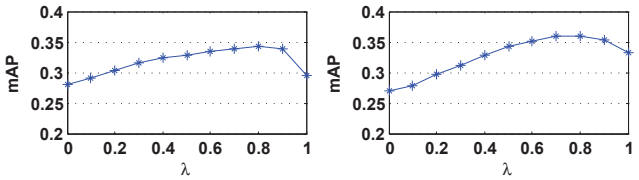

Figure 3: Performance in mAP on the two data sets (Left: PASCAL VOC'07; Right: MIR Flickr) of combining the original kernel and common kernel.

Table 1: Performance in mAP on the two data sets for different learning methods.

	PASCAL VOC'07	MIR Flickr
SVM	0.294	0.333
Co-training	0.323	0.351
MKL+LSR	0.366	0.367
MRMTL+CoTR	**0.385**	**0.386**

amples for each classifier based on the output of the other classifier.

• **MKL+LSR**: the approach proposed in [10]. A multiple kernel learning (MKL) classifier is learned on the labeled examples firstly. Then a classifier (only using the visual information) was obtained by the least squares regression (LSR) on the MKL scores for all examples.

• **MRMTL+CoTR**: simply replace K_v with K_{v-new} in the co-training setup.

The results of MKL+LSR, co-training and baseline SVM presented in [10] are directly used here because we use the same datasets and the same features. The experimental results shown in Table 1 show the superiority of the proposed MRMTL for image classification. The individual APs of the 58 classes are not reported due to the limited page length.

4. CONCLUSION AND DISCUSSION

We present a novel manifold regularized multi-task learning (MRMTL) based algorithm for semi-supervised image classification in this paper. In particular, the algorithm extracts the latent shared subspace among multiple tasks, in which a feature mapping is computed to discover this subspace for each kind of visual feature. Afterward, we transform the data into the shared subspace and a linear combination of the original features and the common features is used to get a new visual kernel. Our experiments demonstrate that 1) MRMTL outperforms popular algorithms for extracting shared subspace on image classification tasks, 2) it is effective to combine both the common features across different tasks with the original visual features.

5. ACKNOWLEDGMENTS

This paper is partially supported by NBRPC 2011CB302400, NSFC 60975014 and NSFB 4102024.

6. REFERENCES

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 2005.

[2] J. Baxter. A model of inductive bias learning. *JAIR*, 2000.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] R. Caruana. Multitask learning. *Machine Learning*, 1997.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[7] B. Geng, C. Xu, D. Tao, L. Yang, and X. Hua. Ensemble manifold regularization. In *CVPR*, 2009.

[8] B. Geng, L. Yang, C. Xu, and X. Hua. Ranking model adaptation for domain specific search. In *CIKM*, 2009.

[9] B. Geng, L. Yang, C. Xu, and X. Hua. Content-aware ranking for visual search. In *CVPR*, 2010.

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.

[11] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR*, 2008.

[12] S. W. Ji, L. Tang, S. P. Yu, and J. P. Ye. Extracting shared subspace for multi-label classification. In *KDD*, 2008.

[13] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[14] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *TCSVT*, 2009.

[15] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *TMM*, 2009.

[16] Z. Zha, X. Hua, T. Mei, J. Wang, G. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *CVPR*, 2008.

[17] Z. Zheng-Jun, Y. Linjun, M. Tao, W. Meng, and W. Zengfu. Visual query suggestion. In *SIGMM*, 2009.