Image Annotation Using Multi-Correlation Probabilistic Matrix Factorization

Zechao Li, Jing Liu, Xiaobin Zhu, Tinglin Liu, Hanqing Lu
Institute of Automation, Chinese Academy of Sciences
Beijing 100086, China
+86-10-62542971
{zcli, jliu, xbzhu, tlliu, luhq}@nlpr.ia.ac.cn

ABSTRACT

The image-word correlation estimation is an essential issue in image annotation. In this paper, we propose a multi-correlation probabilistic matrix factorization (MPMF) algorithm for the correlation estimation. Different from the traditional solutions which treat the image-word correlation, image similarity and word relation independently or sequentially, in the proposed MPMF, these three elements are integrated together simultaneously and seamlessly. Specifically, we have derived two low-dimensional sets by conducting a joint factorization upon the word-to-image relation matrix, the image similarity matrix, and the word relation matrix to derive two low-dimensional sets of latent word factors and latent image factors. Finally, the annotation words of each untagged or noisily tagged image can be predicted by reconstructing the image-word correlations with the both derived latent factors. Experimental results on the Corel dataset and a Flickr image dataset show the superior performance of our proposed algorithm over the state-of-the-arts.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Algorithms, Theory, Experimentation

Keywords

Matrix factorization, Image annotation, Image similarity, Word correlation

1. INTRODUCTION

Image Annotation has attracted extensive researchers owing to its great potentials in image retrieval. The goal of image annotation is to find suitable annotation words to represent the visual content of an untagged or noisily tagged image. In other words, the correlation between images and annotation words is a central problem in view of technical solutions.

To effectively annotate an image, some efforts are conducted on exploring possibly available tagging information (i.e., partial image-word correlation, IWR), image-to-image relation (IIR), and word-to-word relation (WWR), or partial mentioned items to estimate the correlation between the image and each annotation word. Some probabilistic modeling methods, such as CMRM [3], CRM [7] and MBRM [2], attempted to model such correlation over all tagged images, and they have different representations of IWR and IIR. In order to maintain the semantic consistence of annotation words for an image, there are some efforts considering WWR in the annotation process, such as Coherent Language Model [4], Correlated Label Propagation [6] and WordNet-based method [5]. To jointly consider the three types of relations, some efforts [9, 11, 8] adopted a relaxing solution of image annotation by considering image correlations with the assumption of the word independence, and vice versa. Liu et al. [8] explicitly demonstrated such an idea by proposing a graph-learning framework for image annotation, in which two sequential steps of learning processes are conducted, namely image-based graph learning for "basic image annotation" and word-based graph learning for "annotation refinement". Since the sequential learning process as a relaxing solution cannot explore the three relations in a simultaneous and seamless manner, it may be not suitable.

To address the problem, in this paper, we propose a novel image annotation algorithm using Multi-correlation Probabilistic Matrix Factorization (MPMF), in which the information of IWR, IIR and WWR are integrated simultaneously. The basic idea is illustrated as in Fig. 1. Recently,

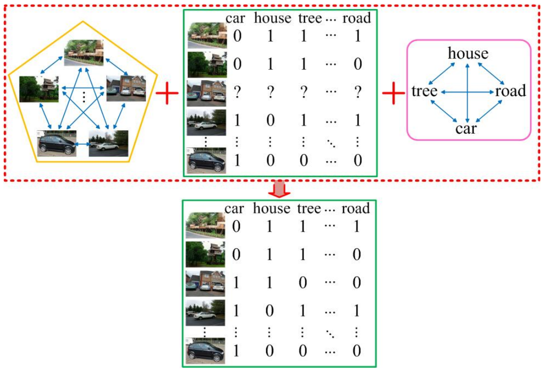


Figure 1: Illustration of our method.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'10, October 25–29, 2010, Firenze, Italy.

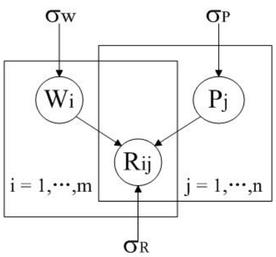


Figure 2: Graphical Model for PMF.

the Probabilistic Matrix Factorization (PMF) algorithm [10] has been applied in recommendation and classification to estimate the relations among items. We extend the typical PMF model by fusing different sources of correlations, which is called MPMF. By conducting latent factor analysis using probabilistic matrix factorization, we learn the low-rank latent feature spaces by employing an IWR matrix, a WWR matrix and an IIR matrix. Different from traditional factor analysis, we connect these three different data resources through the shared word latent feature space and image latent feature space respectively. That is, the word latent factors in WIR space is tied to the ones in WWR space, and the image latent factors in WIR space is tied to the ones in IIR space. Finally, the discovered bases, i.e., the image latent factors and the word latent factors, are used to reconstruct the correlations of images and words. The experimental results on the Corel dataset and a web dataset (crawled from Flickr.com) demonstrate that our approach performs better than the state-of-the-art algorithms.

The rest of paper is organized as follows. Section 2 provides an introduction of typical PMF algorithm. Section 3 presents the improved MPMF for image annotation. The experimental results are reported and discussed in Section 4. We conclude the paper in Section 5.

2. OVERVIEW OF PMF

In this section, we will introduce the framework of probabilistic matrix factorization, which performs well on the large, sparse and very imbalanced datasets and is typically applied in the recommender system. For clarity, we employ the assumption that we have m users, n items, and rating values within the range $[0, 1]$. Let r_{ij} represent the relation between user i and item j. The idea of matrix factorization is to derive a high-quality l-dimensional feature representation W of users and P of items. Let $W \in \mathbb{R}^{l\times m}$ and $P \in \mathbb{R}^{l\times n}$ be the latent user and item feature matrices, with column vectors W_i and P_j representing user-specific and item-specific latent feature vectors, respectively. A probabilistic model with Gaussian observation noise (see Fig. 2) and define the conditional distribution as

$$p(R|W, P, \sigma_R^2) = \prod_{i=1}^m \prod_{j=1}^n \mathcal{N}[(r_{ij}|g(W_i^T P_j), \sigma_R^2)]^{I_{ij}^R} \quad (1)$$

where $\mathcal{N}(x|\mu, \sigma^2)$ denotes the probabilistic density function, in which the conditional distribution is defined as the Gaussian distribution with mean μ and variance σ^2, and I_{ij}^R is the indicator function that is equal to 1 if the user i rated item j and equal to 0, otherwise. The function $g(x)$ is the logistic function $g(x) = 1/(1 + exp(-x))$, which makes it possible to bound the range of $W_i^T P_j$ within the range $[0, 1]$. We place zero-mean spherical Gaussian priors on user and item feature vectors:

$$\begin{aligned} p(W|\sigma_W^2) &= \textstyle\prod_{i=1}^m \mathcal{N}(W_i|0, \sigma_W^2 \mathbf{I}) \\ p(P|\sigma_P^2) &= \textstyle\prod_{j=1}^m \mathcal{N}(P_j|0, \sigma_P^2 \mathbf{I}) \end{aligned} \quad (2)$$

The log of the posterior distribution over the user and item features is given by

$$\begin{aligned} &\ln p(W, P|R, \sigma_R^2, \sigma_W^2, \sigma_P^2) = \\ &- \tfrac{1}{2\sigma_R^2} \textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R (r_{ij} - g(W_i^T P_j))^2 \\ &- \tfrac{1}{2\sigma_W^2} \textstyle\sum_{i=1}^m W_i^T W_i - \tfrac{1}{2\sigma_P^2} \sum_{j=1}^n P_j^T P_j \\ &- \tfrac{1}{2}(\textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R \ln \sigma_R^2 + ml \ln \sigma_W^2 + nl \ln \sigma_P^2) + \mathcal{C} \end{aligned} \quad (3)$$

where $\mathcal{C}$ is a constant independent on the parameters. Actually, maximizing the log-posterior distribution with fixed hyperparameters (i.e. the observation noise variance and prior variances) is equivalent to minimizing the sum-of-squared-errors objective function with quadratic regularization terms:

$$\begin{aligned} E =& \tfrac{1}{2} \textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R (r_{ij} - g(W_i^T P_j))^2 \\ &+ \tfrac{\lambda_W}{2} \|W\|_F^2 + \tfrac{\lambda_P}{2} \|P\|_F^2 \end{aligned} \quad (4)$$

where $\lambda_W = \sigma_R^2/\sigma_W^2$, $\lambda_P = \sigma_R^2/\sigma_P^2$, and $\|\cdot\|_F$ denotes the Frobenius norm. A local minimum of the objective function given by Eq. 4 can be found by performing gradient descent in W and P.

3. MPMF FOR IMAGE ANNOTATION

3.1 Motivations

Image annotation can be understood as a learning process, in which the unknown relations between test images and annotated words are estimated by exploring available resources. Thus, how to estimate and integrate these relations is a key issue. In this section, we will address the issue by proposing an extended PMF algorithm for image annotation.

In the problem of image annotation, there are two media types: image and word. We can have three kinds of relations: word-word relation, word-image relation and image-image relation. The word-image relation in the problem of image annotation can be analogous to user-item relation in recommender system. Furthermore, the available relation of annotated words and images is usually very sparse and imbalanced. Due to the scarce of high-quality image tagged dataset, the probabilistic matrix factorization algorithm as a natural and feasible option is employed to conduct our work. However, the standard probabilistic matrix factorization model can only employ one relation. Then we extend the model to integrate the relation between words and annotated images, word correlations and image similarities, named as Multi-Correlation Probabilistic Matrix Factorization (MPMF). We employ the factor analysis to factorize word-word relation matrix, word-image relation matrix and image-image relation matrix, respectively, and then connect these three different data resources through the shared word latent feature space, that is, the word latent feature space in the word-image relation matrix is the same as in the word co-occurrence space, and the shared image latent feature space, that is, the image latent feature space in the word-image relation matrix is the same as in the image similarity space.

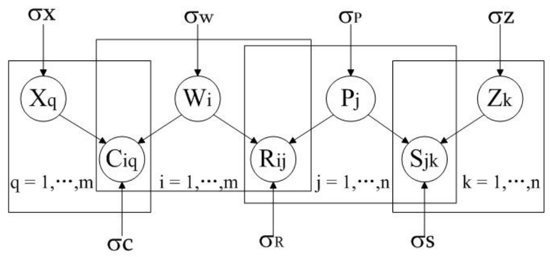


Figure 3: Graphical Model for MPMF

3.2 Image Annotation via MPMF

In this section, will introduce the proposed MPMF model and employ it in task of image annotation. The graphical model for MPMF is illustrated as in Fig. 3.

Suppose we have m annotated words and n images. Let r_{ij} represent the relation of word i and image j within the range $[0, 1]$, $c_{iq} \in [0, 1]$ denote the weight between word i and word q, and $s_{jk} \in [0, 1]$ denote the similarity between image j and image k. Let $W \in \mathbb{R}^{l\times m}$, $P \in \mathbb{R}^{l\times n}$, $X \in \mathbb{R}^{l\times m}$ and $Z \in \mathbb{R}^{l\times n}$ be latent word, image, word factor and image factor feature matrices, with column vectors W_i, P_j, X_q and Z_k representing word-specific, image-specific, word factor-specific and image factor-specific latent feature vectors, respectively. We place zero-mean spherical Gaussian priors on word, image, word factor and image factor feature vectors, similar to Eq. 2.

We can employ probabilistic factor analysis to factorize the matrices R, C and S, respectively. Thus, we can obtain three log of the posterior distribution which are similar to Eq. 3. As described in Fig. 3, we fuse the word-image matrix, word-word matrix and image-image matrix into a consistent and compact feature representation. Based on Fig. 3, the log of the posterior distribution is given by

$$\begin{aligned}
&\ln p(W, P, X, Z|R, C, S, \sigma_R^2, \sigma_C^2, \sigma_S^2, \sigma_W^2, \sigma_P^2, \sigma_X^2, \sigma_Z^2) = \\
&- \tfrac{1}{2\sigma_R^2} \textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R (r_{ij} - g(W_i^T P_j))^2 \\
&- \tfrac{1}{2\sigma_C^2} \textstyle\sum_{i=1}^m \sum_{q=1}^m \mathbf{I}_{iq}^C (c_{iq} - g(W_i^T X_q))^2 \\
&- \tfrac{1}{2\sigma_S^2} \textstyle\sum_{j=1}^n \sum_{k=1}^n \mathbf{I}_{iq}^C (s_{jk} - g(P_j^T Z_k))^2 \\
&- \tfrac{1}{2\sigma_W^2} \textstyle\sum_{i=1}^m W_i^T W_i - \tfrac{1}{2\sigma_P^2} \sum_{j=1}^n P_j^T P_j \\
&- \tfrac{1}{2\sigma_X^2} \textstyle\sum_{q=1}^m X_q^T X_q - \tfrac{1}{2\sigma_Z^2} \sum_{k=1}^n Z_k^T Z_k \\
&- \tfrac{1}{2}((\textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R) \ln \sigma_R^2 + (\sum_{i=1}^m \sum_{q=1}^m \mathbf{I}_{iq}^C) \ln \sigma_C^2) \\
&- \tfrac{1}{2} \textstyle\sum_{j=1}^n \sum_{k=1}^n \mathbf{I}_{jk}^S \ln \sigma_S^2 - \tfrac{1}{2} ml \ln \sigma_W^2 \\
&- \tfrac{1}{2}(nl \ln \sigma_P^2 + ml \ln \sigma_X^2 + nl \ln \sigma_Z^2) + \mathcal{C} \qquad (5)
\end{aligned}$$

where $\mathcal{C}$ is a constant independent on the parameters. Maximizing the log-posterior distribution with fixed hyperparameters is equivalent to minimizing the sum-of-squared-errors objective function with quadratic regularization terms:

$$\begin{aligned}
E = {} & \tfrac{1}{2} \textstyle\sum_{i=1}^m \sum_{j=1}^n \mathbf{I}_{ij}^R (r_{ij} - g(W_i^T P_j))^2 \\
& + \tfrac{\lambda_C}{2} \textstyle\sum_{i=1}^m \sum_{q=1}^m \mathbf{I}_{iq}^C (c_{iq} - g(W_i^T X_q))^2 \\
& + \tfrac{\lambda_S}{2} \textstyle\sum_{j=1}^n \sum_{k=1}^n \mathbf{I}_{jk}^S (s_{jk} - g(P_j^T Z_k))^2 \\
& + \tfrac{\lambda_W}{2} \|W\|_F^2 + \tfrac{\lambda_P}{2} \|P\|_F^2 + \tfrac{\lambda_X}{2} \|X\|_F^2 + \tfrac{\lambda_Z}{2} \|W\|_Z^2 \qquad (6)
\end{aligned}$$

where $\lambda_C = \sigma_R^2/\sigma_C^2$, $\lambda_S = \sigma_R^2/\sigma_S^2$, $\lambda_W = \sigma_R^2/\sigma_W^2$, $\lambda_P = \sigma_R^2/\sigma_P^2$, $\lambda_X = \sigma_R^2/\sigma_X^2$, and $\lambda_Z = \sigma_R^2/\sigma_Z^2$. A local minimum of the objective function given by Eq. 6 can be found by performing gradient descent in W, P, X and Z. In order to reduce the model complexity, in all of the experiments we conduct in Section 4, we set $\lambda_W = \lambda_P = \lambda_X = \lambda_Z$.

In our experiment, we use the graphic model in Fig. 3 for image annotation and annotation refinement. The word-image relation is binary, that is, r_{ij} equals to 1 if word i is an annotated word of image j, and 0 otherwise. The image similarity is calculated by GMM model and word correlation is calculated by the function $c_{iq} = N(i, q)/N(i)$, where $N(i)$ denotes the number of images whose annotated words contain word i. In this paper, the difference between image annotation and annotation refinement is the initiation for the relation between annotated words and the testing images. In image annotation, the initial matrix elements according to the relation between annotated words and the testing images are set to 0. In image annotation refinement, we employ the results of other annotation methods, such as CRM and MBRM, to initialize the relation between annotated words and the testing images.

4. EXPERIMENTAL ANALYSIS

We tested the algorithms using two different datasets, the Corel data set and our web dataset (crawled from Flickr) without any manual label information. The following experiments will demonstrate that in our approach the promising results can be achieved not only with the good training information but also under the circumstance that no training knowledge is available.

4.1 Results on the Corel Dataset

In this section, MPMF will be tested on the Corel dataset [1] for image annotation and annotation refinement. The Corel dataset is a basic comparative dataset for recent research works on image annotation.

To compare with previous works, the quality of AIA is measured by the process of retrieving test images with single keyword. For each keyword, precision and recall are calculated as in [3, 7]. Let A be the total number of images automatically annotated with a given word, B be the number of images correctly annotated and C be the number of correct images under ground-truth annotation. Then $recall = \frac{B}{C}$, and $precision = \frac{B}{A}$. Recall and precision values are averaged over the testing words. Table 1 shows the comparison results of AIA on the Corel dataset. We compared our method with various state-of-the-art algorithms including CMRM [3], CRM [7], MBRM [2], CLM [4], CLP [6], DCMRM [9] and MSC [11]. Results are reported for all (260) words in the testing set. To make comparisons with the methods in [3, 7, 2, 11], the results for the top 49 words are also reported. The annotation length for each testing image is set to be 5. From Table 1, we can draw the conclusion that our method outperforms the state-of-the-art algorithms. Because our method can explore the IWR, WWR and IIR simultaneously and seamlessly, it achieves better performance than DCMRM and MSC, gaining 6 and 8 percent in recall and precision respectively for all words compared with MSC. Compared with CLP, it gains 8%, 61.9% and 35% on Recall, Precision and the number of words with non-zero recall for all words. And the corresponding is 36.0%, 12.5% and 10.7% compared with MBRM. For the top 49 words, it gains 6.4% and 5.4% on Recall and Precision compared with MBRM. Thus, our method is preferable to annotate images when the

Table 1: Image annotation performance comparison on the Corel dataset

Models	CMRM [3]	CRM [7]	MBRM [2]	CLM [4]	CLP [6]	DCMRM [9]	MSC [11]	MPMF
`#words with recall` $\geqslant 0$	66	107	122	79	125	135	**136**	135
`Results on all 260 words`								
`Mean Per-word Recall`	0.09	0.19	0.25	0.10	0.21	0.28	0.32	**0.34**
`Mean Per-word Precision`	0.10	0.16	0.24	0.12	0.20	0.23	0.25	**0.27**
`Results on 49 best words`								
`Mean Per-word Recall`	0.48	0.70	0.78	-	-	-	0.82	**0.83**
`Mean Per-word Precision`	0.40	0.59	0.74	-	-	-	0.76	**0.78**

Table 2: Annotation refinement on the Corel dataset

Models	CRM+MPMF	MBRM+MPMF
`#words with recall`$\geqslant$`0`	136	138
`Results on all 260 words`		
`Mean Recall`	0.34	0.35
`Mean Precision`	0.27	0.28
`Results on 49 best words`		
`Mean Recall`	0.83	0.84
`Mean Precision`	0.79	0.81

Table 3: Mean Average P@m ($m = 5, 10, 15, 20, 50, 100$) on randomly selected 100 words of Web dataset

m	5	10	15	20	50	100
`MAP@`m	0.257	0.221	0.204	0.192	0.173	0.159

training data are available. We also did some experiments on the image annotation refinement on the Corel dataset. We utilized the results of CRM and MBRM to initialize the relation between words and testing images in our method. Table 2 presents the corresponding results. For all words and the top 49 words, we can see that the performances improve significantly on Recall, Precision and the number of words with non-zero recall respectively. We can see that MPMF is effective for annotation refinement.

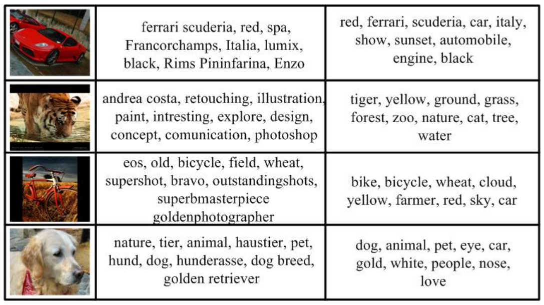
4.2 Results on the Flickr Dataset

In order to test the applicability of our method, another web datebase is built by us. 360 queries were submitted to Flickr searcher and 210 top-ranked images were crawled as well as their corresponding tags for each query. Note that the words occurred less than 10 times and the web pages containing no image were filtered out. Finally, we got a dataset of 74,763 images and 8,037 words totally, which show great diversity. Additionally, a 204-dimensional visual feature vector for each image is extracted, including 36-dimensional color histogram, 24-dimensional texture moment, and 144-dimensional color correlogram.

Generally, web images have extensive semantics and large variation on visual content, so the AIA for web images is a challenge work. Because the acquisition of the ground truth is too expensive, we evaluate the performance by the view of image retrieval. The annotation length is set to 10. We randomly select 100 query words of the web dataset and calculate the mean average $P@m$. Table 3 shows the performance of MPMF on different numbers of retrieved images and Fig. 4 presents some examples of the annotations generated by our method. Considering that all the information are got from web and no human work is provided, the performance is remarkable.

5. CONCLUSIONS

We proposed the MPMF model for image annotation, which integrates the image-word correlation, image similarity and word correlation simultaneously and seamlessly. Different from standard models, MPMF connects these three different data resources through the shared word latent feature space and the shared image latent feature space. The experiments on the Corel dataset and the Flickr image dataset demonstrate that the proposed approach is more preferable than the state-of-the-art algorithms.



	ferrari scuderia, red, spa, Francorchamps, Italia, lumix, black, Rims Pininfarina, Enzo	red, ferrari, scuderia, car, italy, show, sunset, automobile, engine, black
	andrea costa, retouching, illustration, paint, intresting, explore, design, concept, comunication, photoshop	tiger, yellow, ground, grass, forest, zoo, nature, cat, tree, water
	eos, old, bicycle, field, wheat, supershot, bravo, outstandingshots, superbmasterpiece goldenphotographer	bike, bicycle, wheat, cloud, yellow, farmer, red, sky, car
	nature, tier, animal, haustier, pet, hund, dog, hunderasse, dog breed, golden retriever	dog, animal, pet, eye, car, gold, white, people, nose, love

Figure 4: Annotation examples with top 10 returned words. Second column: top 10 annotations in Flickr; Third column: top 10 annotations by MPMF

6. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China (Grant No. 60903146 and 90920303), and 973 Program (Project No. 2010CB327905).

7. REFERENCES

[1] P. Dugulu and K. Barnard. Object recognitions as machine translation: learning a lexicon for a fixed image vocabulary. *ECCV*, 2002.

[2] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. *CVPR*, pages 1002–1009, 2004.

[3] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. *ACM SIGIR*, pages 119–126, 2003.

[4] R. Jin, J. Y. Chai, and L. Si. Effective automatic image annotation via a coherent language model and active learning. *ACM SIGMM*, pages 892–899, 2004.

[5] Y. Jin, L. Khan, L. Wang, and M. Awad. Image annotation by combining multiple evidence & wordnet. *ACM SIGMM*, pages 706–715, 2005.

[6] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. *CVPR*, pages 1719–1726, 2006.

[7] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *NIPS*, 2004.

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

[9] J. Liu, B. Wang, M. Li, M. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. *ACM SIGMM*, pages 605–614, 2007.

[10] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *NIPS*, 20:605–614, 2008.

[11] C. Wang, S. Yan, L. Zhang, and H. Zhang. Multi-label sparse coding for image annotation. *CVPR*, pages 1463–1650, 2009.