Low Rank Metric Learning for Social Image Retrieval

Zechao Li[†], Jing Liu[†], Jiang Yu[†], Jinhui Tang[‡], Hanqing Lu[†]
[†]National Laboratory Of Pattern Recognition, Institute Of Automation, Chinese Academy Of Sciences
[‡]School of Computer Science, Nanjing University of Science and Technology
zechao.li@gmail.com, {jliu, jyu, luhq}@nlpr.ia.ac.cn, jinhuitang@mail.njust.edu.cn

ABSTRACT

With the popularity of social media applications, large amounts of social images associated with rich context are available, which is helpful for many applications. In this paper, we propose a Low Rank distance Metric Learning (LRML) algorithm by discovering knowledge from these rich contextual data, to boost the performance of CBIR. Different from traditional approaches that often use the must-links and cannot-links between images, the proposed method exploits information from the visual and textual domains. We assume that the visual similarity estimated by the learned metric is expected to be consistent with the semantic similarity in the textual domain. Since tags are usually noisy, misspelling or meaningless, we also leverage the preservation of visual structure to prevent overfitting those noisy tags. On the other hand, the metric is straightforward constrained to be low rank. We formulate it as a convex optimization problem with nuclear norm minimization and propose an effective optimization algorithm based on proximal gradient method. With the learned metric for image retrieval, some experimental evaluations on a real-world dataset demonstrate the outperformance of our approach over other related work.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models; I.2.6 [**Artificial Intelligence**]: Vision and Scene Understanding

General Terms

Algorithms, Experimentation, Theory

Keywords

Metric Learning, Low Rank, Social Images Retrieval

MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$15.00.

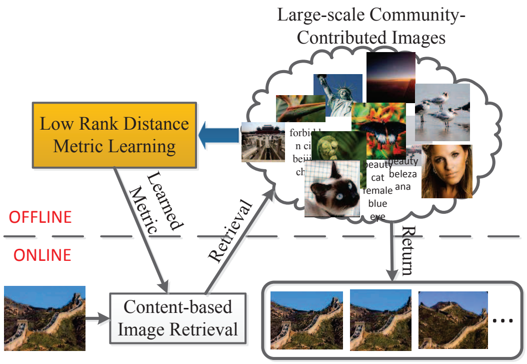
Large-scale Community-Contributed Images
Low Rank Distance Metric Learning
Learned Metric
Retrieval
OFFLINE
ONLINE
Return
Content-based Image Retrieval
...

Figure 1: The framework for image retrieval based on the learned metric. The distance metric is learned from social data with rich context and then applied to content-based image retrieval.

1. INTRODUCTION

With the advance of digital cameras and high quality mobile devices as well as internet technologies, there are increasingly large amounts of images available on the web such as Flickr, Zooomr and Picasa, which necessitates effective and efficient image retrieval techniques [14]. As one paradigm of image retrieval, Content-Based Image Retrieval (CBIR) typically ranks images according to the visual similarities measured with Euclidean distance [13]. However, its retrieval performance is often unsatisfied due to the well-known semantic gap between visual representation and semantic meaning [11, 9, 7, 16]. Fortunately, lots of image sharing sites provide us with plentiful community contributed resources, in particular, images and their associated tags, from which raw correspondences between images and tags are available but possibly noisy. Consequently, how to learn an appropriate distance metric by leveraging the available contextual information is essential to alleviate the semantic gap in CBIR, which is also our focus in this paper.

So far, distance metric learning has been extensively studied in machine learning and data mining work [5, 1, 4, 15], which usually exploits some side information given in the form of either class labels, or pairwise constraints indicating whether two simples are similar (must-link) or dissimilar (cannot-link). The side information is usually collected from users as a kind of exact knowledge. This makes it difficult to be applied in the case of web application provided with a large-scale but noisy collection. Some methods focus on

Area Chair: Zhen Wen

feature selection to measure sample similarity [6, 8]. There are some work [10, 15] devoted to learn distance metric by leveraging community contributed resources. As one of most related to our work, multi-label distance metric learning [10] is proposed to learn a metric from social media data. It considers the user tags and visual content by two linear transformation matrices, which transform the visual features and text features into two latent spaces, respectively. These two latent spaces are assumed to have some common structures. Due to the different characteristics of the both features, such assumption may be ideal and strict too much.

Based on the above considerations, in this paper, we propose a robust low-rank metric learning algorithm by leveraging social media, and then apply it to image retrieval, as shown in Fig. 1. We investigate the problem from the following aspects. First, a more reasonable assumption about the co-constraints from the visual and tagging information are presented. Specifically, we assume that the image similarities with the learned metric should be consistent with the semantic similarities according to the raw tagging information of images. Second, considering the inevitable noise in tagging information, the learned metric is constrained not to deviate from the visual structure, which is regularized with the preservation of typical Euclidean distance. Third, as indicated in the paper title, the learned metric should be (approximately) low rank to better reflect the intrinsic structure of data. Finally, the proposed problem is formulated as a convex optimization problem with nuclear norm minimization and is solved based on proximal gradient method.

2. LOW RANK METRIC LEARNING

In this section, we first define some notions and problem setting in Section 2.1. Then we elaborate our formulation in Section 2.2 and finally the proposed optimization algorithm is detailed in Section 2.3.

2.1 Problem Setup

Throughout this paper, we use bold uppercase characters to denote matrices, bold lowercase characters to denote vectors. For any matrix $\mathbf{A}$, $\mathbf{a}_i$ means the i-th column vector of $\mathbf{A}$, $\|\mathbf{A}\|_F$ denotes the Frobenius norm of $\mathbf{A}$ and $\mathrm{Tr}[\mathbf{A}]$ is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. $\mathbf{A}^T$ denotes the transposed matrix of $\mathbf{A}$.

In metric learning problems, we are often given a set of n data points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathcal{R}^l$ is the l dimensional visual feature vector. In our problem, there exist textual representations of samples $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$. Here $\mathbf{y}_i \in \mathcal{R}^m$ is the m dimensional textual feature vector, which is binary or calculated by TF-IDF model. The goal is to compute the distance function $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)$:

$$d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \quad (1)$$

Here $\mathbf{M} \in \mathbf{R}^{l \times l}$ is the Mahalabobis metric, a symmetric matrix. To satisfy the properties of metric, i.e., non-negativity and triangle inequality, $\mathbf{M}$ must be positive semi-definite (p.s.d.), that is $\mathbf{M} \succeq 0$. Our goal aims at learning an optimal $\mathbf{M}$ by leveraging the knowledge of the visual and textual spaces. Note that when $\mathbf{M}$ is equal to the identity matrix $\mathbf{I}$, the distance in Eq. 1 reduces to the Euclidean distance.

2.2 Formulation

We now elaborate the formulation of the proposed low rank metric learning method. The main idea is that the visual similarity computed by the learned metric is expected to be consistent with the semantic similarity in the textual domain and should not deviate from the original similarity in the Euclidean space. To this end, we formulate our metric learning problem into the following optimization framework.

$$\mathbf{M} = \arg\min_{\mathbf{M}} F(\mathbf{M}) + R(\mathbf{M}) \quad \text{s.t.} \quad \mathbf{M} \succeq 0 \qquad (2)$$

$F(\mathbf{M})$ is the objective function defined over the given data to preserve the similarity in the textual space and the visual space simultaneously. $R(\mathbf{M})$ is the regularization term for the low rank constraint. The learned metric can not only introduce the semantic information but also prevent overfitting the noisy tags in the real-world problem. In practice, each valid metric $\mathbf{M}$ can be decomposed as $\mathbf{M} = \mathbf{W}\mathbf{W}^T$, where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_r] \in \mathbf{R}^{l \times r}$. Hence, $\mathbf{W}$ can be interpreted as a linear mapping function: $\mathbf{R}^l \mapsto \mathbf{R}^r$.

First, to constrain the consistency of the pairwise similarity in the mapping space and the textual space, we encourage the pairwise similarities of images to be similar across these two spaces. For this, we propose the following cost function as a measure of disagreement between the structures with pairwise similarities.

$$\min_{\mathbf{M}} \ell(\mathbf{K}_{\mathbf{M}}, \mathbf{K}_{\mathbf{Y}}) \qquad (3)$$

Here $\ell(\cdot, \cdot)$ is a loss function. $\mathbf{K}$ is the similarity matrix for the corresponding space. We choose linear kernel, i.e., $k_{\mathbf{Y}}(\mathbf{y}_i, \mathbf{y}_j) = \frac{\mathbf{y}_i^T \mathbf{y}_j}{\|\mathbf{y}_i\|_2 \|\mathbf{y}_j\|_2}$ as the similarity measure in the textual space. In the mapping space, the Gaussian kernel is adopted to measure the similarities, i.e., $k_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = e^{-d^2_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)}$, where $d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}\mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j)}$ $= \|\mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|$. By minimizing the above disagreement, we try to keep the semantic similarity under the learned metric. The pairwise similarities across two spaces are guaranteed to be consistent.

However, tags of social images are created by users, which leads to that the tags are noisy, subjective and irrelevant. This may cause that the learned metric $\mathbf{M}$ is inaccurate and overfits noisy semantic similarity. To address this problem, we employ the visual content of images to prevent overfitting the noisy tags and enhance generalization and robustness of the learned metric. We expect that the learned metric enables to preserve the original visual similarity. Specifically, we aim to regularize $\mathbf{M}$ as close as possible to the identity matrix $\mathbf{I}$. In this work, we adopt Bregman divergence [4] to measure the closeness between $\mathbf{M}$ and $\mathbf{I}$ as

$$D(\mathbf{M}||\mathbf{I}) = g(\mathbf{M}) - g(\mathbf{I}) - \langle \nabla_g(\mathbf{I}), \mathbf{M} - \mathbf{I} \rangle, \qquad (4)$$

where $g(\cdot)$ is a strict convex and continuously differentiable function. In this work, we use the $\log det$ function to define $g(\cdot)$, i.e., $g(\mathbf{M}) = -\log det(\mathbf{M})$. Consequently, we have

$$D(\mathbf{M}||\mathbf{I}) = \mathrm{Tr}(\mathbf{M}) - \log det(\mathbf{M}) - n. \qquad (5)$$

Combining Eq. 3 and Eq. 5, we obtain:

$$F(\mathbf{M}) = \frac{\gamma}{2} \ell(\mathbf{K}_{\mathbf{M}}, \mathbf{K}_{\mathbf{Y}}) + \lambda D(\mathbf{M}||\mathbf{I}). \qquad (6)$$

Here γ and λ are two trade-off parameters. The above function is convex, sine the first term and the Bregman divergence is convex obviously.

For the regularization term $R(\mathbf{M})$, we constrain it to be low rank and it is intuitive to minimize its rank. However, $\mathrm{rank}(\mathbf{M})$ is a non-convex function with respect to $\mathbf{M}$

and hard to optimize due to the combinational nature. To address this problem, we replace rank($\mathbf{M}$) with its nuclear norm, which is a surrogate of matrix rank and convex [2]. The nuclear norm of $\mathbf{M}$ is defined as the sum of its singular values, i.e., $\|\mathbf{M}\|_* = \sum_{i=1}^{r} \sigma_i(\mathbf{M})$, where σ_i is the i-th singular value of $\mathbf{M}$ and r is the rank of $\mathbf{M}$. Hence, we have

$$R(\mathbf{M}) = \|\mathbf{M}\|_*. \tag{7}$$

Substituting $F(\mathbf{M})$ and $R(\mathbf{M})$ by the above definitions, our objective function in Eq. 2 can be rewritten as:

$$\min_{\mathbf{M}\succeq 0} \frac{\gamma}{2}\ell(\mathbf{K_M}, \mathbf{K_Y}) + \lambda(\mathrm{Tr}(\mathbf{M}) - \log det(\mathbf{M})) + \|\mathbf{M}\|_* \tag{8}$$

Given a convex loss function, we can see that this objective function is convex with respect to $\mathbf{M}$ and has a clear closed form. We will detail its optimization in the next subsection.

2.3 Optimization Algorithm

To solve the optimization problem (8), we first decide the function $\ell(\cdot, \cdot)$. In this work, we use the function $\ell(x, y) = (x - y)^2$ to measure the disagreement. To optimize the above objective function with the nuclear norm regularizer, we utilize the proximal gradient method [12]. Instead of directly minimizing our objective function, proximal gradient algorithms minimize a sequence of separable quadratic approximations to it by Taylor expansion at current value of $\mathbf{M} = \mathbf{M}_\tau$ and Lipschitz coefficient α, denoted as $Q(\mathbf{M}, \mathbf{M}_\tau)$.

$$\begin{aligned} Q(\mathbf{M}, \mathbf{M}_\tau) &= F(\mathbf{M}_\tau) + \langle \nabla F(\mathbf{M}_\tau), \mathbf{M} - \mathbf{M}_\tau \rangle \\ &\quad + \frac{\alpha}{2}\|\mathbf{M} - \mathbf{M}_\tau\|_F^2 + \|\mathbf{M}\|_* \end{aligned} \tag{9}$$

Here $\nabla F(\mathbf{M}_\tau)$ is the gradient computed as follows.

$$\begin{aligned} \nabla F(\mathbf{M}_\tau) = \gamma \sum_{i,j=1}^{n} &[(k_\mathbf{M}(\mathbf{x}_i, \mathbf{x}_j) - k_\mathbf{Y}(\mathbf{y}_i, \mathbf{y}_j))k_\mathbf{M}(\mathbf{x}_i, \mathbf{x}_j) \\ &\times (-(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j))] + \lambda(\mathbf{I} - \mathbf{M}^{-1}) \end{aligned} \tag{10}$$

In the above derivation, we use $\frac{\partial \log det(\mathbf{M})}{\partial \mathbf{M}} = \mathbf{M}^{-1}$ and $\frac{\partial \mathrm{Tr}[\mathbf{M}]}{\partial \mathbf{M}} = \mathbf{I}$. Let $\mathbf{G}_\tau = \mathbf{M}_\tau - \alpha^{-1}\nabla F(\mathbf{M}_\tau)$. We obtain

$$\begin{aligned} Q(\mathbf{M}, \mathbf{M}_\tau) &= \frac{\alpha}{2}\|\mathbf{M} - \mathbf{G}_\tau\|_F^2 + \|\mathbf{M}\|_* \\ &\quad + F(\mathbf{M}_\tau) - \frac{1}{2\alpha}\|\nabla F(\mathbf{M}_\tau)\|_F^2, \end{aligned} \tag{11}$$

$$\mathbf{M} = \arg\min_{\mathbf{M}} Q(\mathbf{M},\mathbf{M}_\tau) = \arg\min_{\mathbf{M}} \frac{\alpha}{2}\|\mathbf{M} - \mathbf{G}_\tau\|_F^2 + \|\mathbf{M}\|_*. \tag{12}$$

$\mathbf{M}$ can be updated by minimizing $Q(\mathbf{M}, \mathbf{M}_\tau)$ with fixed $\mathbf{M}_\tau$ iteratively. We can see that $\mathbf{M}$ can be learned by a fixed-pointed iterative method involving two alternating steps:

(1) (gradient step) $\mathbf{G}_\tau = \mathbf{M}_\tau - \alpha^{-1}\nabla F(\mathbf{M}_\tau)$,

(2) (shrinkage step) $\mathbf{M}_{\tau+1} = S_{\frac{\mu}{\alpha}}(\mathbf{G}_\tau)$

In the shrinkage step, $S_{\frac{\mu}{\alpha}}(\mathbf{G}_\tau)$ is a matrix shrinkage operator on $\mathbf{G}_\tau$. This can be solved by singular value thresholding since $\mathbf{G}_\tau$ is a symmetric and p.s.d. matrix. Letting $\mathbf{G}_\tau = \mathbf{U\Lambda U}^T$ be the eigenvalue decomposition of $\mathbf{G}_\tau$, $S_{\frac{\mu}{\alpha}}(\mathbf{G}_\tau) = \mathbf{U}\max\{\mathbf{\Lambda} - \frac{\mu}{\alpha}, 0\}\mathbf{U}^T$, where max is element-wise. This step truncates any eigenvalue less than $\frac{\mu}{\alpha}$ to 0, which reduces the nuclear norm as well. We summarize the proximal gradient method based method in Algorithm 1.

Algorithm 1 Proximal Gradient Method for LRML

```
Input:
    Visual Representation $\mathbf{X}$ and Text Representation $\mathbf{Y}$;
    Parameters $\gamma$ and $\lambda$.
$\tau = 1$; Initialize $\mathbf{M}_\tau$ as an identity matrix;
repeat
    Initialize $\alpha = \alpha_0$;
    repeat
        Set $\mathbf{G}_\tau = \mathbf{M}_\tau - \alpha^{-1}\nabla F(\mathbf{M}_\tau)$;
        Decomposition $\mathbf{G}_\tau = \mathbf{U\Lambda U}^T$
        Update $\mathbf{M}_{\tau+1} = \mathbf{U}\max\{\mathbf{\Lambda} - \frac{\mu}{\alpha}, 0\}\mathbf{U}^T$;
        Update $\alpha = \eta\alpha$
    until $F(\mathbf{M}_{\tau+1}) + \|\mathbf{M}_{\tau+1}\|_* \leq Q(\mathbf{M}_{\tau+1}, \mathbf{M}_\tau)$
    $\tau = \tau + 1$;
until Convergence criterion satisfied
Output:
    Metric $\mathbf{M} = \mathbf{M}_\tau$
```

3. EXPERIMENTS

In this section, we evaluate LRML on the NUS-WIDE-Lite dataset [3], which is a challenging collection of real-word web images from Flickr. These social images contain rich information, including user tags and other metadata. This dataset contains 55,615 images with 5,018 unique tags. The ground-truth annotations over 81 concepts have been provided. For visual feature representation, we use features provided by the dataset: 64-D color histogram (LAB), 144-D color auto-correlation (HSV), 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments (LAB). We sequentially combine these 5 groups into 634-D features. For the text domain, the textual feature vectors are represented by binary vectors. For performance evaluation, this set is randomly divided into three parts: 5,000 images for learning $\mathbf{M}$, $2,000$ images as query images and the remaining images as retrieval database.

We apply the learned metric to image retrieval and adopt Normalized Discounted Cumulative Gain at top k (NDCG@k), a ranking-based evaluation measure, to evaluate the retrieval performance. It measures the different levels of relevance and prefers the retrieved ranking results that follow the actual relevance order. Due to the space limit, we omit the definition of NDCG@k (please refer to [10]). To evaluate the entire ranking list, we use Average Precision (AP), a good combination of precision and recall, based on the groundtruths over 81 concepts. If the returned image has a common concept with the query image, we treat it relevant. Mean Average Precision (MAP) is obtained by averaging the APs on all test images.

The two parameters γ and β are important, which trade off the importance of visual information and tag information. We tune them in the range $\{0.001, 0.01, 0.1, 1, 10\}$ by cross validation. For MLML, we also tune its trade-off parameter in the same range as ours by cross validation.

Next, we compare the proposed method with the state-of-the-art algorithms: 1) Euclidean (Eu), 2) Relevant Component Analysis (RCA) [1], 3) Neighborhood Components Analysis (NCA) [5] and 4) Multi-Label Metric Learning (MLML) [10]. The compared experimental results are shown in Table 1 and Table 2 in terms of NDCG@k and MAP, respectively. From the results, first we can see that all the distance metric learning approaches significantly outperforms

Table 1: NDCG@k of our proposed LRML and the compared algorithms.

NDCG@k	Eu	RCA	NCA	MLML	LRML
5	0.4240	0.4241	0.4356	0.4506	**0.4549**
10	0.4451	0.4461	0.4502	0.4661	**0.4832**
50	0.4689	0.4701	0.4721	0.4821	**0.4966**
100	0.4829	0.4843	0.4842	0.5213	**0.5488**
500	0.5359	0.5369	0.5350	0.5473	**0.5793**
1000	0.5647	0.5674	0.5713	0.5938	**0.6199**

Table 2: MAP of our proposed LRML and the compared algorithms.

	Eu	RCA	NCA	MLML	LRML
MAP	0.6791	0.6798	0.6812	0.7272	**0.7508**

the Euclidean distance. It does not utilize any tagging information. This shows that the distance metric learning using tags is beneficial and important for image retrieval. Second, the proposed metric learning algorithm achieves the overall best performance among other metric learning methods. This demonstrates the advantages of our method to learn a low rank distance metric leveraging the visual and textual information. Finally, compared with MLML, our method gains better results, which reveals that the pairwise similarity preservation across domains and low rank constraint are suitable to learn a semantic metric.

Finally, the qualitative image retrieval performance achieved by different methods are evaluated by randomly choosing several test images. Figure 2 illustrates top 5 relevant images returned by different distance metrics. The irrelevant images are marked with red boundary. We can observe that our metric learning method often achieves better quality.

4. CONCLUSION

In this work, we study the metric learning problem to boost the performance of CBIR, and propose a low rank distance metric for social image retrieval by exploiting knowledge from community contributed images associated with tags. The learned metric can preserve the sematic similarity in textual space and the visual similarity in visual space, which can enable to learn a robust distance metric. The proposed problem is formulated as a convex optimization with the nuclear norm regularization and then an effective optimization method is provided. Finally, We apply the learned metric to image retrieval and conduct extensive experiments, which show that our method is effective and promising.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China 973 Program (Project No. 2012CB316304) and the National Natural Science Foundation of China (Grant No. 60903146, 90920303).

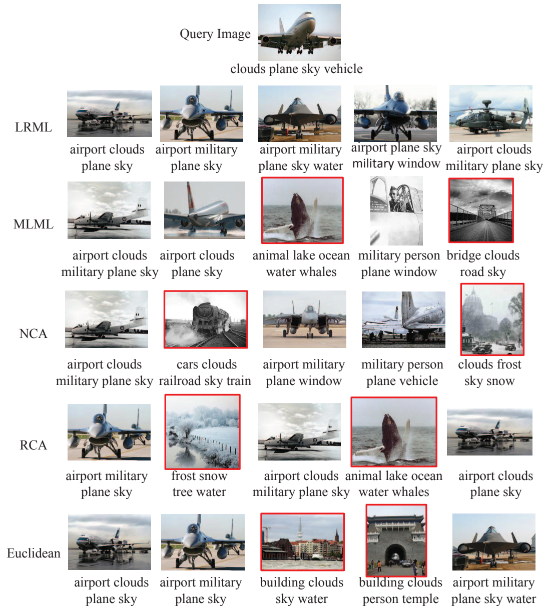
Query Image
clouds plane sky vehicle
LRML
airport clouds plane sky
airport military plane sky
airport military plane sky water
airport plane sky military window
airport clouds military plane sky
MLML
airport clouds military plane sky
airport clouds plane sky
animal lake ocean water whales
military person plane window
bridge clouds road sky
NCA
airport clouds military plane sky
cars clouds railroad sky train
airport military plane window
military person plane vehicle
clouds frost sky snow
RCA
airport military plane sky
frost snow tree water
airport clouds military plane sky
animal lake ocean water whales
airport clouds plane sky
Euclidean
airport clouds plane sky
airport military plane sky
building clouds sky water
building clouds person temple
airport military plane sky water

Figure 2: Examples showing the retrieved results by 5 different methods. For each method, top 5 ranked images are presented. Red rectangle denotes irrelevant images to the queries. Best viewed in color.

6. REFERENCES

[1] A. Bar-Hillel and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

[2] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *ACM CIVR*, 2009.

[4] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[5] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2005.

[6] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, 2009.

[7] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu. Image annotation using multi-correlation probabilistic matrix factorization. In *ACM MM*, 2010.

[8] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, 2012.

[9] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *PR*, 42(2):218–228, 2009.

[10] G.-J. Qi, X.-S. Hua, and H.-J. Zhang. Learning semantic distance from community-tagged media collection. In *ACM MM*, 2009.

[11] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22(12):1349ÍC1380, 2000.

[12] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 2010.

[13] M. Wang, X.-S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. MM*, 11(3):465–476, 2009.

[14] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang. Towards a relevant and diverse search of social images. *IEEE Trans. MM*, 12(8):829–842, 2010.

[15] P. Wu, S. C. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. In *WSDM*, 2011.

[16] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans. PAMI*, 34(4):723–742, 2012.