Image Tag Completion and Refinement by Subspace Clustering and Matrix Completion

Yuqing Hou [#†1], Zhouchen Lin [#*2]
[#] *Key Lab. of Machine Perception (MOE), School of EECS, Peking University, P. R. China*
[*] *Cooperative Medianet Innovation Center, Shanghai Jiaotong University, P. R. China*
[†] *HTC Research*
[1] yqhou@pku.edu.cn [2] zlin@pku.edu.cn

***Abstract*—Tag-based image retrieval (TBIR) has drawn much attention in recent years due to the explosive amount of digital images and crowdsourcing tags. However, the TBIR applications still suffer from the deficient and inaccurate tags provided by users. Inspired by the subspace clustering methods, we formulate the tag completion problem in a subspace clustering model which assumes that images are sampled from subspaces, and complete the tags using the state-of-the-art Low Rank Representation (LRR) method. And we propose a matrix completion algorithm to further refine the tags. Our empirical results on multiple benchmark datasets for image annotation show that the proposed algorithm outperforms state-of-the-art approaches when handling missing and noisy tags.**

***Index Terms*—Image Annotation, Subspace Clustering, Low Rank Representation, Matrix Completion, Tag Completion, Tag Refinement**

I. Introduction

The prevalence of social network and digital photography in recent years makes image retrieval an urgent need. Image retrieval methods can be classified into two categories: content-based image retrieval (CBIR) and tag-based image retrieval (TBIR). The performance of CBIR algorithms are limited due to the semantic gap between the low-level visual features used to represent images and the high-level semantic meaning behind images.

Tags can represent the semantics of image more precise than low-level visual features, giving rise to research on TBIR. Basing on text processing techniques, TBIR systems are usually more accurate and efficient in identifying relevant images and retrieving relevant images [1]. However, tags are usually noisy and incomplete due to the arbitrariness of user tagging behaviors, leading to performance degradations of TBIR systems. What's more, manual annotation is laborious, error prone, and subjective, making automatic image annotation an attractive research field.

Many machine learning methods have been developed for image annotation. They can be roughly grouped into three categories: discriminative methods, generative methods and search-based methods.

Discriminative methods use the classified images and tags to train a dictionary of concept models and formulate image annotation as a supervised learning problem. They annotate images using the likelihood between images and tags.

Generative methods learn the joint probability of image regions and words. Images are represented by properties of each of their segments, or blobs. Once all the images are segmented, quantization can be used to obtain a finite vocabulary of blobs. Thus, the images are treated as bags of words and blobs, each of which are assumed to be generated by hidden variables. Once the joint word-blob probabilities are learned, the annotation problem for a query image is formulated as a likelihood problem relating blobs to words. [2], [3] extend the LDA model and propose the Correlation LDA method. [4] introduces the cross-media relevance models (CMRM).

Search-based methods always search in the feature space to find the most relevant images to the query image, and transfer tags to it using various tag transfer algorithms [5], [6], [7]. JEC [6] demonstrates that simple baseline algorithm can achieve high performance. TagProp [5] applies metric learning in the neighborhood of the feature space to annotate query images.

Most annotation methods only complete tags or refine tags, or tackle both simultaneously. However, tag completion aims at adding missing tags and tag refinement focuses on removing noisy ones. To resolve the contradiction, we propose a Subspace Clustering and Matrix Completion (SCMC) method. SCMC performs tag completion and refinement sequentially. The refinement benefits from the completion.

We first perform tag completion for further refinement. We cluster images and share tags in each cluster. Here we model the tag completion task in a subspace clustering framework, which can model the distribution of the image features more precisely than classical clustering algorithms. Besides, subspace clustering algorithms do not need to measure similarity between different features. We assume that images are sampled from a union of multiple linear (or affine) subspaces and images, as well as their corresponding tags, should form a compatible (low rank) submatrix. Thus we can segment the subspaces and cluster images by the state-of-the-art method LRR [8]. Then we adopt a tag transfer algorithm [6] to complete tags in each cluster separately.

Most tag refinement methods are region-based, depending heavily on the image segmentation accuracy. To avoid the segmentation procedure, we introduce an inductive matrix completion (IMC) [9] model to refine the tags, making the

model robust and efficient. IMC has been successfully applied to predict gene-disease associations, which share the same nonlinear property with image-tag associations. The main contributions of our research are summarized as follows:

- We formulate the tag completion task as a subspace clustering framework, which is different from all the other subspace clustering based algorithms in the tag completion field.
- We refine the tag matrix using a matrix completion model to overcome the semantic gap as well as the sparsity of the tag matrix. Our approach is a novel application of the IMC method to the tag refinement field.

II. Annotation based on Subspace Clustering and Matrix Completion

The proposed annotation framework is illustrated in Fig. 1. We summarize the flowchart as follows:

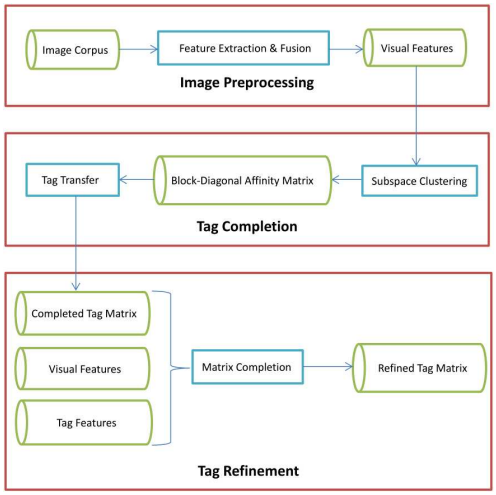
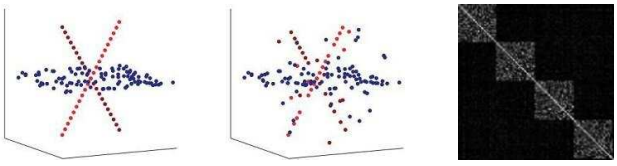


Fig. 1: The flowchart of the proposed SCMC



(a) (b) (c)

Fig. 2: Subspace clustering and the block-diagonal property of the affinity matrix: a mixture of subspaces consisting of a 2D plane and two 1D lines. (a) The samples are strictly drawn from the underlying subspaces. (b) The samples are approximately drawn from the underlying subspaces. (c) The block-diagonal property of the affinity matrix, each submatrix corresponds to a subspace [8].

A. Image Preprocessing

We adopt a subset of the image features exploited in [5]. Namely, 1 GIST descriptor and 8 bag-of-features (2 features types $\times$ 2 descriptors $\times$ 2 layouts). These features include global descriptors such as GIST and local descriptors such as SIFT and robust HUE descriptor. We adopt PCA to perform dimensionality reduction separately for all features, which are then concatenated to form the unique visual feature vectors for corresponding images.

B. Tag Completion

Here we apply LRR to cluster the visual feature vectors into different subspaces. The algorithm outputs a block-diagonal affinity matrix, each submatrix of which corresponds to a subspace (cluster), as shown in Fig. 2c [8]. Then we can cluster images according to the affinity matrix and perform tag completion by transferring tags in each cluster separately. Here we use only visual features for clustering since the tags may be too noisy and too incomplete.

1) Subspace Clustering: In the image annotation field, one usually needs a parametric model to characterize the user-provided images. Researchers have exploited the RPCA model [1] to decompose the tag matrix into a low-rank refined tag matrix and a sparse error matrix. However, a given image dataset is seldom well described by a single subspace. It is more reasonable to assume that images belonging to different categories are approximately sampled from a mixture of several low-dimensional subspaces, as shown in Fig. 2b [8], and the membership of the data points to the subspaces might be unknown, leading to the challenging problem of subspace clustering. Here, the goal is to cluster data into k clusters with each cluster corresponding to a subspace. When the cluster number is one, the subspace clustering model reduce to the RPCA model. Note that by clustering tagged images into clusters, we perform classification simultaneously.

A number of approaches to subspace clustering have been proposed in the past two decades. One of the state-of-the-art method is the LRR model [8], which performs robust subspace clustering and error correction in an efficient and effective way. LRR seeks the lowest rank representation among all the candidates that can represent the data samples as linear combinations of the basis in a given dictionary [8].

We denote the set of image feature vectors as $X = [x_1, x_2, \ldots, x_n]$, drawn from a union of k subspaces $\{S_i\}_{i=1}^{k}$. Each column of X is a feature vector in R^D and can be represented by a linear combination of the basis in a "dictionary". The LRR model just uses the matrix X itself as the dictionary and takes error into consideration:

$$\min_{Z,E} \|Z\|_* + \mu\|E\|_{2,1},$$
$$s.t., X = XZ + E.$$

where $Z = [z_1, z_2, \ldots, z_n]$ is the coefficient matrix with each z_i being the representation of x_i and E is the sparse error matrix. LRR solves the problem by LADM [10] efficiently. We can define the affinity matrix of an undirected graph using the lowest-rank representation (denoted by Z^*). The data vectors correspond to the vertices and affinity between x_i and x_j is computed by $|[Z^*]_{ij}| + |[Z^*]_{ji}|$. LRR uses the spectral clustering algorithm Normalized Cuts to perform the final segmentation. Fig. 2c demonstrates the block-diagonal property

of the affinity matrix, where each submatrix corresponds to a subspace [8]. Images belonging to the same subspace are clustered together.

2) Tag Transfer: In this work, we just construct tag matrices for each cluster and improve the simple and intuitive algorithm proposed in [6] to transfer tags in each cluster separately. For each cluster, we rank all the tags in the cluster, taking tag frequency, tag co-occurrence and local frequency into consideration. Then we can transfer the highest ranking tags to each image depending on the original tags. The tag matrix after tag sharing is no longer a $0/1$ valued matrix and the values (between 0 and 1) can represent their confidence level.

C. Tag Refinement

Tag refinement aims to correct noisy tags. The problem can be regarded as designing a *recommender system* where the goal is to predict the 'preference' that a user (image) would give to an item (tag). An important formulation used in recommender systems is matrix completion, where the problem is to 'delete' the noisy ones in the user-item preference matrix and 'complete' the missing ones given a sample of observed preferences. However, the tag matrix may be so sparse that some columns have at most one known entries and some rows have no known entries. The extreme sparsity makes traditional matrix completion methods [1] not applicable. Since we have performed tag completion to make the tag matrix much more complete to overcome the extreme sparsity, we can overcome the difficulty and employ the matrix completion method.

We construct a tag matrix $P \in R^{N_{im} \times N_{tg}}$, where each row corresponds to one image (the number of images is N_{im}), and each column corresponds to one tag (the number of tags is N_{tg}), such that $P_{ij} = 1$ if image i is annotated with tag j and 0 otherwise. Denote the set of observed entries by Ω, i.e. $\Omega = \{(i,j)|P_{ij} > 0\}$. We adopt the IMC method for tag refinement, which assumes that the tag matrix is generated by applying feature vectors associated with its row as well as column entities to the underlying low-rank matrix Z [9].

Let $x_i \in R^{f_{im}}$ denote the feature vector of image i, and $y_j \in R^{f_{tg}}$ denote the feature vector of tag j, which could be computed from pre-trained word2vec [11]. Let $X \in R^{N_{im} \times f_{im}}$ denote the feature matrix of N_{im} images, where the i-th row is the image feature vector x_i, and $Y \in R^{N_{tg} \times f_{tg}}$ denote the feature matrix of N_{tg} tags, where the i-th row is the tag feature y_i. Our goal is to recover the low-rank matrix $Z \in R^{f_{im} \times f_{tg}}$ using the observed entries from the tag matrix P, where P_{ij} is modeled as $P_{ij} = x_i^T Z y_j$. The idea is illustrated in Figure 3.

We formulate the matrix completion problem in a multi-label regression framework:

$$\min_{Z \in R^{f_{im} \times f_{tg}}} \sum_{(i,j) \in \Omega} loss(P_{ij}, x_i^T Z y_j) + \lambda rank(Z)$$

The loss function $loss$ penalizes the deviation of estimated entries from the observations. The regularization parameter λ trades off losses on observed entries and the low-rankness constraint. A common choice for loss function is the squared loss function given by $loss_{sq}(a,b) = (a-b)^2$. The low-rankness constraint on Z is NP-hard to solve. So we replace it with the standard relaxation, the trace norm, i.e. sum of singular values. Then we get the final object function:

$$\min_{Z \in R^{f_{im} \times f_{tg}}} \sum_{(i,j) \in \Omega} (P_{i,j} - x_i^T Z y_j)^2 + \lambda \|Z\|_*$$

Note that the object function is convex. So we can use the LADM [10] method to solve the problem.

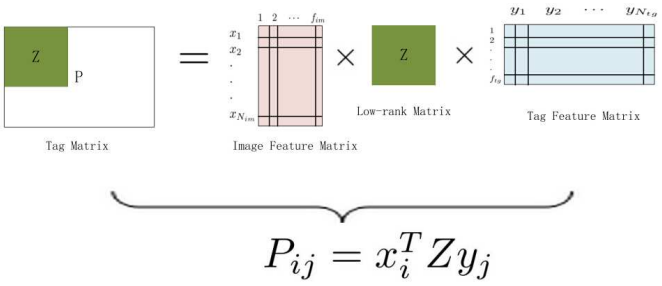


Fig. 3: The idea of the low-rank modeling of tag matrix. The shaded region in the tag matrix P corresponds to the underlying low-rank matrix Z.

III. Experimental Evaluation

The SCMC algorithm is evaluated on two well known benchmark datasets: MIRFlickr-$25K$ and Corel$5K$.

A. Datasets and Experimental Setup

The MIRFlickr-$25K$ dataset is collected from Flickr. Compared to the Corel$5K$ dataset, tags in MIRFlickr-$25K$ are rather noisy and many of them are misspelled or meaningless words. Hence, a pre-processing procedure is performed. We match each tag with entries in a Wikipedia thesaurus and only retain the tags in accordance with Wikipedia. We use the pre-trained word and phrase vectors [11] to extract tag vectors from the tags in these two datasets. As to the parameters in the LRR algorithm, we just adopt their default values. We may extend our model by establishing the latent connection between sub-images and corresponding tags by means of statistical machine translation [12], [13], [14], or exploring embedding approaches [15], [16], [17] that simultaneously learn the distributed representations for both the image and its tags in the further.

TABLE I: Statistics of Two Datasets

Statistics	**Corel**$5K$	**MIRFlickr-**$25K$
No. of images	4,918	25,000
Vocabulary Size	260	1,386
Tags per Image (mean/max)	3.4/5	12.7/76
Images per Tag (mean/max)	65.3/1,120	416.5/76,890

B. Comparisons to State-of-the-art Annotation Methods

We compare the proposed SCMC algorithm to the state-of-the-art methods, including LRR based model DFC-LRR [18], RPCA based model LRES [1], search-based algorithms

(i.e. JEC [6], TagProp [5], and TagRelevance [7]), mixture models (i.e. CMRM [4] and MBRM [19]), tag recommendation approaches (i.e. Vote+ [20] and Folk [21]) and Bayesian network model InfNet [22]. Note that the parameters of adopted baselines are also carefully tuned on the validate set of Corel$5K$ with corresponding proposed tuning strategy. We further compare the tag transfer algorithm employed by SCMC with the graph-based tag propagation algorithm proposed by an LRR-based method, DFC-LRR, which is much more complex. To make a fair comparison, the two algorithms run on the same affinity matrix calculated by LRR.

We measure all the algorithms in terms of *average precision*@N (i.e. $AP@N$), *average recall*@N (i.e. $AR@N$) and *coverage*@N (i.e. $C@N$). In the top N completed tags, *precision*@N is to measure the ratio of correct tags in the top N competed tags, *recall*@N is to measure the ratio of missing ground-truth tags, both averaged over all test images. *Coverage*@N is to measure the ratio of test images with at least one correctly completed tag.

Table II and III demonstrate comparisons on performance. Due to the limitation of space, we only report results when $N = 2, 3$. We observe that: 1) Generally algorithms achieve

TABLE II: Performance Comparison on Corel$5K$

	Corel$5K$					
	(N = 2)			(N = 3)		
	AP	**AR**	**C**	**AP**	**AR**	**C**
SCMC	**0.27**	**0.42**	**0.50**	**0.23**	**0.50**	**0.59**
DFC-LRR [18]	0.26	**0.41**	**0.50**	0.22	**0.50**	**0.59**
LRES [1]	**0.27**	0.39	0.47	**0.23**	0.47	0.57
JEC [6]	0.23	0.34	0.39	0.19	0.40	0.47
TagProp [5]	**0.27**	0.40	**0.50**	0.22	**0.48**	0.57
TagRel [7]	**0.27**	**0.41**	0.48	0.22	0.47	0.57
CMRM [4]	0.16	0.20	0.23	0.13	0.24	0.27
MBRM [19]	0.20	0.29	0.35	0.17	0.34	0.42
Vote+ [20]	0.23	0.34	0.40	0.19	0.40	0.48
Folk [21]	0.19	0.29	0.34	0.16	0.34	0.41
InfNet [22]	0.15	0.19	0.24	0.12	0.22	0.29

TABLE III: Performance Comparison on MIRFlickr-$25K$

	MIRFlickr-$25K$					
	(N = 2)			(N = 3)		
	AP	**AR**	**C**	**AP**	**AR**	**C**
SCMC	**0.25**	**0.38**	**0.42**	**0.20**	**0.41**	**0.54**
DFC-LRR [18]	**0.25**	0.34	0.40	0.19	0.40	**0.53**
LRES [1]	**0.25**	**0.35**	**0.42**	**0.20**	0.39	**0.53**
JEC [6]	0.20	0.30	0.32	0.16	0.38	0.45
TagProp [5]	0.23	**0.35**	0.39	0.19	0.42	0.51
TagRel [7]	0.24	0.34	0.37	**0.20**	**0.43**	0.52
CMRM [4]	0.12	0.15	0.16	0.11	0.21	0.24
MBRM [19]	0.13	0.16	0.18	0.14	0.30	0.35
Vote+ [20]	0.19	0.29	0.33	0.14	0.33	0.40
Folk [21]	0.12	0.16	0.19	0.13	0.22	0.36
InfNet [22]	0.09	0.10	0.14	0.07	0.18	0.24

better performance on Corel$5K$, since tags in MIRFlickr-$25K$ are more noisy. 2) Compared to the traditional mixture model baselines, search-based methods generally have remarkably better performance. 3) Subspace-based method, such as SCMC and LRES, always achieve the best performances, confirming our assumption on the subspace clustering property of the image datasets. 4) The tag transfer procedure in SCMC is a simpler algorithm with comparable performance. 5) SCMC nearly outperforms all the other algorithms in all cases. 6) Performance on MIRFlickr-$25K$ in some sense provides an evidence for the robustness of SCMC.

IV. Conclusion

In this paper, we propose an effective approach SCMC for automatic image tag completion and refinement. SCMC performs tag completion and tag refinement sequentially. It clusters images using LRR and shares tags using voting algorithm, then refines tags by IMC. Our model achieves the state-of-the-art performance in extensive experiments conducted on benchmark datasets for image annotation.

References

[1] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *ACM MM*, 2010.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *JMLR*, 2003.

[3] C. Wang, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *CVPR*, 2009.

[4] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *ACM SIGIR*, 2003.

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

[6] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *ECCV*, 2008.

[7] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *TM*, 2009.

[8] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *TPAMI*, 2013.

[9] P. Jain and I. S. Dhillon, "Provable inductive matrix completion," *arXiv preprint arXiv:1306.0626*, 2013.

[10] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *NIPS*, 2011.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] M. Fan, Q. Zhou, and T. F. Zheng, "Content-based semantic tag ranking for recommendation," in *WI-IAT*, 2012.

[13] M. Fan, Y. Xiao, and Q. Zhou, "Bringing the associative ability to social tag recommendation," in *Workshop Proceedings of TextGraphs-7 on Graph-based Methods for Natural Language Processing*, 2012.

[14] M. Fan, Q. Zhou, and T. F. Zheng, "Mining the personal interests of microbloggers via exploiting wikipedia knowledge," in *CICLing*, 2014.

[15] M. Fan, Q. Zhou, T. F. Zheng, and R. Grishman, "Large margin nearest neighbor embedding for knowledge representation," *arXiv preprint arXiv:1504.01684*, 2015.

[16] M. Fan, Q. Zhou, E. Chang, and T. F. Zheng, "Transition-based knowledge graph embedding with relational mapping properties," in *PACLIC*, 2014.

[17] M. Fan, K. Cao, Y. He, and R. Grishman, "Jointly embedding relations and mentions for knowledge population," *arXiv preprint arXiv:1504.01683*, 2015.

[18] A. Talwalkar, L. Mackey, Y. Mu, S.-F. Chang, and M. I. Jordan, "Distributed low-rank subspace segmentation," in *ICCV*, 2013.

[19] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *CVPR*, 2004.

[20] B. Sigurbjörnsson and R. Van Zwol, "Flickr tag recommendation based on collective knowledge," in *ACM WWW*, 2008.

[21] S. Lee, W. De Neve, K. N. Plataniotis, and Y. M. Ro, "Map-based image tag recommendation using a visual folksonomy," *PRL*, 2010.

[22] D. Metzler and R. Manmatha, "An inference network approach to image retrieval," in *CIVR*, 2004.