IMAGE ANNOTATION AND TAG COMPLETION VIA KERNEL METRIC LEARNING AND NOISY MATRIX RECOVERY

By

Zheyun Feng

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Doctor of Philosophy

2016

ABSTRACT

IMAGE ANNOTATION AND TAG COMPLETION VIA KERNEL METRIC LEARNING AND NOISY MATRIX RECOVERY

By

Zheyun Feng

In the last several years, with the ever-growing popularity of digital photography and social media, the number of images with user-provided tags has increased enormously. Due to the large amount and content versatility of these images, there is an urgent need to categorize, index, retrieve and browse these images via semantic tags (also called *attributes* or *keywords*). Following this trend, image annotation or tag completion out of missing and noisy given tags over large scale datasets has become an extremely hot topic in the interdisciplinary areas of machine learning and computer vision.

The overarching goal of this thesis is to reassess the image annotation and tag completion algorithms that mainly capture the essential relationship both between and within images and tags even when the given tag information is incomplete or noisy, so as to achieve a better performance in terms of both effectiveness and efficiency in image annotation and other tag relevant tasks including tag completion, tag ranking and tag refinement.

One of the key challenges in search-based image annotation models is to define an appropriate similarity measure (distance metric) between images, so as to assign unlabeled images with tags that are shared among similar labeled training images. Many kernel metric learning (KML) algorithms have been developed to serve as such a nonlinear distance metric. However, most of them suffer from high computational cost since the learned kernel metric needs to be projected into a positive semi-definite (PSD) cone. Besides, in image annotation tasks, existing KML algorithms require to convert image annotation tags into binary

constraints, which lead to a significant semantic information loss and severely reduces the annotation performance.

In this dissertation we propose a robust kernel metric learning (RKML) algorithm based on regression technique that is able to directly utilize the image tags. RKML is computationally efficient since the PSD property is automatically ensured by the regression technique. Numeric constraints over tags are also applied to better exploit the tag information and hence improve the annotation accuracy. Further, theoretical guarantees for RKML are provided, and its efficiency and effectiveness are also verified empirically by comparing it to state-of-the-art approaches of both distance metric learning and image annotation.

Since the user-provided image tags are always incomplete and noisy, we also propose a tag completion algorithm by noisy matrix recovery (TCMR) to simultaneously enrich the missing tags and remove the noisy ones. TCMR assumes that the observed tags are independently sampled from unknown distributions that are represented by a tag matrix, and our goal is to recover that tag matrix based on the partially revealed tags which could be noisy. We provide theoretical guarantees for TCMR with recovery error bounds. In addition, a graph Laplacian based component is introduced to enforce the recovered tags to be consistent with the visual contents of images. Our empirical study with multiple benchmark datasets for image tagging shows that the proposed algorithm outperforms state-of-the-art approaches in terms of both effectiveness and efficiency when handling missing and noisy tags.

To my parents, Shiying and Xuexiang.

ACKNOWLEDGMENTS

First and foremost, I feel indebted to my advisor, Professor Rong Jin, for his guidance, encouragement, and inspiring supervision throughout the course of this research work. His patience, prudential attitude, extensive knowledge, and creative thinking have been the source of inspiration for me. He was available for advice or academic help whenever I needed and gently guided me for deeper understanding, no matter how late or inconvenient the time was. When I was struggling to stop my career and leaving Paris four years ago, I was not sure about my decision, but now I am so happy and proud to say that I made a so wise decision. It is extremely hard to express how grateful I am for his unwavering support over the last years and in the coming future.

I would like to take on this opportunity to thank my thesis committee members Professor Anil K. Jain, Professor Joyce Y. Chai, and Professor Selin Aviyente who have accommodated my timing constraints despite their full schedules, and provided me with precious feedback for the presentation of the results, in both written and oral form.

During my Ph.D. studies, I had the pleasure of collaborating with many researchers from each and every one of which I had things to learn, and the quality of my research was considerably enhanced by these interactions. I would like to thank Professor Anil K. Jain who generously gave me so much general guidance in my research direction and paper polishing. I would like to thank Songhe Feng, Tianbao Yang and Fengjie Li for all the discussions we had and the fun moments we spent together on doing research and future planning. I also spent two wonderful summers as intern at Comcast Lab, DC with Dr. Jan Neumann. I learned a lot from his team and would like to express my gratitude for having me as a research intern. I also would like to thank Serhat Selçuk Bucak and Radha Chitta for some helpful discussions and suggestions.

Living in East Lansing without my good friends would not have been easy. I want to thank all my friends in the department and outside the department. I wish I could name you all.

Last but definitely not least, I want to express my deepest gratitude to my beloved parents and dearest boyfriend. Their love and unwavering support have been crucial to my success, and a constant source of comfort and counsel. Special thanks to my parents for abiding by my absence in last four years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Chapter 1

Introduction

We are facing the problem of image explosion: massive images have been provided through different sources including the Internet, camera network, research laboratories, personal digital devices and many other photo management applications. The Internet has greatly promoted the ability to release and access all manner of multimedia information especially images. For instance, KPCB analyst Mary Meeker's annual Internet Trends report states that all internet-connected citizens share over 1.8 billion photos each day [143] through multi-platform services such as Snapchat [1], Instagram [2], Facebook [3], WhatsApp [4], etc., as shown in Figure 1.1. Therefore, such a proliferation of images poses an urgent challenge for large scale image categorization, indexing, retrieval and browser.

In the image retrieval community, most methods can be categorized into two groups: content based image retrieval (CBIR) [176] and tag based image retrieval (TBIR) [135]. CBIR matches the query image and the gallery images based on their visual similarities that could be computed from a group of visual features including color, texture, shape [75], LBP [152], HOG [37], SIFT [137], GIST [154] and the list goes on [204]. Despite the elaborate system designing and computational efforts, the performance of CBIR is still prohibited by the notorious semantic gap between the low level visual features that reflect the image

[1] https://www.snapchat.com/
[2] https://www.instagram.com
[3] https://www.facebook.com/
[4] https://www.whatsapp.com/

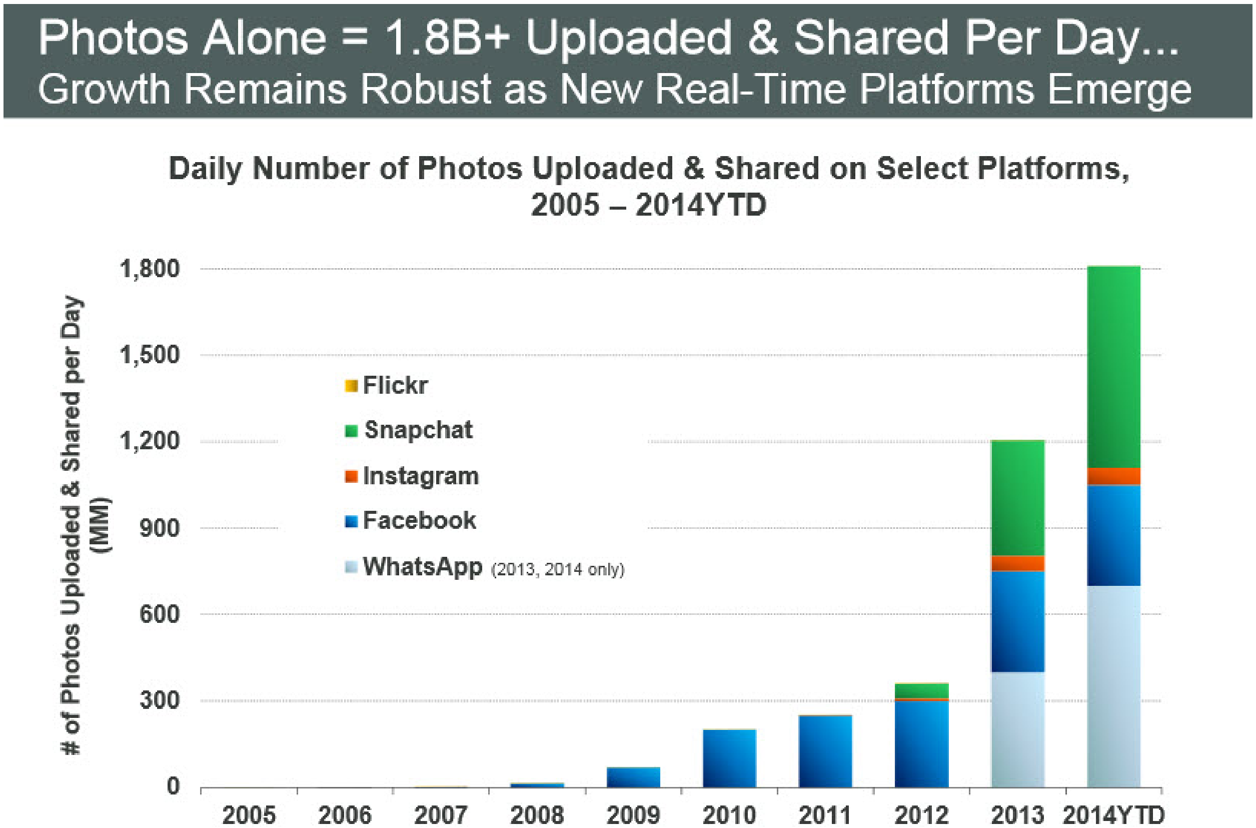


Figure 1.1: Daily number of images uploaded to the Internet through selected apps [143].

contents and the high level semantics behind images [67, 201]. To considerably improve the performance of CBIR, substantial advancements in terms of the involved technologies and designs are still required, which includes feature extraction, feature selection, indexing, query rephrasing and completion [132].

To overcome the limitations of CBIR in terms of both retrieval effectiveness and efficiency, the TBIR was proposed accordingly. Instead of the visual features which take great computation cost, TBIR represents images with a set of tags (also called keywords, labels or attributes). The user gives out the query as a sequence of semantic words, and then the relevant images are retrieved based on the matches between the textual query and the image tags. Compared with CBIR, TBIR has two significant advantages. First, it allows the users to better express their query needs with semantic words, which alleviates the semantic gap and improves the retrieval accuracy. And secondly, TBIR formulates the image retrieval problem as a document retrieval problem, which allows to use the inverted index technique [227] and greatly improves the retrieval efficiency.

Besides tag based image retrieval, there is also many other tag dependent tasks that categorize [116], indexing and browse [168] images via semantic tags for similar reasons. However, the good performance of these tasks substantially relies on the image set which is supposed to have sufficient high quality tags.

However, among the great amount of available images, only a small portion of images are associated with appropriate tags. Generally images are annotated manually, either by professional annotators or simply by the photo takers and reviewers. The professionally annotated tags are elegant and reliable, but cost tremendous label efforts and time. Typical such image datasets include CCUB NABirds 700 Dataset[5] and Microsoft COCO Dataset [125], which took years to collect, annotate and build up by a group of researchers. Apparently, this is prohibitively labor costing in terms of the proliferation of images [202]. Fortunately in most cases, the image tags are provided by the users who upload the image to social media (e.g. Flickr[6]) and the reviewers of this image, or directly crawled from the accompany descriptions/titles of that image. However tags generated in this way are far away from reliable, since they are usually general, ambiguous, biased, and sometimes even inappropriate, incomplete or redundant for many reasons according to [70, 101]. All these factors could severely prevent the performance of TBIR and other tag-based tasks.

As a result, the need for reliable tags over large scale images becomes profitable and emergent, which motivates the research community to develop effective and efficient auto-

[5] `http://info.allaboutbirds.org/nabirds/`.

[6] `https://www.flickr.com/`.

matic tagging systems [56]. Among them, image annotation and tag completion are two big branches that catch the most eyes.

The differences between these two branches come from the two types of the supervised information. In the image annotation framework, a subset of images is associated with appropriate tags and the other images are not assigned with any tag. And the goal is to predict tags for the unlabeled images. In the tag completion task, all the images are associated with certain number of tags. However some tags are appropriate while some others are not, and there are also some tags supposed to be observed but actually not. And its final purpose is to update the whole tag matrix to make it better describe the image's visual contents.

So basically, the goal of image annotation and tag completion work is to learn from labeled examples in order to predict the labels of other examples or the scores of the other labels. That is, given a training set of supervised information (examples or labels), it is aimed to learn a hypothesis that assigns each label a confidence score to associate a sample where the label or the sample could have never been observed by the algorithm. Efficiently finding such an effective hypothesis based on the training set and the observed labels, which minimizes some validation measure of performance, is the main focus of this learning.

This chapter is devoted to an overview of these two broad topics of tag assigning and amelioration, aiming to develop a general correspondence between or within the image visual contents and the semantic tags. Here we move towards to the definitions in a fairly non-technical manner and the formal detailed definitions will be given in Chapter 2.

1.1 Image Annotation

The objective of image annotation is to automatically annotate an image with appropriate tags, which exclusively reflect its visual content. Image annotation has been a hot topic of on-going research for more than a decade, and many techniques have been developed. Conventionally, image annotation is tackled as a machine learning problem, where two major components are included: visual feature extraction and mapping those features to semantic tags [132]. Feature extraction obtains significant patterns from images, and then the patterns are mapped to keywords in the semantic space via a set of machine learning algorithms, which capture the relationship between visual contents and semantic tags in one of three ways: (i) formalizing a statistical model between tags and visual features [22, 49, 119, 132]; (ii) casting the problem into a set of binary classification ones [47, 72]; and (iii) representing the tags as a matrix and treating the annotation problem as a matrix factorization [223] or matrix completion problem [126, 201]. The key of these methods is to train a reliable model with sufficiently accurate tags by optimizing image compactness and separability in a global sense. However, the semantic gap, and the imperfect tags usually lead to a biased model and result in a suboptimal solution. That means the discriminatory power of input images might vary between different neighborhood, and a global model hence cannot fit well the visual-semantic relation over the data manifold. Besides, many parametric models are not rich enough to effectively capture the complicate dependencies between image content and tags.

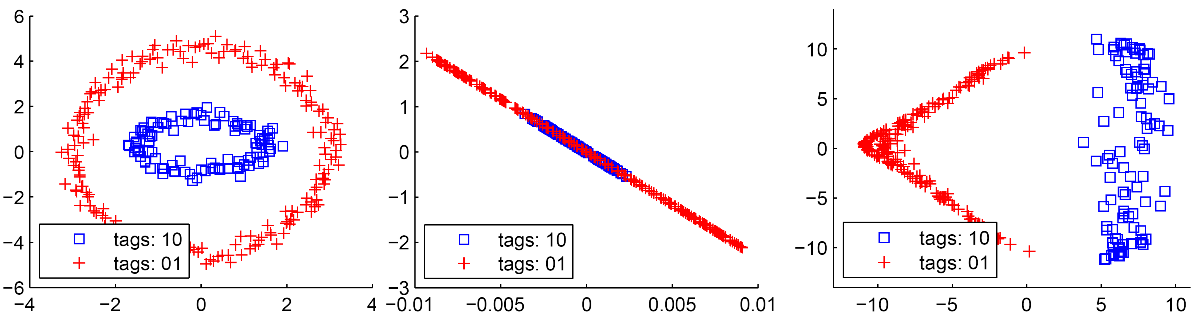
Recently, a local non-parametric model, the search based approach, has been proved to be quite effective, particularly for large image datasets with many keywords [67, 93, 139, 194]. Its key idea is to annotate a test image $\mathcal{I}$ with the common tags shared by the subset of training images that are visually similar to $\mathcal{I}$.

The crux of search based annotation methods is to effectively measure the visual similarity between images. *Distance metric learning* (DML) [78, 210, 205] tackles this problem by learning a metric that pulls semantically similar images close and pushes semantically dissimilar images far apart. Many studies on DML are restricted to learning a linear Mahalanobis distance metric in a finite dimensional space, which is expected to be consistent with the associated tags.

However, most distance metric learning algorithms assume all data is of linear separability [26], and they usually fail to capture the nonlinear relationships among images. To address this problem, several nonlinear DML algorithms have been proposed. Their key idea is to map data points from the original vector space to a high (or even infinite) dimensional space through a nonlinear mapping, which can be either explicitly constructed using boosting methods [73, 74, 172], or implicitly derived through kernel functions. And the latter is referred to as *Kernel Metric Learning* (KML) [26, 39, 187], which has been widely used to settle image similarity problems in image classification [39, 60, 187], clustering [2, 26, 205], and retrieval [77, 78]. Figure 1.2 illustrates an example comprised of two groups of data points annotated by different tags, and Figure 1.2(b) shows the distributions of data points adjusted by a learned linear distance metric, which fails to separate the objects with different tags, while in Figure 1.2(c) the newly learned nonlinear (kernel) distance metric successfully separates the data with different tags.

1.2 Image Tagging

Recently many user-provided tags are automatically generated, and thus are incomplete or inaccurate in describing the visual content of images [201]. In particular, these tags are



(a) Original data distribution (b) Linear distance metric (c) Nonlinear distance metric

Figure 1.2: An illustrative example for the comparison of linear and nonlinear distance metric learning algorithms. (a), (b) and (c) show the original data distribution, the distribution adjusted by a learned **linear** distance metric, and the distribution adjusted by a learned **kernel** metric, respectively.

crawled from the descriptions of the uploader and reviewers, and in most cases only a small number of tags are provided for each image and some are even irrelevant to the visual content of images. This disadvantage makes it difficult to fully utilize these accompany tags and limits their applications in tag dependent tasks such as tag based image retrieval and tag recommendation [126, 129, 226]. To better benefit from the tags, there is an urgent demand for efficient algorithms that are able to improve the tagging quality for a large scale of images, specifically, effective algorithms that can simultaneously recover the missing tags and remove or down weight the noisy tags.

Generally there are two groups of algorithms can fulfill this desire to re-weight the given tags: the fine-grained ones and the coarse ones. And both groups model mainly the correlations among the partially observed tags, then apply this model to the whole dictionary and re-weight all tags in the dictionary.

The fine-grained ones [40, 44, 116, 115] aim at the images with a specific contents and carefully manually annotated tags. The contents of images are simple and concentrated, while the semantic knowledge are complete and clean, and sometimes even manually associated to the correspondent segments in the images. Typical datasets for these algorithms are Animal with Attributes dataset [115] and CUB-200-2011 dataset [189]. The fine-grained group includes transfer learning [115], label propagation [40], zero shot recognition [116] and fine-grained categorization [44]. This group is able to learn new attributes by exploiting their graphical or probabilistic relationship via both an intermediate-level semantic representation and the low-level mapping between tags and their correspondent segments in the image repository.

However, natural images are usually assorted and involve a large scope of topics including scene, people, animal, action, objects and many other aspects of life and environment. Besides, because of the large amount and versatility of the images, the tags are usually user-provided instead of manually annotated, and thus contain many missing annotations and errors. And probably, the tag dictionary might be very long. Typical datasets could be referred to Flickr1M dataset [201] and NUS-WIDE [33] dataset. In this case the fine grained methods are no longer capable to capture the tag-tag or tag-image dependencies since on the one hand, the manual segmentation and tag localization are labor costing; and on the other hand the observed tags are usually too problematic to train a reliable model. In the contrast, the coarse methods are able to simultaneously address the challenges of missing and noisy tags [33] for a huge variety of images using machine learning techniques, which rely mainly on the semantic information that covers a wide range of knowledge, and as well as the auxiliary visual content information.

We refer to this problem as **tag completion** [33] to distinguish it from previous coarse image tagging work. Although the final objective of those tagging works is to assign an image with complete and exact tags with reasonable confidence scores, their initial setups vary, as illustrated in Figure 1.3. *Image annotation* [28, 67, 139] automatically assigns unla-

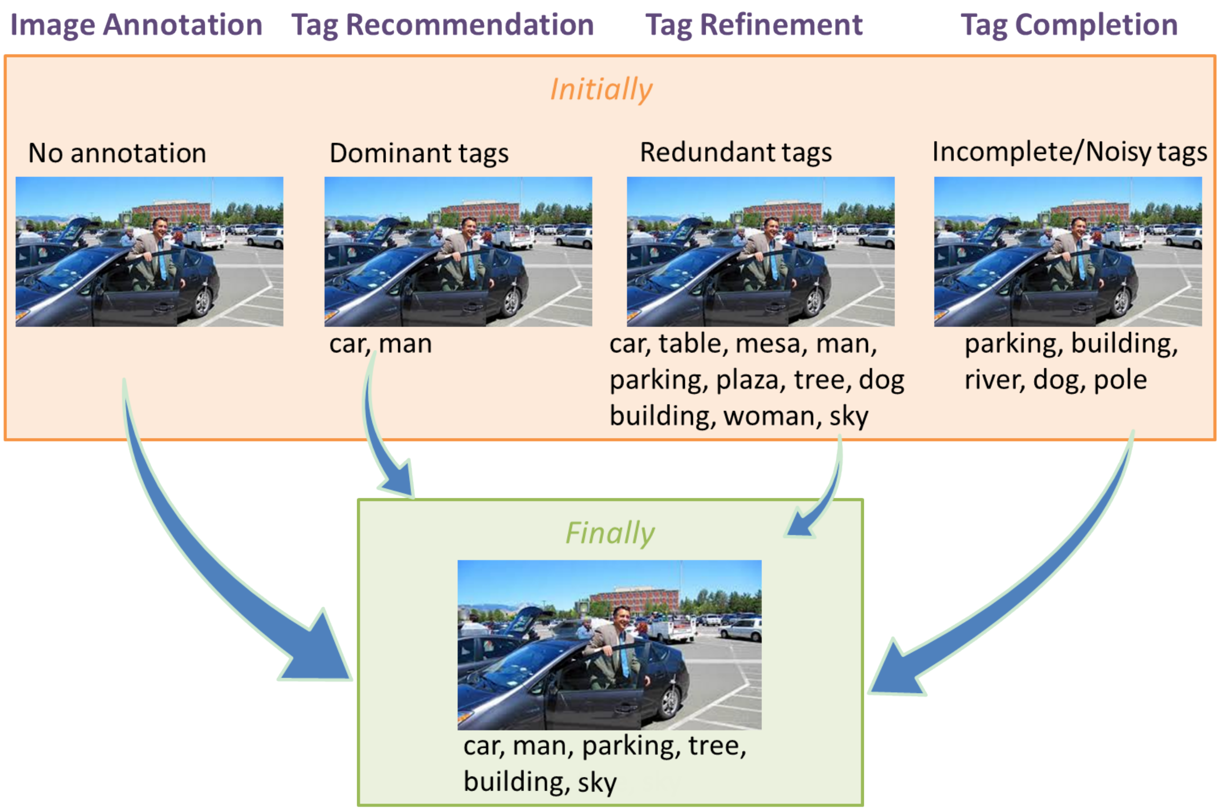


Figure 1.3: Exemplar illustration of tag completion and other image tagging works including image annotation, tag recommendation and tag refinement. The upper box shows the initially given information (both visual and semantic), and the bottom box indicates the ultimate objective of all four tasks.

beled images with appropriate keywords. As a state-of-the-art image annotation approach, search based algorithms [52, 67] rely on the quality of tags assigned to training images [52]. *Tag recommendation* suggests candidate tags to online annotators in order to improve the efficiency and quality of the tagging process [108, 158, 206]. It usually identifies missing tags by topic models (e.g. *Latent Dirichlet Allocation* (LDA)) [12, 108, 225], but does not address the noisy tag problem, an important issue in exploiting user-provided tags. *Tag refinement* applies various techniques, including topic model, tag propagation, sparse training and partial supervision [28, 131, 206], to select a subset out of the user-provided tags based on image features and tag correlation [224]. Although it is able to handle noisy tags, it cannot explicitly enrich the missing tags.

1.3 Thesis Contributions

In this section we shall elaborate on the main problems considered in this thesis and our key contributions to address these problems.

This dissertation mainly deals with the image annotation and tag completion problems, giving theoretical guarantees and providing empirical comparisons with state-of-the-art baseline algorithms. Generally, we attempt to delve into the image-image correlation, image-tag mapping and tag-tag interaction to capture their underlying relationship. In particular, the main contributions can be summarized as follows.

- **New effective image distance metric and its theoretical foundation**. The dissertation proposes an novel kernel based distance metric learning algorithm (RKML) specifically for image annotation in Chapter 3. This algorithm achieves success by fully exploring the image-tag dependency, which is consequently used to better capture the nonlinear complexities among images. Many strategies are applied to guarantee the annotation accuracy, including the incorporation of soft semantic constraints which better explore the semantic information between tags, and the adoption of rank based regularization term which effectively reduces the overfitting risk to the training data. Besides, the theoretical guarantee for the proposed kernel distance metric learning is provided.

- **Efficient kernel metric learning and related computation**. The proposed RKML algorithm explores the regression technique to avoid the projection to PSD cone, which is necessary in distance metric learning and is intensively computationally expensive.

Besides, Nyström approximation is applied in the kernel computation to speed up the implementation. These skills as well as the rank based regularization greatly reduce the computation burden for the proposed RKML algorithm.

- **Novel image tag completion work that effectively dealing with missing and noisy tags**. The dissertation also proposes a novel image tag matrix completion (TCMR) framework in Chapter 4 that effectively recovers the expected tags from incomplete and noisy given tags. This algorithm focuses to capture the tag-tag correlation, and then uses it to reversely update the tag confidence score matrix. Based on the idea of topic model, TCMR assumes that the observed tags of any image are drawn independently from a mixture of a small number of multinomial distributions, which can be straightforwardly interpreted as the low rank matrix completion theory. So following this theory, the nuclear norm is applied to simultaneously capture the interactions among tags in two ways, either between different tag keywords or between tag vectors associated with different images. Maximum likelihood component is also employed as the loss function, which successfully connects probabilistic models and matrix completion theory. All these techniques are applied to ensure the performance of tag matrix recovery out of missing and noisy tags.

- **Theoretical guarantee of image tag completion by noisy matrix recovery**. The final objective function of the optimization problem is convex, which guarantees that the global optimal solution exists and it would be efficient to find this optimal solution. The error bounds between the recovered matrix and the statistically optimal one are also provided theoretically.

1.4 Thesis Overview

The remainder of this dissertation is organized as follows. Chapter 2 lays out the foundation for the rest of the dissertation. In particular, we provide a survey on some of the background materials including image tagging tasks like image annotation and image tag completion, distance metric learning (both linear and kernel), statistical models applied in image tagging work, and as well as low rank matrix recovery theory. It will become clear in this chapter that there exist deep connections between these topics.

The first part of the thesis focuses on the image annotation problem. In Chapter 3 we focus on the kernel distance metric learning problem, investigate how it affects the image annotation performance, and propose strategies to solve the limitations in existing kernel distance metric learning algorithms and their applications in real-world.

The second part of the thesis deals with the image tag completion problem. Chapter 4 discusses its relationship to the statistical/topic models and matrix completion theory. The effectiveness of the proposed algorithm is justified both theoretically by the recovery error bounds and empirically on a bunch of datasets in terms of several of setups.

Finally, Chapter 5 summarizes this work by concluding the main contributions, some potential extensions and the future research directions. Besides, the appendix summarizes rather standard things on relevant topics of this work, and gives the error bounds that are used in the proof of results in the thesis and is mainly for reference. In order to facilitate independent reading of various chapters, some of the definitions are repeated for multiple times.

1.5 Notation

This section serves as a glossary for the main mathematical symbols used throughout the thesis. Vectors are shown by lower case bold letters, such as $\mathbf{x} \in \mathbb{R}^d$. Such a vector usually represents the visual feature or tag vector of an image. Matrices are indicated by uppercase letters such as A and their pseudo-inverse is represented by $A^\dagger$. We use $[m]$ as a shorthand for the set of integers $\{1, 2, \ldots, m\}$. Throughout the paper we denote by $|\cdot|$, $|\cdot|_1$, $|\cdot|_F$ and $|\cdot|_*$ the ℓ_2 (Euclidean) norm, ℓ_1-norm, Frobenius norm and spectral norm, respectively.

1.6 Bibliographic Notes for Previous Publications

Some of the results in this dissertation have appeared in prior publications.

The material in Chapter 3 is based on a work published in the International Conference on Computer Vision [52] (ICCV), the content of Chapter 4 comes from [51] which is published at the European Conference on Computer Vision (ECCV), and [50] which is published at the IEEE Transactions on Image Processing.

Chapter 2

Background

The goal of this chapter is to give a general and formal overview of the materials related to the work that has been done in this thesis. In particular, we will discuss the key concepts and questions relevant to problems of image annotation, image tag completion, kernel distance metric learning and noisy matrix recovery. The exposition given here is necessarily very brief and the detailed discussion will be provided in the relevant chapters.

2.1 Image Representation

In the computer vision area, image representation plays an important and ineluctable role. Specifically, appropriate feature representation significantly improves the performance of typical image relevant tasks including image classification, image clustering, image understanding, video understanding, etc. Since an image consists of an unstructured array of pixels,the first step of image representation is to extract efficiently certain types of discriminative visual features from these pixels, either colorful or grayscale [220]. Various feature extraction techniques will be reviewed in detail in the following sub-sections.

2.1.1 Color Feature

Color feature is one of the most basic and fundamental features to capture the image characteristics, which is usually defined subject to a particular color space, such as RGB, HSV, and $L\alpha\beta$ spaces [41, 65]. Within these spaces, color features could be extracted, including color histogram [204], color moments [220] and color coherence vector [220].

2.1.2 Texture Feature

Unlike color features which measured the property of a single pixel, texture features explore the traits of a group of pixels. According to the extracted domain, texture features can be divided into two groups including spatial texture feature and spectral texture features [220].

Spatial texture features are usually extracted by computing the pixel statistics, searching local pixel patterns or converting with stochastic/generative models in the original image space. Typical spatial features include texon histogram [140] and Markov random field [97]. Generally, since spatial features are directly generated in the original image space, they could be straightforwardly extracted from irregular shaped regions, while they usually suffer severely from the noise, mutation and distortions of images [220].

Spectral texture features serve as significant image analysis tools in the Computer Vision area in early 2000s, and they are usually extracted in the frequency domain that is transformed from the original image space. Common spectral texture features includes Fourier transform (FT) [122], discrete cosine transform (DCT) [164], wavelet [85] and Gabor filters [128]. Among them, FT and DCT are efficient but sensitive to scale and rotation, wavelet is fast computed but limited to orientations, and Gabor feature is robust to scale and orientation but would lose certain spectral information due to the incomplete cover of spectrum [183].

2.1.3 Typical Features in Image Tagging

Here, several state-of-art image visual features are summarized and compared in detail, which are potentially useful for image level tasks including image annotation and image tagging.

SIFT feature [137] is initially proposed for object recognition. It first extracts the SIFT descriptors from a set of reference images at different scales with Gaussian filters and then uses bag-of-words model to computer the histogram of the descriptors to form the final image feature. There are various versions of SIFT features including sparse SIFT, dense SIFT and SURF. Sparse SIFT [218] builds the features at Hessian-affin and Dense SIFT [120] extract the descriptors within a flat window. SURF [7] is a speed-up version of SIFT which take care of the scale problem by a convolution with box filters and handle the orientation problem with wavelet responses.

Gist feature [155] is initially described as a low dimensional representation of the scene and specifically for scene recognition, which requires no image segmentation as in tradition. It summarizes the gradient information, both scales and orientations, by convolving the image with a bank of Gabor filters [128], which provides a rough description of the image characteristics.

HOG feature [38, 48] is reported to provide excellent performance for object and human detection. It first densely extracts the histogram of oriented edges (HOG) descriptors and stacks the neighboring HOG descriptors together to increase the feature dimension and the descriptive power as well. Bag-of-words model is used later to finally compute the HOG feature for an image.

LBP feature [153], short for Local Binary Patterns, is a powerful texture feature based on occurrence histogram of local binary patterns. Basically LBP divides the image into blocks, for example 3×3, then threshold the block with the center pixel value and encode it into a sequence of binary digits. The sequence is then converted to a decimal number which is set as the value of the center pixel. Thus the histogram of each block can be computed and concatenated together to form a feature vector of the representing image. Essentially, LBP encodes the local contrast and patterns, making it highly discriminative while computed efficiently.

2.2 Image Annotation

Once sufficient visual features are extracted from the image, high level semantics like annotations and tags could be learned immediately from the given information. According to [67], traditional automatic image annotation methods can be categorized into three groups, while recently new deep neural network based models have also gained more and more attention in the annotation community.

2.2.1 Generative Models

This type of models usually trains global probabilistic models to explain the co-occurrence between image visual features and semantic labels, and then predict new tags with the newly learned relationship. Among them, many are borrowed from the techniques of natural language and text-based document processing. Duygulu et al. tried to translate image blobs into label keywords directly using a machine translation model [46], which inspired several relevance models. These early works, including Cross-Media Relevance Model [92], Continuous-Space Relevance Model [119] and Multiple Bernoulli Relevance Model [49], assumes the blobs and tags are conditionally independent given an specific image. Besides, an algorithm in [22] is designed to model the joint distribution between tags and visual features with a mixture distribution, while [145] models the visual and semantic relationship via Bayesian network.

Meanwhile, latent space models derived from natural language and text processing, including Latent Semantic Analysis [45] and Probabilistic Latent Semantic Analysis [76], and variants of Latent Dirichlet Allocation models [4, 148, 132, 108, 158] have been successfully applied to image annotation.

Besides the previous large groups, in [195] the authors propose a semi-supervised formulation based on linear regression with a tag-biased regularization.

These methods usually have unsatisfactory performance since the probabilistic models are too global to capture the nonlinear relation between images.

2.2.2 Discriminative Models

Image annotation can also be viewed as a classification problem where each keyword is treated as an independent class. As a state-of-the-art classifier, Support Vector Machine (SVM) has been shown with high effectiveness when handling high dimensional data like image. An SVM classifier is basically a binary classifier, so in order to be adaptive to the image annotation tasks which requires multiple classifier, some SVM-based annotation models first train a separate SVM for each concept with each classifier generating a probability, and later fuse all the SVM classifiers together to get a final confidence score for each tag [36, 47]. Further, a batch mode re-tagging method is proposed in [27], where a SVM with augmented features is proposed to learn adapted a set of classifiers to refine the existing noisy tags.

Besides SVM, there is a group of other discriminative models that have been successfully applied in image annotation. [139] assigns tags by a k nearest neighbor classifier combining with multiple distance metrics. [144] applies a structural model to attribute-based image classification, and transfers the user inputs as well as the attribute-class mapping results to predicted tags. [13] learns the class labels by exploiting the group lasso technique and minimizing the ranking errors. Commonly for these methods, both the training and testing phases are computationally expensive. But [72] raises a max-margin formulation that models the dense pairwise label correlations, and reduces the complexity from exponential to linear. [157] also learns a multi-label classifier that explicitly and efficiently models the dependencies between submodular pairwise labels via graph-cut, and directly optimizes the F-score.

In [133], a multiview Hessian discriminative sparse coding is presented, which exploits Hessian regularization to steer the solution which varies smoothly along geodesics in the manifold, and treats the label information as an additional view of feature for incorporating the discriminative power for image annotation. In [79], R. Hong et al. explore both the positive and negative tag correlations and propose an method with discriminative feature mapping, which selects the effective features from a large and diverse set of low-level features for each concept under multiple-instance learning settings.

Despite the considerable performance in learning image annotations, this group of algorithms shares the same shortcomings that they have poor scalability on large datasets or when the tag dictionary is large; and they also perform unsatisfactory especially when the training tags are incomplete or noisy.

2.2.3 Search based Models

Since image annotation is a highly nonlinear problem, parametric models might not be sufficient to capture the complex distribution of the data, recent works on image tagging have mostly focused on nonparametric nearest-neighbor methods, which offer higher expressive power. Search based approaches have gained much popularity in the exploring of tag relevance due to its feasibility on large scale data. Recent studies on image annotation show that search based approaches are more effective than both generative and discriminative models [67, 202]. Here, we briefly review the most popular search-based approaches developed for image annotation.

TagProp [67] constructs a similarity graph for all images, and propagates the label information from the training images to testing images via this graph. In [123] a majority voting scheme among the neighboring images is proposed. [130] obtains the tag relevance score using kernel density estimation, and then performs random walk to boost the primary tag relevance score over the tag proximity graph that is constructed from the neighboring images. A sparse coding scheme is proposed in [59] to select semantically related images for tag propagation, and then local and global ranking agglomeration is adopted to down weight the noisy tags. Besides, conditional Random Field model is adopted in both [93] and [203] to capture the spatial correlation between annotations of neighboring images, but [93] embeds the kernelized logistic regression with multiple visual distance metric learning while [203] optimizes the model by maximizing margins of the hinge loss function.

This category of works usually concerns more on search technique or visual-semantic consistency problems, where much attention has been paid to learn effective and efficient distance metrics.

2.2.4 Neural Network based Models

Neural network based image annotation models typically includes conventional artificial neural network (ANN) [55] and recent developed deep convolutional neural network (CNN) [109].

An ANN consists of multiple layers of nodes called neurons, and nodes in different layers are connected by edges with correspondent weights. Each neuron works by inputting the outputs of the previous layers and the weights of its connecting edges into an activation function to generate a final output. Figure 2.1 shows how an ANN annotates an image with three tags. As an example, four 3-layer ANNs are used in [112] to annotate image regions.

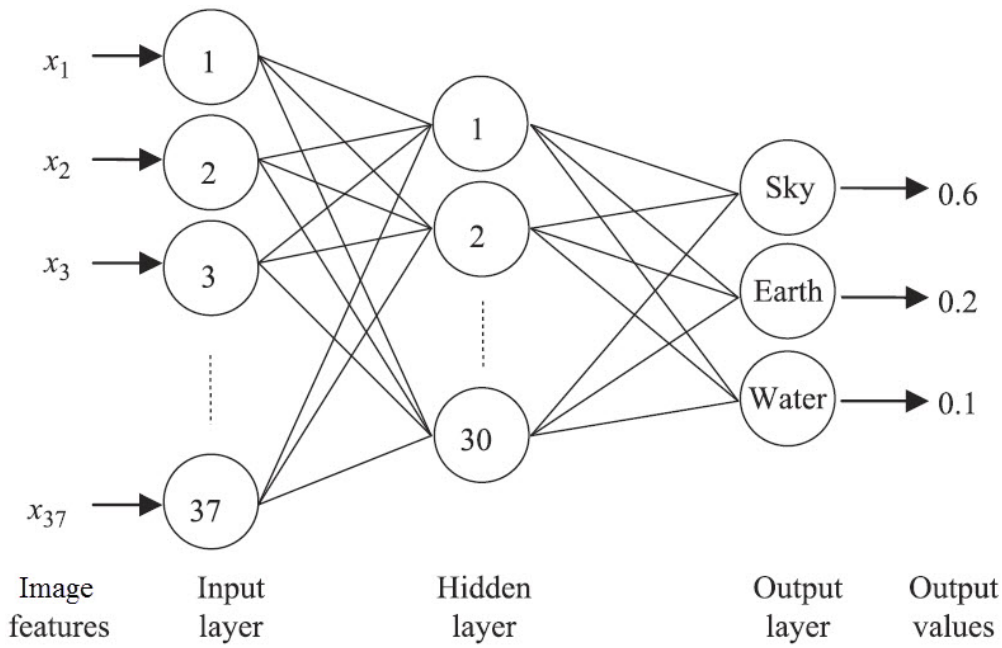


Figure 2.1: Annotate an image with ANN [220].

Very recently, deep convolutional neural networks (CNN) have demonstrated promising results for image classification [109], and features based on CNNs have also shown potential to significantly boost performance in terms of image annotation and tagging [64, 180]. [64] proposes a feature based on DNN that combines convolutional architectures with approximate top-k ranking objectives, and finally overwhelmingly outperforms the traditional visual feature in multiple image annotation jobs. [34] builds many sparsely connected neural layers by training only the winner-take-all neurons, which yields large network depth and excellent performance on image annotation tasks. In [214], Yang et al. tactfully apply deep neural network to establish the correlations between visual features and semantics, and address the imbalanced keyword distribution by incorporating the keyword frequencies and log-entropy.

This group has some distinctive pros and cons. For massive input and output data, when we have no idea what the function mapping between the two together is, neural network can learn this function without having to explicitly provide it. And it also well handles defective data sets with noise and missing variables. Nevertheless, emerging from a neural network's weights can be difficult to understand, specifically, it may work, but it is hard to explain the literal and physical meaning, and there is no theoretical guarantees. Sometimes, its training takes longer than certain other methods of machine learning.

2.3 Image Tagging

Image tag was initially applied to improve the performance of content-based image retrieval [114], and then image tagging works were developed to generate tags by associating semantic words to unlabeled images [5]. Probabilistic and language models are widely used in early models that match the semantics and images [4, 5, 119]. As the image annotation problem, the image tagging problem can also be formulated as a multi-label classification problem where each image can be assigned to more than one class simultaneously [14]. And following this idea, there are many multiclass techniques, including SVM, CRF, and some other works such as [14, 15, 115], that has been modified to adapt to the image tagging problem.

Similar as literature on image annotation, most existing image tagging works explore only the relationship between the visual features the tags, for instance, the direct mapping between visual and tag spaces, the probabilistic dependencies and the graphical model between visual contents and tags[67, 126, 129, 201].

To achieve a better tagging performance, some works try to learn a better mapping by studying the precise tag localization or an adaptive distance metric. [15] and [14] factorize the Bags-of-Words feature as a weighted sum of class histograms plus an error to model the image content, and thus pose the multi-label classification problem as a rank minimization problem. [123] proposes to scalably and reliably learn the tag relevance vector of an image by accumulatively votes the tags associated to its similar images (nearest neighbors). [52] and [202] apply distance metric learning methods to capture the dependency between visual and textual contents.

However, since compared to image annotation, additional tag information are observed in image tagging tasks, some other works delve into the textual correlations among tags. [40] introduces a so-called *Hierarchy and Exclusion graph* to encode the rich semantic relations including mutual exclusion, overlap and subsumption. [42] maps both images and text to a common semantic space using *word embedding*, which improves the tagging performance by avoiding direct cross-modal mapping that is always impractical to be constructed.

Besides, other works mainly follow the ideas of topic model and matrix completion. They usually explore the mutual dependencies between tags and then solve an optimization problem derived from the image-tag relation [58, 151, 193, 215, 224].

And in our study, we focus on the essential correlations among different tags, which can be effectively recovered out of the incomplete and noisy observed tags by the noisy matrix recovery model. In order to provide a more comprehensive presentation on this model, we further review the image tag completion works as well as the closely relevant topic model based image tagging approaches as follows.

2.3.1 Image Tagging with Topic Models

Topic model is originally designed for document clustering [12, 11] which discovers the abstract ”topics” that occur in a collection of documents. Figure 2.2 shows a typical topic model pipeline. It is first assumed that there is a hand of ’topics’ in the collection of documents, as shown in the left column, and each topic could be modeled by a distribution over a set of words. Then the generation of a document can be described as follows. First, a distribution over the topics (the histogram at right) should be chosen, and then for each word, we choose a topic assignment and choose the word from the corresponding topic [11].

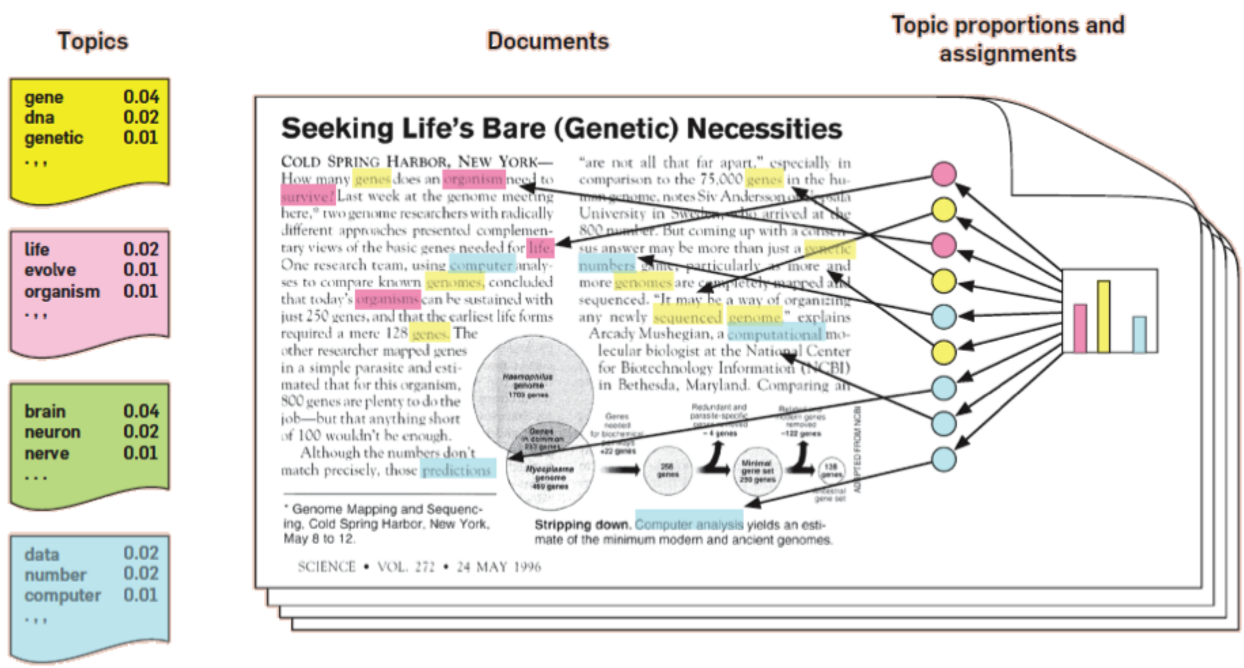


Figure 2.2: The illustration of how topic model works [11], where each topic is highlighted by a specific color.

In the last decade, Topic models has been widely applied in image understanding and tag recovery applications [151, 225]. [206] applies topic model to tag refinement by jointly modeling tag similarity and tag relevance. [108] uses LDA [12] to discover latent topics from resources with complete tag annotations, and discovered topics are then used to recommend topics for new resources that are annotated with only a few tags. [158] presents a topic-regression multi-modal LDA for image annotation. However all these methods focus on the simple co-occurrence of tags and fail to capture their underlying dependencies, and thus work poorly on imperfect tags. Recently, [151] encodes the textual tags as relations among the images, and then uses topic model to learn the image content and modify their encoded relations. [87] extends traditional LDA to noisy tags by additionally introducing a general distribution unrelated to the image content which leads to the noisy tags. The key limitation of these proposed topic models are (i) they have to solve a non-convex optimization, and (ii) they usually do not have any theoretical guarantee on the learned models.

2.3.2 Image Tag Completion

There are only a handful studies fitting the category of tag completion with both incomplete and noisy tags. [226] proposes a data-driven framework for tag ranking that optimizes the correlation between visual cues and assigned tags. In [129] the noisy tags are first removed based on the visual and semantic similarities, and then tags are obtained by expanding the observed tags with their synonyms and hypernyms using WordNet. [201] proposes to search for the optimal tag matrix that is consistent with both observed tags and visual similarity. [190] proposes to complete the missing tags by a *local linear learning*, which constructs a unified objective function to calculate the tag scoring vector for each image among its neighborhood. In [208], the authors propose an image-tag re-weighting scheme to adjust the penalty of each tag and image based on both image similarities and tag associations, and therefore formulate a unified re-weighted empirical loss function to handle the defective setting with both incomplete and noisy tags. Despite the successful application, none of these studies provides any theoretical guarantee for their approaches.

Besides, matrix decomposition is adopted in literature including [15, 149, 223, 224] to handle both missing and noisy tags. [134] formulates tag completion into a non-negative data factorization problem. [126] exploits sparse learning techniques to reconstruct the tag matrix. The key limitation of these approaches is that they require a full observed tag matrix with a small number of errors, making it inappropriate for tag completion problem.

2.4 Image Annotation by Metric Learning

It is ubiquitous to find appropriate measures to represent the distance or similarity between data in research and engineering communities including machine learning, computer vision, information retrieval and data mining, which increases the emergence of distance metric learning (DML) [9]. Euclidean distance is the simplest and most generally used distance metric, but despite easily used, hardly it is able to capture the irregularities and idiosyncrasies of the complicated and versatile data. The studies of DML can be traced back to 2002 [205], and immediately it becomes a hot topic and inspires many research work. Yang et al. [210], Kulis [110] and Bellet et al. [9] have comprehensive yet detailed surveys on this topic including problem formulation, optimization and applications. Given the rich literature on this subject, we only discuss the metric learning studies closely related to image annotation, we refer the readers to [9, 67, 202, 210] for more detailed surveys on the focused topic, if necessary.

The goal of distance metric learning is to take advantage of the prior information in form of labels/tags or pairwise constraints to create a projection of the data into another space such that the relevant images have smaller distances and share more labels while irrelevant images have larger distances and share fewer labels. According to the linearity of projection, we roughly categorize DML into two groups: linear and nonlinear distance metric learning.

Besides, there are also some extensions including online learning and local metric learning. Certain parts of these groups may overlap.

2.4.1 Linear Distance Metric Learning

Most linear DML methods assume data points lie in a finite linear space, and focus on Mahalanobis metric learning problem setting, written as

$$d_M(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^\top M(\mathbf{x} - \mathbf{x}')},$$

where the metric M should be symmetric and positive semidefinite (PSD). Most notable works for learning such a Mahalanobis distance fall into several groups [210].

The first group learns metrics with explicit *class labels* (may also be referred to *tags*, *concepts* or *keywords*). For instance, NCA [63] explicitly learns metrics through a k-nearest neighbor classification, MLCC [60] constructs a convex problem leading to a metric that collapses same class samples to a single point and pushes samples in the other classes infinitely far away, LMNN [196] extends the K-NN based works by achieving maximal margin nearest neighbor classification, and LDML [68] models the image similarity using posteriori class probabilities and obtains the distance metric by maximizing the log-likelihood.

The second group learns metrics from *pairwise constraints* and typically includes following examples. NMC [205] proposes a convex formulation that maximizes the sum of distance between dissimilar points while keeping the sum of distance between similar examples small. RCA [3] learns a distance metric through a set of positive constraints (must-link), and later DCA [78] and ERCA [216] extend RCA by additionally introducing negative constraints (cannot-link) at the cost of a more expensive algorithm. LRML [77] provides a semi-supervised metric by integrating the unlabeled data information and a graph regularization. ITML [39] introduces LogDet divergence regularization and minimizes the differential entropy under both positive and negative constraints. In ITML, a Bregman divergence defined as

$$D_{ld}(M, M_0) = tr(MM_0^{-1}) - \log\det(MM_0^{-1}) - d,$$

where d is the dimension of the input space, is introduced to keep the learned metric to be as identical as possible to the Euclidean metric ($\mathbf{I}$). In ITML, it is automatical and pretty easy to guarantee the positive semidefiniteness of M by minimizing $D_{ld}(M, M_0)$, due to the fact that the LogDet divergence is finite if and only if M is positive definite. Furthermore, [91] and SDML [160] follow this idea and propose more efficient Mahalanobis distance learning algorithms. Besides the LogDet divergence regularization, SDML [160] also employs an extra L_1 regularization on the off-diagonal elements of M to speedup the computation in high dimensional space while make it theoretically descent. And moreover, to handle the noisy constraints, RML [82] minimizes the worst-case violation over all possible sets of correct constraints.

Exceptionally, despite the popularity of DML algorithms that taking care of class labels and constraints, only a few works are designed to handle other types of supervised information such as annotated tags. For image annotation tasks, [93, 200, 202] propose to explore metrics from implicit side information instead of class assignments or pairwise constraints. KCRF [93] embeds a Kernelized Logistic Regression (KLR) with multiple visual distance learning into a unified Conditional Random Fields (CRF) framework. PRCA [200] first proposes a probabilistic metric learning out of the probabilistic side information based on a graphical model. UDML [202] unifies both inductive and transductive metric learning techniques to effectively exploit both visual and textual image contents. Besides, MLR [142] learns a metric for solving ranking and retrieval tasks, and its extension R-MLR [124] additionally deals with the noisy features using a mixed $L_{2,1}$ norm to ignore most of the irrelevant features.

Furthermore, the linear similarity metric learning is usually an alternative of linear distance metric learning. The only difference is that similarity measure does not necessary have distance properties, especially the PSD and symmetric requirements, and as a result it is usually more flexible and scalable to large data. Typical linear similarity metric learning algorithms include SiLA [159], OASIS [25], SLLC [8] and RSL [30].

2.4.2 Nonlinear Distance Metric Learning

Due to the multimodal distributions of real-world data, recently a number of nonlinear distance metric learning approaches have been developed to tackle these nonlinear patterns. The main idea of nonlinear metric learning is to learn a linear metric in a reproduced nonlinear feature space. Depending on how the nonlinear mapping is constructed, the nonlinear DML family is usually classified into two categories, boosting based approaches [73, 74, 172] and kernel based approaches [39, 78, 187].

Typical boosting methods are listed as follows. BoostDist [73] combines boosting hypotheses over the product space with a weak learner that is based on partitioning the original feature space. BoostMetric [173] applies a set of positive semidefinite matrices with trace and rank being one as weak learners to an boosting based learning process. And GB-LMNN [98] applies gradient boosting to learn a nonlinear mapping directly in the function space.

Initial kernel metric learning (KML) algorithms, such as KPCA [169], Kernel NMC [205], Kernel MCML [60], Kernel DCA [78], KLMCA [187], Kernel ITML [39] and KernelBoost [74], directly extend their linear or boosting based counterparts to kernel metric learning using the kernel tricks. Although several approaches have been empirically shown to be able to kernelizable, in general kernelizing setting, a specific metric learning is not trivial. It involves a new problem formulation where the interface of data is limited to inner products, and a $n \times n$ matrix is ineluctable to learn. Besides, as the number of training examples n increases, the problem becomes intractable. These problems together yield an extremely different and difficult solution as in the linear space. To address this problem, a hand of general kernelization extension works [24, 219] have been developed based on KPCA [169]. A so-called *KPCA trick*, which introduces a kernel to project the data into a nonlinear space followed by a dimensionality reduction strategy, is adopted and its soundness is justified theoretically through representer theorems [24]. It is also possible to obtain general kernelization through the equivalence between Mahalanobis distance learning and linear transformation kernel learning with spectral regularizers [90, 89]. In preactical implementation, such an appropriate kernel function could be select through a multiple kernel framework that is proposed in [191].

In parallel, some other KML works straightforwardly propose new kernel based metric learning frameworks. [2] proposes an explicit kernel transformation to tackle a constrained trace ratio optimization problem. It exploits both positive and negative constraints and as well as the topological structure of data. The suggested implementation of this KML algorithm is quite efficient since it is not necessary to learn all entries in the $n \times n$ metric matrix. NAML [26] formulates a trace maximization problem to joint kernel learning, dimension reduction and clustering together and solves it in a EM framework. [209] proposes a support vector metric learning (SVML) that co-joints a Mahalanobis distance and the SVM model with a RBF kernel, where the PSD constraint is automatically guaranteed and the metric can be made low rank.

However, although literature has shown that kernel metric learning may dramatically improve the quality of learned distance over highly nonlinear data, it also suffers from the computational burden and easily cause data overfitting, which results in a poor generalization performance.

2.4.3 Online Metric Learning

As previously stated, a main challenge in linear or nonlinear distance metric learning is to enforce the learned metric to be positive semidefinite (PSD), which turns out to be very computationally expensive in terms of both time and space, especially when dealing with large scale problems. Online learning is contrary very useful in handling these problem by getting rid of the bottleneck of PSD requirements and thus gains great popularity, though it occasionally performs a bit inferior to batch algorithms. Prominent online works can be referred to POLA [171], LEGO [91], OASIS [25], RDML [95] and MDML [111]. POLA [171] is the first online Mahalanobis distance learning approach, which provides a regret bound and is done quite efficiently. LEGO [91] learns metrics in an online setting using a LogDet regularization, and OASIS [25] is a similarity metric learning which scales linearly with the data size through online learning of a bilinear model using a margin criterion and an efficient hinge loss. RDML [95] solves a convex quadratic program in each iteration step instead of doing eigenvalue computation like POLA, and it performs comparably to LMNN and ITML yet much faster. MDML [111] is based on composite mirror descent and can accommodate a large class of loss functions and regularizers for which efficient updates are derived. Besides, both MCML [60] and ITML [39] have online versions with excellent performance.

2.4.4 Local Metric Learning

The previous studies learn a global linear or nonlinear metric, which may incapable to capture the complexity if the data is heterogeneous. However it may beneficial to use local metrics that vary across the space, which have been shown to significantly outperform global methods at the expense of higher time and memory requirements. [211] presents a Local Distance Metric (LDM) that aims to optimize local compactness and local separability in a probabilistic framework. Multiple Metric LMNN (M^2-LMNN) [197, 198] learns several Mahalanobis distances in different parts of the space that are partitioned by clustering algorithms. GLML [113] leverages the power of generative model in the context of metric learning, by locally and simultaneously minimizing the asymptotic probability of misclassification and as well as the bias caused by finite sampling. In [192], PLML is proposed which learns local metrics as linear combinations of basis metrics defined on anchor points over different regions of the instance space, and it is quite robust to overfitting due to its global manifold regularization. Further, [86] extends PLML by regularizing the anchor metrics to be low rank, which allows a better optimization to achieve the optimal metric.

2.4.5 Other Metric Learning

Besides, there are also a few approaches that are outside the scope of the previous categories. For instance, the multi-task metric learning is designed for multi-task setting, where given a set of related tasks a metric is learned for each task in a coupled fashion in order to improve the performance on all tasks. Typical multi-task metric learning algorithms include mt-LMNN [156], MLCS [212], GPML [213] and TML [222]. And as for sparse metric learning, typical examples include LPML [166] and SML [217], which favor the sparsity through L_1

norm and $L_{2,1}$ norm regularization, respectively. However, LPML is not guaranteed to be low rank while SML suffers from the complexity issue in high dimensional problems. Besides, an unified and general framework for sparse metric learning is proposed in [83, 84].

2.5 Image Tagging by Matrix Completion

Literally matrix completion means completing partially specified matrices to fully specified matrices satisfying certain prescribed properties. The matrix completion problem can be dated to back 1990, when Johnson claims in [96] that given a few assumptions about the nature of the matrix, the expected matrix is allowed to be reconstructed. These assumptions include positive semidefinite property, contraction property and given rank assumption [96, 118].

A breakthrough occurs in 2009 when Candès and Recht [20] prove that a low-rank matrix can be reconstructed based on convex optimization of the nuclear norm. Until now, low rank matrix completion has become a recurring problem in many fields, for example, collaborative filtering [62] (notably, the Netflix challenge) and computer vision problems including structure-from-motion [186], multi-classification [1, 15], global positioning [174], among many others. We refer to [19] for a discussion of more applications.

2.5.1 Low Rank Matrix Recovery with Nuclear Norm Minimization

Since finding the lowest rank matrix satisfying the equality constraints is NP-hard [31] and the function of matrix rank is non-convex, a popular approach is to replace it with the nuclear norm, the tightest convex relaxation of matrix rank [19, 21]. The theoretical base for such relaxation is provided in [21, 162] that under favorable conditions, the minimization of the rank function can be achieved by the nuclear norm, which lays the foundation for later matrix completion problem learning. And with the nuclear norm, it is possible to accurately recover a low rank matrix from a small fraction of its entries even if they are corrupted with noise [19, 20, 105].

In the noiseless setting, the matrix completion problem is considered as exact or near-exact recovery, where relevant works [21, 66, 99, 161, 174] discover the minimum required number of random observations to exactly reconstruct a low rank matrix by a constrained nuclear norm minimization. [21, 99, 174] prove that $O\left(nr\mathrm{poly}(\ln n)\right)$ observed samples are required to recover a r-rank $n \times n$ matrix in special case. [66] develops more general methods and improves that result by introducing a *degree of incoherence* ν between the unknown matrix and the basis, and finally indicates that $O(nr\nu \ln^2 n)$ randomly sampled entries is sufficient to recover any low-rank matrix with high probability. And [161] simplifies the previous arguments and sharpens the results of [21, 99, 174] by providing a bound on that number which is optimal up to a small numerical constant and one logarithmic factor. These results thus provide theoretical guarantees for the nuclear norm constrained minimization methods.

In a parallel line of work, noisy matrix completion, which is more common and where a few observed entries are corrupted with noise, has also been extensively studied [19, 53, 54, 100, 102, 103, 104, 105, 107, 150, 165]. The observed noisy matrix is usually regarded as $A = L + S$, where L is an unknown low rank matrix and S corresponds to the noisy corruptions. Compared to the noiseless setting, noise could severely harm the matrix completion results, as shown in [207] that the nuclear norm minimization could fail to recover the low rank matrix even if S contains only a single non-zero column.

When all entries of A are observed, the matrix completion problem becomes a matrix decomposition problem. [23] assumes S is sparse and proves that L and S can be perfectly recovered under additional sufficient identifiability conditions, and milder conditions are further given in [81]. RPCA [18] studies the same model based element-wise sparse S where the corruption positions are sampled uniformly at random, while [207] considers column-wise sparse S, where the uncorrupted columns are chosen uniformly at random and guaranteed to recover as long as L is low rank.

When only partial entries of A are observed, the matrix completion problem is regarded as approximate matrix recovery. It is first systematically addressed in [18] in the noiseless framework with element-wise sparse S, where the corruption positions are sampled uniformly at random. And [29] improves by considering column-wise sparse S, and it proves that the uncorrupted columns of L can be recovered and the corruption positions in S can be identified as well, as long as the following assumptions are satisfied: The uncorrupted columns are chosen uniformly at random, L is low rank, the number of corrupted columns are limited and the number of observed uncorrupted entries are sufficient. Recently, both element-wise and column-wise corruptions are simultaneously addressed in [105], where the high probability recovery of L requires only an upper bound on the maximum of the absolute values of L and S, instead of the rank of L and the sparsity level of S as in previous studies.

In the noisy/approximate matrix recovery setting, most works delve into low rank matrix reconstruction by minimizing the nuclear norm with uniform sampling [19, 54, 100, 102, 107]. Keshavan *et al.* [100] improves over the results of [19] and achieves reconstruction guarantees that are order-optimal in a variety of circumstances. Foygel *et al.* [54] presents reconstruction guarantees based on analysis on the Rademacher complexity of the nuclear norm unit ball [179]. Koltchinskii *et al.* [107] proposes a nuclear norm penalization with fast convergence rate that is shown to be optimal up to logarithmic factors in a minimax sense and is equipped with a non-minimax lower bound. And later, unknown noise variance is focused on in [102], where the author proposes a reconstruction estimator that achieves, up to a logarithmic factor, optimal rates of convergence under the Frobenius risk. And this estimator yields comparable matrix completion performance as the previous studies [19, 107, 150, 165] with known noise deviation.

A common strategy to solve the convex optimization problem is the iterative scheme, and typical algorithms include [16, 138, 185]. Besides, a low complexity algorithm OptSpace based on a combination of spectral techniques and manifold optimization is first introduced by [99] to handle the exact recovery problem, and its robustness to noisy matrix problem setting is theoretically proved in [100]. And other efficient nuclear norm minimization solvers [57, 88, 94, 141] have also been intensively learned. However, most of them fail to solve large scale problems, making nuclear norm regularization less feasible in practice, despite its strong theoretical guarantees. Fortunately, it is recently claimed that large scale matrix completion problem could be solved through a parallel stochastic gradient algorithm [163], or by an efficient nuclear solver via active space selection [80].

The nuclear norm has been applied as a regularizer to image classification [15, 61, 71], visual recovery [136, 149] and tag relevant tasks including image tag refinement [224] and image tag completion [51], where the nuclear norm is used to enforce correlations between classifiers or tags. In the matrix completion/recovery scheme, most studies adopt smooth losses, including common squared loss [54, 61, 184], sparse ℓ_1-norm loss [57, 224], logistic loss [15, 61], maximum margin estimator [136], and even ℓ-Lipschitz loss [53].

2.5.2 Low Rank Recovery under Other Constraints and Sampling Distributions

Most matrix completion works focus on the uniform sampling and the nuclear norm regularization, which might be unrealistic since in practice the observed entries are not guaranteed to follow uniform scheme and its distribution is not known exactly. To overcome this limitation, some researchers search for better sampling distributions [104, 105] while some others develop more suitable surrogate of the matrix rank [53, 54, 150, 17].

In [104, 105], the uniform sampling is replaced by a general and unknown sampling distribution within the nuclear norm minimization framework. Nevertheless, the condition needed in [104] is much milder that it requires only an upper bound on the maximum absolute values of the entries in A, instead of both L and S as done in [105].

It is shown in [178] that the standard nuclear norm might perform poorly, and a common alternative is the empirically weighted nuclear norm [53, 150, 178], which incorporates the prior knowledge of sampling distribution that can be computed based on the locations of the observed entries. Besides, a direct rank penalized estimator, obtained by hard thresholding of the singular values of A, is proposed in [103], where general oracle inequality for the prediction error is established. And in parallel, [165] introduces the Schatten-p norm based penalization and also establishes the prediction error bounds for matrix completion. Recently, a so-called *max norm* [127] recently has been proposed as another convex surrogate to the rank of the matrix, which is defined as

$$|M|_{\max} = \min_{M=UV^\top} |M|_{2,\infty} |V|_{2,\infty},$$

where $|\cdot|_{2,\infty}$ is the maximum ℓ_2 row norm of a matrix. The max norm is first applied to matrix completion under the uniform sampling distribution in [54]. And later a max norm constrained minimization method is proposed in [17] for noisy matrix completion under a general sampling model, which is shown to be minimax rate-optimal and yields a unified and robust approximate recovery guarantee.

Chapter 3

Image Annotation with Kernel Distance Metric Learning

A *Regression based Kernel Metric Learning (RKML)* algorithm is proposed in the image annotation framework in this Chapter.

The remainder of the chapter is organized as follows. Section 3.1 motivates the problem and main intuition behind the proposed algorithm, and as well as setups the notations. Section 3.3 is devoted to the detailed description of the proposed RKML algorithm and its extensions. The theoretical properties and guarantee, *i.e.*, the bounds of error between the computed kernel distance metric and its statistical optimal one, is given in Section 3.4, and the omitted proofs are deferred to Section 3.5. Section 3.6 presents the detailed implementation issues. Section 3.7 describes the intensive experimental setup, results and analysis. Section 3.8 summarizes the chapter and Section 3.2 surveys the closely related works.

3.1 Motivation and Setup

Among the huge volume of image annotation algorithms, the search based approach has been proved to be quite effective, particularly for large image datasets with many keywords [67, 93, 139, 194]. Their key idea is to annotate a test image $\mathcal{I}$ with the common tags shared by the subset of training images that are visually similar to $\mathcal{I}$, which gives rise to an emergent need of an effective visual similarity measure between images. Due to the intricate complexities and nonlinear dependencies between image visual contents, we resort to *Kernel Metric Learning* (KML) [26, 39, 187] to tackle this problem by learning a kernel based distance metric that pulls semantically similar images close and pushes semantically dissimilar images far apart. Figure 3.1 illustrates empirical effects of applying appropriate kernel distance metric to images associated with proper tags, indicating that as two images share more tags, their visual distance is shortened by a learned kernel distance metric.

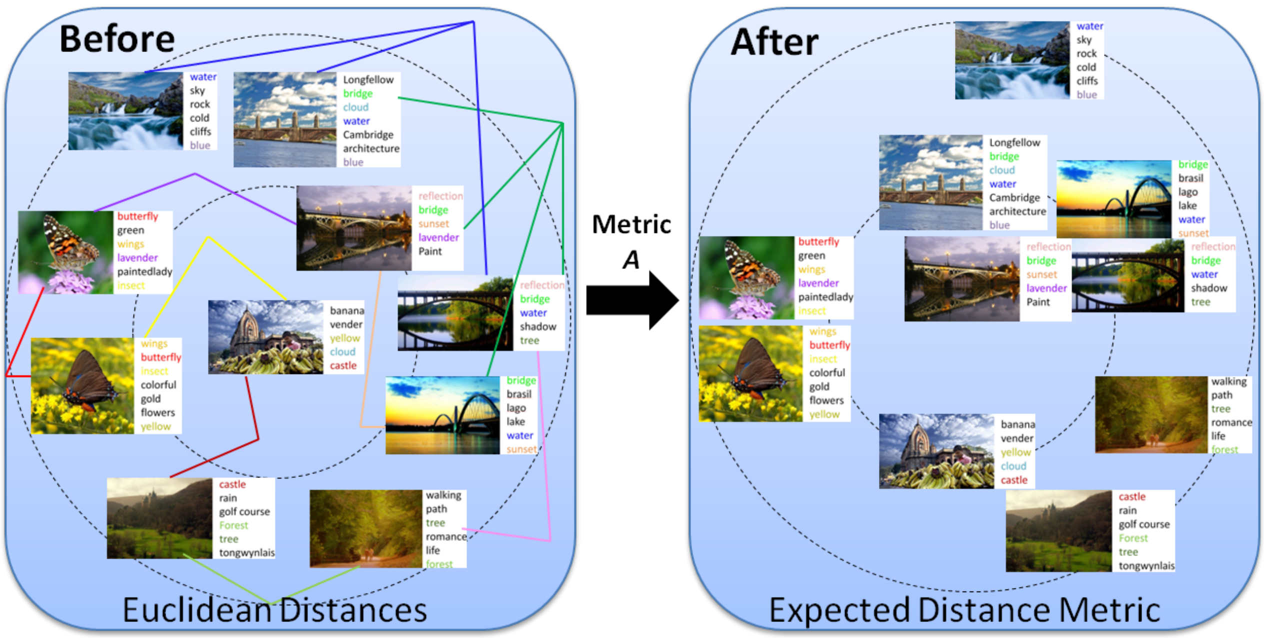


Figure 3.1: Illustration of how kernel distance metric works to images with appropriate tags. In the left box, images share the tags marked in the same color as the lines connecting them.

Kernel metric learning has been widely used to settle image similarity problems in image classification [39, 60, 187], clustering [2, 26, 205], and retrieval [77, 78]. Traditional kernel distance metric learning approaches [39, 60, 78, 187, 205] are usually extended from existing linear distance metric learning, and they usually find the optimal kernel metric by minimizing the distance between must-link images and simultaneously maximizing the distance between cannot-link images.

Despite the success of KML algorithms in those applications, they still suffer from two significant limitations. First, the high dimensionality of KML, denoted by d, usually leads to a high computational cost in solving the related optimization problems. In particular, to ensure the learned metric to be *Positive Semidefinite* (PSD), the existing methods need to project the learned matrix into a PSD cone whose computational cost is $O(d^3)$, which is relatively computationally expensive. Although online learning algorithms [25, 39, 91] are able to get rid of the PSD requirements, they need to train a considerable amount of data and still cost remarkable time to earn a reasonable performance. Secondly, the high dimensionality in kernel metric learning process may lead to the overfitting of training data [95], and finally reduces the annotation performance. To address the over-fitting problem, some studies try to find better kernels with boosting methods [74, 172], some straightforwardly reduce the dimensionality of the projected data [26, 187], and some others directly add a regularizer [95]. However, none of them has a solid theoretic support in dealing with the overfitting problem.

Unlike most linear or kernel metric learning algorithms in similar setup including image classification, clustering and retrieval, which deal with binary semantic constraints (must-link or cannot-link to a label), the proposed RKML algorithm is able to handle the numeric semantic constraints, which better represent the complex semantic relationship between images and thus make better use of the supervised information. Besides, the proposed RKML algorithm avoids the time consuming PSD cone projection step by exploiting the special property of regression, where the PSD property is automatically guaranteed. Additionally the overfitting risk that is easily caused by the high dimensionality and commonly exists in kernel metric learning is alleviated in RKML by appropriately regularizing the rank of the learned kernel metric matrix, instead of an independent norm (Frobenius or Absolute norm) of the learned metric matrix. This strategy also facilitates the further implementation by connecting RKML with the Nyström approximation, and thus speeds up the computation with limited storage memory requested in the computation phase. Finally, the proposed RKML is equipped with theoretical guarantees, the bounds of error between the learned metric and the statistical optimal one, which is original and constructive for kernel distance metric learning.

3.2 Related Work

Due to the rich literature in both areas of image annotation and distance metric learning, here we only survey the studies closely related to this work. For more comprehensive and detailed background review, please refer to Chapter 2.

According to [67], automatic image annotation methods can be categorized into three groups: (i) generative models [22, 49], which are designed to model the joint distribution between tags and visual features, (ii) discriminative models [47, 144] that view image annotation as a classification problems where each keyword is treated as an independent class, and (iii) search based approaches [139, 194]. Recent studies on image annotation show that search based approaches are more effective than both generative and discriminative models. Here, we briefly review the most popular search-based approaches developed for image annotation. TagProp [67] constructs a similarity graph for all images, and propagates the label information via the graph. In [123] a majority voting scheme among the neighboring images is proposed. A sparse coding scheme is proposed in [59] to facilitate label propagation. Conditional Random Field model is adopted in [93] to capture the spatial correlation between annotations of neighboring images.

Many algorithms have been developed to learn a linear distance metric from pairwise constraints [210], and some of them are designed exclusively for image annotation [93, 200, 202]. Recently, a number of nonlinear DML approaches have been developed to handle nonlinear and multimodal patterns. They are usually classified into two categories, boosting based approaches [73, 74, 172] and kernel based approaches, depending on how the nonlinear mapping is constructed. Many KML algorithms, such as Kernel DCA [78], KLMCA [187] and Kernel ITML [39], directly extend their linear counterparts to KML using the kernel trick. To handle the high dimensionality challenge in KML, a common approach is to apply dimensionality reduction before learning the metric [26, 187]. Although these studies show dimensionality reduction helps alleviate the overfitting risk in KML, no theoretical support is provided.

3.3 Annotate Images by Regression based Kernel Metric Learning (RKML)

To begin, let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ be a set of training instances, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional instance. Let m be the number of classes, and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^\top$ be the class assignments of the training instances, where $\mathbf{y}_i \in \{0, 1\}^m$ with $y_{i,j} = 1$ if $\mathbf{x}_i$ is assigned to class j and zero, otherwise. In image annotation, each image can be assigned to multiple classes, and thus each vector $\mathbf{y}_i$ may contain multiple ones. Let

$$\kappa(\mathbf{x}, \mathbf{x}') : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$$

be a kernel function, and $\mathcal{H}_\kappa$ be the corresponding *Reproduced Kernel Hilbert Space*.

Without a metric, the similarity between two instances $\mathbf{x}_a$ and $\mathbf{x}_b$ could be assessed by the kernel function as

$$\langle \kappa(\mathbf{x}_a, \cdot), \kappa(\mathbf{x}_b, \cdot)\rangle_{\mathcal{H}_\kappa} = \kappa(\mathbf{x}_a, \mathbf{x}_b).$$

Similar to linear distance metric learning algorithms, we modify the similarity measure with kernel distance metric as

$$\kappa(\mathbf{x}_a, \mathbf{x}_b) = \langle \kappa(\mathbf{x}_a, \cdot), T[\kappa(\mathbf{x}_b, \cdot)]\rangle_{\mathcal{H}_\kappa},$$

where $T : \mathcal{H}_\kappa \mapsto \mathcal{H}_\kappa$ is a linear operator learned from the training examples. The objective of kernel metric learning is to learn a PSD linear operator T that is consistent with the image tag assignments of training examples. Note that this is different from similarity learning [25] because in distance metric learning we require T to be PSD.

3.3.1 Regression based Kernel Metric Learning

The proposed RKML is a kernel metric learning algorithm based on the regression technique. Let $s_{i,j} \in \mathbb{R}$ be the similarity measure between two images $\mathbf{x}_i$ and $\mathbf{x}_j$ based on their annotations $\mathbf{y}_i$ and $\mathbf{y}_j$. We note that $s_{i,j}$ is a real-valued measurement, which is different from the conventional studies of distance metric learning that only consider a binary relationship between two instances. The discussion of $s_{i,j}$ will be delayed to Section 3.6.1. We adopt a regression model to learn a kernel distance metric consistent with the similarity measure $s_{i,j}$ by solving the optimization problem:

$$\widehat{T} = \arg\min_{T \succeq 0} \sum_{i,j=1}^{n} \frac{1}{2} \left(s_{i,j} - \langle \kappa(\mathbf{x}_i, \cdot), T[\kappa(\mathbf{x}_j, \cdot)]\rangle_{\mathcal{H}_\kappa}\right)^2.$$

Following the representer theorem of kernel learning [170], it is sufficient to assume that $\widehat{T}$ only operates in the subspace spanned by $\kappa(\mathbf{x}_i, \cdot), i = 1, \ldots, n$, leading to the following definition for $\widehat{T}$:

$$\widehat{T}[f](\cdot) = \sum_{i,j=1}^{n} \kappa(\mathbf{x}_i, \cdot) A_{i,j} f(\mathbf{x}_j), \tag{3.1}$$

where $A \in \mathbb{R}^{n \times n}$ is a PSD matrix. Using (3.1), we can change the optimization problem for $\widehat{T}$ into an optimization problem for A as follows:

$$\min_{A \succeq 0} \quad \mathcal{L}(A) = \frac{1}{2} |\mathcal{S} - KAK^{\top}|_F^2, \tag{3.2}$$

where $K = [\kappa(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n}$ is the kernel matrix and $\mathcal{S} = [s_{i,j}]_{n \times n}$ includes all the pairwise semantic similarities between any two training images, and $|\cdot|_F$ represents the Frobenius norm of a matrix.

It is straightforward to verify that

$$A = K^{\dagger} \mathcal{S} K^{\dagger}$$

is an optimal solution to (3.2), where $K^{\dagger}$ stands for the pseudo inverse of K. Note that when the semantic similarity matrix $\mathcal{S}$ is PSD, A will also be PSD, thus no additional projection is needed to enforce the linear operator $\widehat{T}$ to be PSD. To avoid overfitting, we replace K with K_r, the best rank r approximation of K, and express A as

$$A = K_r^{-1} \mathcal{S} K_r^{-1}. \tag{3.3}$$

Evidently, the rank r makes the tradeoff between bias and variance in estimating A: the larger the rank r, the lower the bias and higher the variance. This will become clearer in our theoretical analysis in Section 3.4.

3.3.2 Extension to Image Feature Dimension Reduction

Using the learned linear operator $\widehat{T}$, the similarity between any two data instances $\mathbf{x}_a$ and $\mathbf{x}_b$ is given by

$$\begin{aligned}\kappa(\mathbf{x}_a, \mathbf{x}_b) &= \sum_{i,j=1}^{n} \kappa(\mathbf{x}_a, \mathbf{x}_i)\kappa(\mathbf{x}_b, \mathbf{x}_j)A_{i,j} \\ &= \Phi(\mathbf{x}_a)^\top A \Phi(\mathbf{x}_b) = \left[A^{1/2}\Phi(\mathbf{x}_a)\right]^\top \left[A^{1/2}\Phi(\mathbf{x}_b)\right],\end{aligned}$$

where $\Phi(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}^n$ is given by $\Phi(\mathbf{x}) = [\kappa(\mathbf{x}, \mathbf{x}_1), \ldots, \kappa(\mathbf{x}, \mathbf{x}_n)]^\top$. Thus, the proposed RKML algorithm maps a vector of d dimensions into one with at most m dimensions, *i.e.*, the length of tag dictionary. More justification about the dimension reduction details can be referred to Section 3.6.1.

3.4 Theoretical Guarantee of RKML

We will show that the linear operator learned by the proposed algorithm is stochastically consistent, *i.e.*, the linear operator learned from finite samples provides a good approximation to the optimal one learned from an infinite number of samples. To simplify our analysis, we assume that the semantic similarity measure $s_{i,j} = \mathbf{y}_i^\top \mathbf{y}_j$ [1].

[1] We note that our analysis can be easily extended to the case when $s_{i,j} = \hat{\mathbf{y}}_i^\top \hat{\mathbf{y}}_j$, where $\hat{\mathbf{y}}_i$ is a deterministic transformation of $\mathbf{y}_i$.

Define the optimal linear operator T_* that minimizes the expected loss as follows,

$$\min_{T'} \mathrm{E}_{(\mathbf{x}_a,\mathbf{x}_b,\mathbf{y}_a,\mathbf{y}_b)}\left[\left(\mathbf{y}_a^\top \mathbf{y}_b - \langle \kappa(\mathbf{x}_a,\cdot), T'[\kappa(\mathbf{x}_b,\cdot)]\rangle_{\mathcal{H}_\kappa}\right)^2\right].$$

Let $T_*(r)$ be the best rank-r approximation of T_*, and $\widehat{T}$ be the linear operator constructed by A given in (3.3). We will show that under appropriate conditions,

$$|T_* - \widehat{T}|_*$$

is relatively small, where $|\cdot|_*$ measures the spectral norm.

Let $g_k(\cdot)$ be the prediction function for the k-th class, *i.e.*, $y_{i,k} = g_k(\mathbf{x}_i)$. We make the following assumption for $g_k(\cdot)$ in our analysis:

$$\mathbf{A1}: \quad g_k(\cdot) \in \mathcal{H}_\kappa, \quad k = 1, \ldots, m.$$

Assumption **A1** essentially assumes that it is possible to accurately learn the prediction function $g_k(\cdot)$ given sufficiently large number of training examples. We also note that assumption **A1** holds if $g_k(\cdot)$ is a smooth function and $\kappa(\cdot,\cdot)$ is a universal kernel [146]. The following theorem shows that under assumption **A1**, with a high probability, the difference between T_* and $\widehat{T}$ will be small, provided n is sufficiently large.

Theorem 3.1. *Assume* **A1** *holds, and* $\kappa(\mathbf{x},\mathbf{x}) \le 1$ *for any* $\mathbf{x}$. *Let* $r < n$ *be a fixed rank, and* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues of kernel matrix* K/n *ranked in the descending order.*

For a fixed failure probability $\delta \in (0,1)$, *we assume* n *is large enough such that*

$$\lambda_r \ge \lambda_{r+1} + \frac{8}{\sqrt{n}} \ln(1/\delta). \tag{3.4}$$

Then, with a probability $1-\delta$, we have

$$|\widehat{T} - T_*(r)|_* \leq \varepsilon,$$

where $|\cdot|_*$ is the spectral norm of a linear operator and ε is given by

$$\varepsilon = \frac{8\ln(1/\delta)/\sqrt{n}}{\lambda_r - \lambda_{r+1} - 8\ln(1/\delta)/\sqrt{n}}.$$

The detailed proof of Theorem 3.1 can be found in Section 3.5.1.

3.4.1 Analysis of the Low Rank Approximation Affects to RKML

Using the result from Theorem 3.1, we can analyze how rank r affects $|\widehat{T}-T_*|_*$, the difference between the estimated linear operator and the optimal one represented in spectral norm. We have

$$|\widehat{T} - T_*|_* \leq |\widehat{T} - T_*(r)|_* + |T_* - T_*(r)|_*.$$

As indicated by Theorem 3.1,

$$|\widehat{T} - T_*(r)|_* \leq O\left(\frac{1}{\sqrt{n}(\lambda_r - \lambda_{r+1})}\right),$$

provided

$$\lambda_r \geq \lambda_{r+1} + \frac{16\ln(1/\delta)}{\sqrt{n}}.$$

By choosing a small r, we would expect a large $\lambda_r - \lambda_{r+1}$ and consequentially a small $|\widehat{T} - T_*(r)|_*$, implying a small variance in approximating $T_*(r)$. On the other hand, as the r goes smaller, the $|T_* - T_*(r)|_*$ becomes larger, implying a large bias in approximating T_*.

Thus, rank r essentially makes the tradeoff between the bias and variance in the estimation of the optimal linear operator T_*.

3.5 Proofs of Error Bounds

In this section, we give out the proofs of the main theorems proposed in Section 3.4.

3.5.1 Proof of Theorem 3.1

We here give the sketch of the proof and refer the readers to Section A.1 for more detailed analysis. We first rewrite T into the following form using the expression of A in (3.3)

$$\widehat{T}[f](\cdot) = \sum_{k=1}^{m} \widehat{h}_k(\cdot)\langle \widehat{h}_k(\cdot), f(\cdot)\rangle_{\mathcal{H}_\kappa},$$

where

$$\widehat{h}_k(\cdot) = \sum_{i=1}^{n} \kappa(\mathbf{x}_i, \cdot)[K_r^{-1}\mathbf{y}^k]_i,$$

and $\mathbf{y}^k \in \mathbb{R}^n$ is the k-th column vector of matrix Y.

Using the definition of $g_k(\cdot)$ and assumption **A1**, as well as the reproducing property of kernel function [170], we have

$$y_{i,k} = g_k(\mathbf{x}_i) = \langle g_k(\cdot), \kappa(\mathbf{x}_i, \cdot)\rangle_{\mathcal{H}_\kappa}.$$

Based on these preparations, we develop the following theorem for $\widehat{h}_k(\cdot)$.

Theorem 3.2. *Under assumption* **A1**, *we have*

$$\widehat{h}_k(\cdot) = \sum_{i=1}^{r} \widehat{\varphi}_i(\cdot)\langle \widehat{\varphi}_i(\cdot), g_k(\cdot)\rangle_{\mathcal{H}_\kappa},$$

where $\widehat{\varphi}_i(\cdot), i = 1, \ldots, r$ *are the first* r *eigenfunctions of the linear operator*

$$L_n[f] = \frac{1}{n}\sum_{i=1}^{n} \kappa(\mathbf{x}_i, \cdot) f(\mathbf{x}_i).$$

The proof of Theorem 3.2 can be referred to Section A.1.1.

Using similar analysis as Theorem 3.2, we can express T_* as

$$T_*[f] = \sum_{k=1}^{m} h_k(\cdot)\langle h_k(\cdot), f(\cdot)\rangle,$$

where

$$h_k(\cdot) = \sum_{i=1}^{r} \varphi_i(\cdot)\langle \varphi_i(\cdot), g_k(\cdot)\rangle_{\mathcal{H}_\kappa},$$

and it is the projection of prediction function $g_k(\cdot)$ into the subspace spanned by $\{\varphi_i\}_{i=1}^{r}$. Here $\varphi_i(\cdot)$, $i = 1, \ldots, r$ are the first r eigenfunctions of the integral operator

$$L[f] = \mathrm{E}_{\mathbf{x}}\left[\kappa(\mathbf{x}, \cdot) f(\mathbf{x})\right].$$

Therefore the following theorems bound $|\widehat{T} - T_*|_*$ and $|L - L_n|_*$ by the following two theorems, respectively.

Theorem 3.3. *Let* λ_r *and* λ_{r+1} *be the* r*-th and* $r+1$*-th eigenvalues of kernel matrix* K.

For a fixed failure probability $\delta \in (0, 1)$, assume

$$\frac{\lambda_r - \lambda_{r+1}}{n} > |L - L_n|_*,$$

where $|\cdot|_*$ measures the spectral norm of a linear operator. Then, with a probability $1 - \delta$, we have

$$\max_{f \in \mathcal{H}_\kappa} \quad \|(\widehat{T} - T_*)[f]\|_{\mathcal{H}_\kappa} \leq \gamma \|T_*[f]\|_{\mathcal{H}_\kappa},$$

where γ is given by

$$\gamma = \frac{2\|L - L_n\|_2}{(\lambda_r - \lambda_{r+1})/n - \|L - L_n\|_2}.$$

The proof of Theorem 3.3 can be referred to Section A.1.2.

Theorem 3.4. *[175] Assume $\kappa(\mathbf{x}, \mathbf{x}) \leq 1$. With a probability $1 - \delta$, we have*

$$\|L - L_n\|_{HS} \leq \frac{4\ln(1/\delta)}{\sqrt{n}}.$$

Theorem 3.1 follows immediately from Theorem 3.4 and 3.3.

3.6 Implementation

Regarding implementation, we have two important issues to address: (1) how to appropriately measure the semantic similarity $s_{i,j}$, and (2) how to efficiently compute K_r, the best rank r approximation of K, without computing the full kernel matrix K. The second issue is particularly important for applying the proposed algorithm to large datasets consisted of millions of annotated images. Below, we will discuss these two issues separately.

3.6.1 Computing Semantic Similarity $s_{i,j}$

The most straightforward approach is to measure the semantic similarity as $s_{i,j} = \mathbf{y}_i^\top \mathbf{y}_j$. We improve upon this approach by incorporating the log-entropy weighting scheme [117] which has been used for document retrieval. It computes the weighted class assignment $\tilde{y}_{i,j}$ as

$$\tilde{y}_{i,j} = \left(1 + \sum_{k}^{n} \frac{p_{k,j} \log p_{k,j}}{\log n}\right) \cdot \log(y_{i,j} + 1), \tag{3.5}$$

where $p_{k,j} = y_{k,j} / \sum_i^n y_{i,j}$. We apply Latent Semantic Analysis (LSA) [117] to further enhance the estimation of semantic similarity, which allows us to remove the noise and correlation in/between annotations. Let $\tilde{Y} = [\tilde{y}_{i,j}]_{n \times m}$ include the weighted class assignments for all the training images, and $\hat{Y} \in \mathbb{R}^{n \times m'}$ include the first m' singular vectors of $\tilde{Y}$ with each of its row L_2-normalized by 1. This operation projects $\tilde{Y}$ onto a space of reduced dimensionality m', and this space representation has been empirically shown to capture to some degree the semantic relationship across annotations corpus [147]. We then compute the semantic similarity as

$$\mathcal{S} = \hat{Y}\hat{Y}^\top.$$

3.6.2 Efficiently Computing K_r by Random Projection

The proposed RKML algorithm requires computing the full kernel matrix K and its top r singular vectors. Since the cost of computing K is $O(n^2)$, it will be expensive when the number of training instances n is large. We can improve the computational efficiency by exploiting the Nyström method [43] to approximate K_r. To this end, we randomly sample $n_s < n$ instances from the collection of n training examples, denoted by $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_{n_s}$, then compute the rectangle matrix $K^b \in \mathbb{R}^{n\times n_s}$, and approximate K_r by

$$\tilde{K}_r = K^b[K_r^s]^{-1}[K^b]^\top, \tag{3.6}$$

where K_r^s is the best rank r approximation of $K^s = [\kappa(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j)]_{n_s\times n_s}$, the kernel matrix for the sampled data. According to [32], with a high probability, we have

$$\|\tilde{K}_r - K_r\|_2 \le O\left(\frac{1}{\sqrt{n_s}}\right),$$

implying that $\tilde{K}_r$ is an accurate approximation of K_r provided the number of samples n_s is sufficiently large. This is also supported by our empirical study, *i.e.*, kernel matrix K can be well approximated by the Nytröm method when n_s is a few thousands. According to our implementation, we observe that further approximating K^b in (3.6) to rank r usually yields more accurate prediction for tags. Thus, our final approximation of K_r is given by

$$\hat{K}_r = K_r^b[K_r^s]^{-1}[K_r^b]^\top.$$

3.6.3 Application of RKML to Image Annotation

Given the learned kernel metric $\widehat{T}$, the similarity between the query image $\mathbf{x}$ and the images in gallery $\mathcal{G}$ could be computed as follows:

$$\mathcal{S}(\mathbf{x}, \mathcal{G}) = \sum_{i=1}^{n} \langle \kappa(\mathbf{x}, \cdot), \widehat{T}[\kappa(\mathbf{x}_i^{\mathcal{G}} \cdot)]\rangle = \mathbf{k}_x * A * K_{\mathcal{G}}$$

where $K_{\mathcal{G}} = [\Phi(\mathbf{x}_1), \cdots, \Phi(\mathbf{x}_n)]^\top$, and $\mathbf{k}_x = \Phi(\mathbf{x})$. Consequently we conduct the estimated similarity and obtain the neighbor list of $\mathbf{x}$ as $\mathcal{N}_{\mathbf{x}} = \{\mathbf{x}_i^{\mathcal{N}} | \mathcal{S}(\mathbf{x}, \mathbf{x}_i^{\mathcal{N}}) > \mathcal{S}(\mathbf{x}, \mathbf{x}_j), \forall i \in [1, k], j \in [1, n], \mathbf{x}_j \neq \mathbf{x}_i^{\mathcal{N}}\}$.

Thus the relevance of keywords for $\mathbf{x}$ can be estimated over $\mathcal{N}_{\mathbf{x}}$ by either majority voting, or weighted voting, i.e.,

$$\widehat{\mathbf{y}} = \sum_{i=1}^{k} \langle \kappa(\mathbf{x}, \cdot), \widehat{T}[\kappa(\mathbf{x}_i^{\mathcal{N}} \cdot)] \rangle \mathbf{y}_i^{\mathcal{N}} = \mathbf{k}_x A K_{\mathcal{N}} \tilde{Y}_{\mathcal{N}} \tag{3.7}$$

The keywords with the t-largest relevance scores will be regarded as the annotation for the test image. Algorithm 1 summarizes the key steps of the image annotation algorithm using RKML.

Algorithm 1 Automatic Image Annotation with RKML

Input:

- Training images: $X \in \mathbb{R}^{n \times d}$, labels $Y \in \mathbb{R}^{n \times m}$
- Testing images: $X_q \in \mathbb{R}^{nq \times d}$
- Parameters: smooth parameter γ and approximation rank r.

1: Randomly sample r images $\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_r$ from the training set
2: Compute kernel matrices $K_g = [\kappa(\mathbf{x}_i, \widehat{\mathbf{x}}_j)]_{n \times r}$ and $\widehat{K} = [\kappa(\widehat{\mathbf{x}}_i, \widehat{\mathbf{x}}_j)]_{r \times r}$
3: Compute singular value decomposition: $V \Lambda V^T = K_b$
4: Select the r largest singular values $\lambda_i \in \Lambda_r$ and their corresponding right-singular vectors $u_i \in U_r$
5: Kernel metric: $A = \sum_i^r u_i u_i^T / (\lambda_i^2 + \gamma)$
6: Relevance score matrix: $T_q = K_q A (K_g^T T_g)(T_g^T T_g)$
7: **Output:** Matrix of tag relevance score $T_q \in \mathbb{R}^{n_q \times m}$

Figure 3.2 highlights the key components of a kernel metric learning algorithm based framework for image annotation. It first constructs a *Reproduced Kernel Hilbert Space* (RKHS) $\mathcal{H}$ based on either the whole set or a subset of images that are randomly sampled from the training image set. It then maps the training images to $\mathcal{H}$, and learns a kernel distance metric A from the mapped images. Given a test image $\mathcal{I}$, it first maps $\mathcal{I}$ to $\mathcal{H}$,

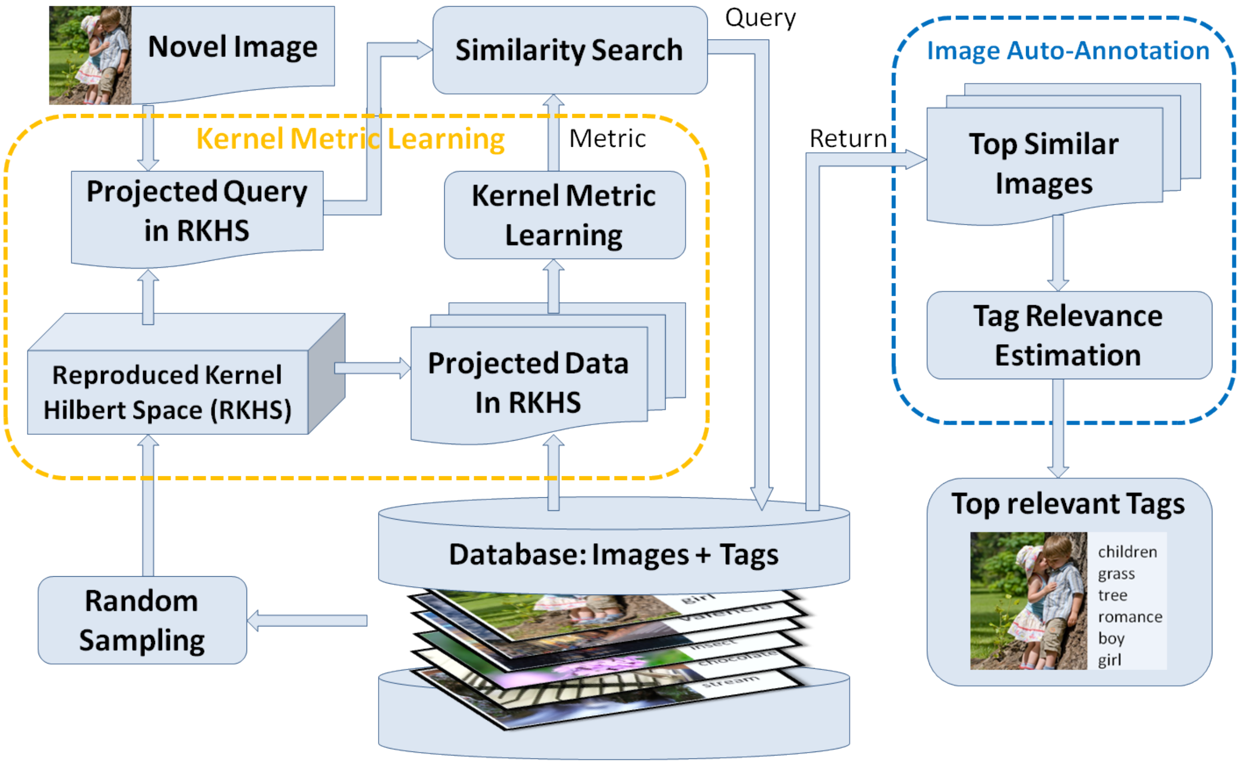


Figure 3.2: The proposed kernel metric learning scheme, *i.e.*, RKML, for automatic image annotation.

and measures the similarities between $\mathcal{I}$ and all the training images in $\mathcal{H}$ using the learned distance metric A. Based on the occurrence of keywords in the subset of training images that share the largest similarities with $\mathcal{I}$, it estimates relevance scores for each keyword and returns the ones with the largest scores as the predicted annotation tags.

3.7 Experiments

3.7.1 Datasets and Experimental Setup

Three benchmark datasets for image annotation are used in our study and their statistics are summarized in Table 3.1. For both ESP Game and IAPR TC12 datasets[2], a bag-of-words

[2]The visual features and tags of both the datasets were obtained from [67] `http://lear.inrialpes.fr/people/guillaumin/data.php`.

model based on densely sampled SIFT descriptors is used to represent the visual content. Flickr1M dataset [202] is comprised of more than one million images crawled from the *Flickr* website that are annotated by more than $700,000$ keywords. Since most keywords are only associated with a small number of images, we only keep the $1,000$ most popular ones. We follow [200, 202] and represent each image with following features: grid color moment, local binary pattern, Gabor wavelet texture, and edge direction histogram.

	ESP Game	IAPR TC12	Flickr1M
No. of Images	20,768	19,627	999,764
Dimensionality	1000	1000	291
Vocabulary size	268	291	1,000
Tags per image	4.69/15	5.72/23	5.98/202
Images per tag	363/5,059	386/5,534	5,976/76,531

Table 3.1: Statistics for the datasets used in the experiments. The bottom two rows are given in the format mean/maximum.

We randomly select 90% of images from each dataset as training and use the remaining 10% for testing. Given a test image, we first identify the k most visually similar images from the training set using the learned distance metric, and then rank the tags by a majority vote over the k nearest neighbors, where k is chosen by cross-validation.

An RBF kernel is used in our study for all KML algorithms. In RKML we set $n_s = 5,000$ and $m' = 0.38m$ based on our experience, and determine the kernel width and rank r by cross-validation. Parameters for the baselines are directly set to their default values suggested by the original authors. Besides, annotation based on the Euclidean distance, denoted by *Euclid*, is used as a reference in our comparison. Since most DMLs are developed against must-links and cannot-links, we apply the procedure described in [200] to generate the binary constraints by performing a probabilistic clustering over the images based on their tags. More details of this procedure can be found in [200].

We evaluate the annotation accuracy by the average precision for the top ranked image tags. Following [201, 202], we first compute the precision for each test image by comparing the top 10 annotated tags with the ground truth, and then take the average over the test set. Average recall and F1 score are reported in the supplementary document. The computational efficiency is measured by the running time[3]. Both the mean and standard deviation of evaluation metrics over 20 experimental trials are reported in this paper.

3.7.2 Comparison with State-of-the-art distance metric learning (DML) and Image Annotation Algorithms

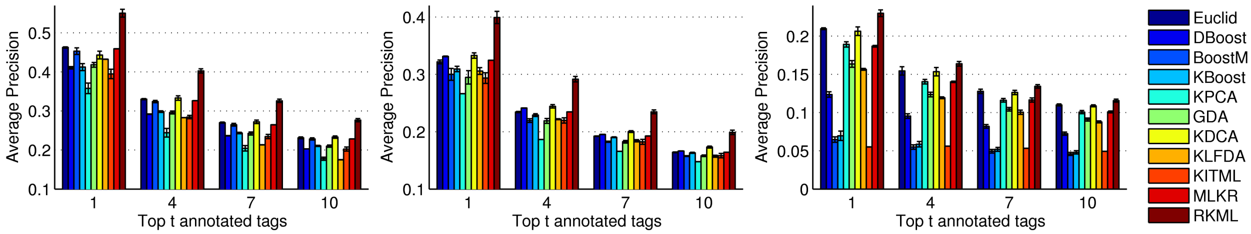
3.7.2.1 Comparison to nonlinear DML algorithms

We first compare the proposed RKML[4] algorithm to six state-of-the-art **kernel** distance metric learning methods: (1) Kernel PCA (*KPCA*) [169], (2) Generalized discriminant analysis (*GDA*) [6], (3) Kernel discriminative component analysis (*KDCA*) [78], (4) Kernel local Fisher discriminant analysis (*KLFDA*) [182], (5) Kernel information theoretic based metric learning (*KITML*) [39], and (6) Metric learning for kernel regression (*MLKR*) [199]. We also include three boosting DML algorithms, *i.e.*, Distance Boost (*DBoost*) [73], Kernel Boost (*KBoost*) [74], and metric learning with boosting (*BoostM*) [172], for comparison.

Figure 3.3, 3.4 and 3.5 show the average precision, average recall and average F1 score, respectively, of the top t annotated tags obtained by nonlinear distance metric learning (DML) baselines and the proposed RKML. Surprisingly, we observe that most of the nonlinear DML algorithms are only able to yield performance similar to that based on the Euclidean

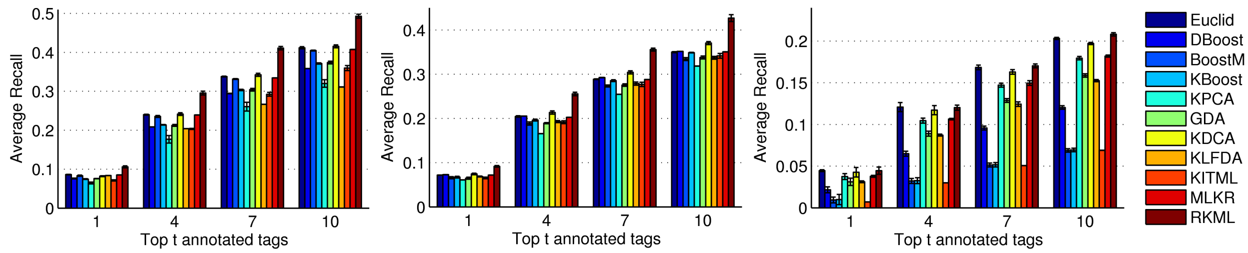
[3]All the codes are downloaded from the authors' websites, and run in Matlab on the AMD 2 core @2.7GHz and 64 GB RAM machine.

[4]Without specific notification, RKML stands for the proposed RKML algorithm with Nyström approximation. And its source code can be found in `http://www.cse.msu.edu/~fengzhey/research/rkml.html`.



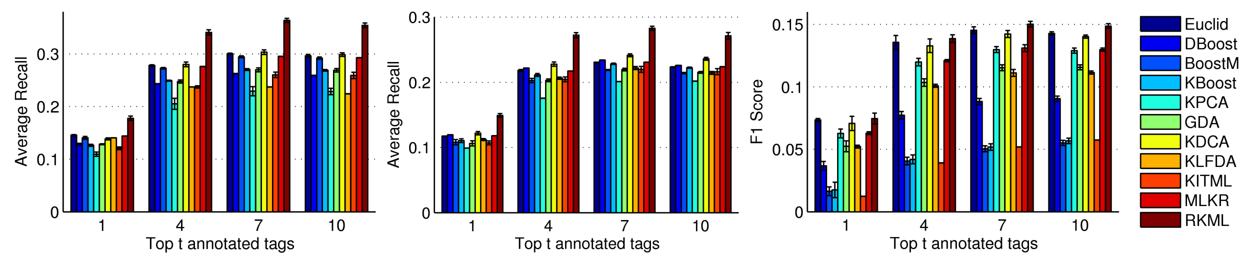
(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.3: Average precision for the top t annotated tags using nonlinear distance metrics.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.4: Average recall for the top t annotated tags using nonlinear distance metrics.



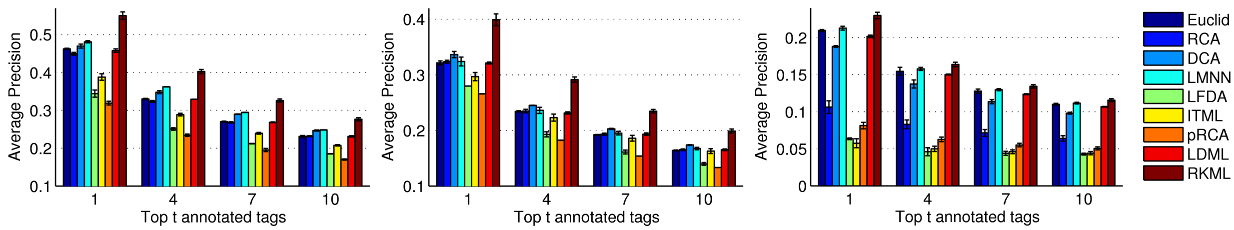
(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.5: Average F1 score for the top t annotated tags using nonlinear distance metrics.

distance, and more disturbingly, some of the nonlinear DML algorithms even perform significantly worse than the Euclidean distance. On the other hand, the proposed algorithm performs significantly better than the Euclidean distance for almost all cases. Relevant analysis of this phenomena is provided in Section 3.7.3.3.

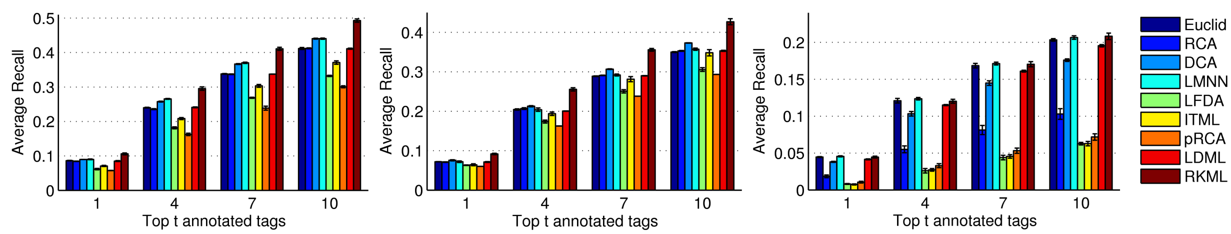
3.7.2.2 Comparison to linear DML algorithms

We compare our RKML to seven state-of-the-art **linear** distance metric learning algorithms, including Relevant component analysis (RCA) [3], Discriminative component analysis (DCA) [78], Large margin nearest neighbor classifier ($LMNN$) [196], Local Fisher discriminant analysis ($LFDA$) [182], Information theoretic based metric learning ($ITML$) [39], Probabilistic RCA ($pRCA$) [200], and Logistic discriminant-based metric learning ($LDML$) [68].



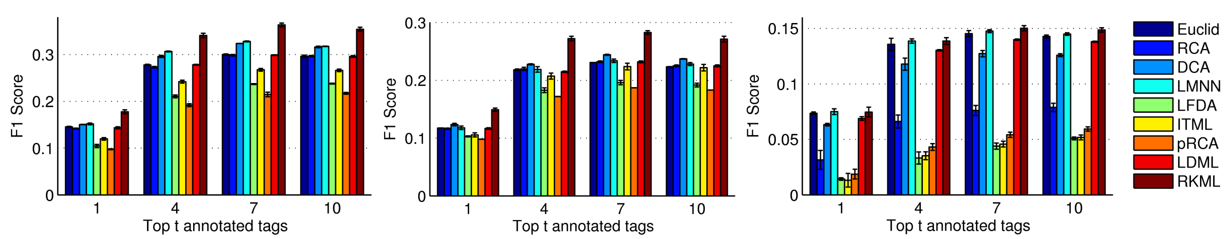
(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.6: Average precision for the top t annotated tags using linear distance metrics.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.7: Average recall for the top t annotated tags using linear distance metrics.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

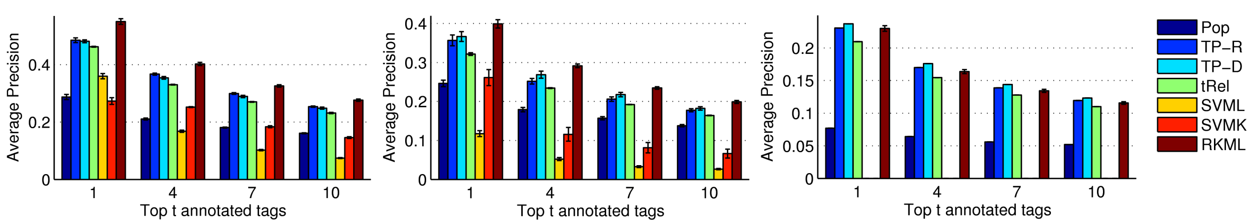
Figure 3.8: Average F1 score for the top t annotated tags using linear distance metrics.

Figure 3.6, 3.7 and 3.8 show the average annotation precision, average recall and average

F1 score, respectively, for the linear distance metric learning (DML) baselines. Similar to KML, we observe that even the best linear DML algorithm is only slightly better than the Euclidean distance, while RKML significantly outperforms all linear DML baselines. Again, we believe that the failure of linear DML is likely due to the binary constraints generated from image annotations, which is explained in Section 3.7.3.3.

3.7.2.3 Comparison with State-of-the-art Image Annotation Methods

Additionally, we compare RKML algorithm to several state-of-the-art image annotation models including: (1) Two versions of the TagProp method [67], using either rank-based weights (*TP-R*) or distance-based weights (*TP-D*), (2) TagRelevance (*tRel*) [123] based on the idea of neighbor voting, (3) 1-vs-1 SVM classification, using either linear (*SVML*) or RBF kernel (*SVMK*) classifiers[5]. We include *Pop* as a comparison reference which simply ranks tags based on their occurring frequency in the training set.

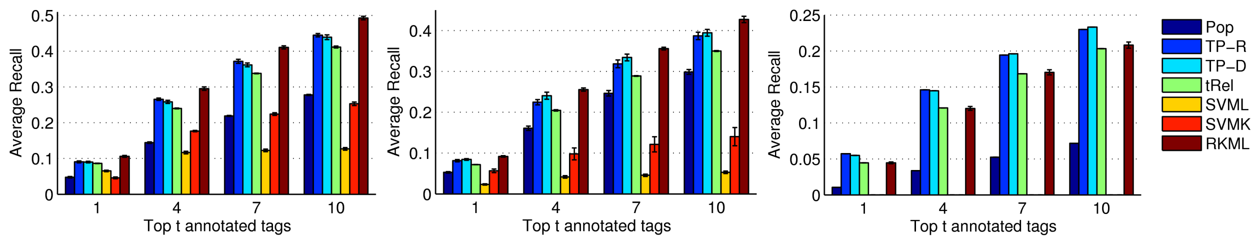


(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.9: Annotation performance in terms of *AP@t* with different annotation models.

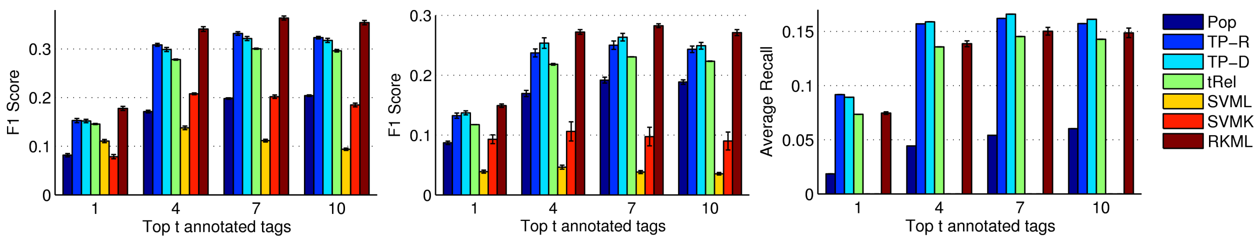
Figure 3.9, 3.10 and 3.11 show the comparison of average precision, average recall and average F1 score that obtained by different image annotation models, respectively. It is not surprising to observe that most annotation methods significantly outperform Pop, while the proposed RMKL method outperforms all the state-of-the-art image annotation methods on

[5]SVM methods were unable to perform over *Flickr 1M* due to its large size and high computational cost, and the results of SVM methods are excluded.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.10: Average recall for the top t annotated tags using different annotation models.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.11: Average F1 score for the top t annotated tags using different annotation models.

IAPR TC12 and ESP Game datasets, and only performs slightly worse than TP-D on the Flickr 1M dataset.

3.7.2.4 Comparison of Annotation Results on Exemplar Images

In order to straightforwardly compare the empirical performance of image annotation between various linear and kernel distance metric learning algorithms as well as image annotation approaches, we include the comparison of annotation results on certain images in Table 3.2, which shows the annotations of exemplar images by different DML and image annotation algorithms.

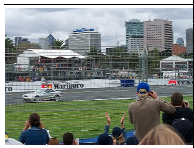
	Ground	Euclid	DCA	LMNN	LDML	DBoost	BoostM	KBoost	KLDA	KPCA	KDCA	KLFDA	MLKR	TP-R	TP-D	RKML
	fog	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**	tree	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**	**mountain**
	front	**wall**	**wall**	**wall**	**wall**	**wall**	**tourist**	**wall**	**range**	sky	**wall**	**range**	**wall**	man	man	**terrace**
	mountain	**terrace**	tree	**fog**	**terrace**	**terrace**	**front**	**terrace**	**wall**	**front**	**terrace**	**terrace**	**terrace**	**fog**	**wall**	**wall**
	range	**fog**	cloud	**terrace**	woman	cloud	**wall**	cloud	cloud	man	**fog**	**wall**	cloud	**wall**	tree	**range**
	ruin	**range**	**terrace**	**tourist**	cloud	**fog**	woman	**range**	sky	**wall**	**range**	grey	man	tree	people	**ruin**
	terrace	cloud	**tourist**	woman	**range**	**range**	group	sky	**terrace**	house	cloud	**ruin**	woman	**terrace**	woman	**fog**
	tourist	tree	**fog**	tree	**fog**	**ruin**	**range**	tree	**fog**	**mountain**	hill	sky	**range**	slope	**front**	slope
	wall	**ruin**	people	forest	man	sky	man	man	people	people	man	**fog**	tree	summit	**tourist**	sky
	building	**sky**	**sky**	**sky**	**sky**	**sky**	people	**sky**	**sky**	**sky**	**sky**	**sky**	**sky**	**sky**	**sky**	**meadow**
	front	**tree**	**tree**	**tree**	**tree**	sea	**sky**	**tree**	cloud	**tree**	**tree**	**tree**	**tree**	**tree**	**tree**	**sky**
	hill	cloud	**building**	**meadow**	cloud	cloud	man	cloud	house	**wall**	sea	**building**	cloud	**wall**	house	**tree**
	meadow	**building**	man	**building**	**building**	beach	mountain	**building**	**tree**	**front**	beach	people	mountain	**ruin**	cloud	**building**
	ruin	house	house	cloud	house	rock	**tree**	bush	**hill**	mountain	bush	square	house	slope	**front**	**hill**
	sky	**hill**	**front**	bush	people	**meadow**	**front**	**meadow**	mountain	people	cloud	column	man	**meadow**	**meadow**	**wall**
	tree	people	cloud	landscape	bush	**tree**	bush	sea	**building**	man	**meadow**	flag	**meadow**	**building**	man	terrace
	wall	bush	**meadow**	**ruin**	**hill**	coast	rock	house	sea	house	house	**front**	people	house	people	**front**
	bike	**road**	man	sky	**road**	tree	man	sky	tree	tree	sky	sky	**road**	**landscape**	tree	**road**
	cycling	man	wall	bush	man	sky	sky	snow	sky	sky	snow	tree	man	man	sky	sky
	cyclist	**cyclist**	desert	man	**cyclist**	**short**	**rock**	tree	meadow	front	**cycling**	**landscape**	**cyclist**	**grass**	man	**landscape**
	helmet	**jersey**	front	**road**	**helmet**	**jersey**	people	building	man	wall	**cyclist**	**rock**	**helmet**	sea	front	**cyclist**
	jersey	**short**	sky	tree	**jersey**	meadow	tree	front	**cyclist**	man	man	bush	**jersey**	tree	**road**	**short**
	landscape	**bike**	floor	**bike**	**short**	sock	bush	people	**landscape**	people	**short**	building	**short**	cactus	wall	**bike**
	mountain	**cycling**	**road**	car	**cycling**	lawn	**landscape**	cloud	**road**	house	**bike**	cloud	sky	**road**	bush	**cycling**
	road	**helmet**	tree	**cycling**	sky	man	cliff	man	rock	**mountain**	front	front	**cycling**	sky	meadow	**jersey**
	short	car	tourist	**cyclist**	**bike**	spectator	front	street	cloud	woman	**helmet**	**grass**	**bike**	**rock**	people	**helmet**
	door	building	**wall**	building	building	front	building	**house**	building	**sky**	front	**house**	building	**house**	**house**	**door**
	house	front	table	street	table	**house**	**tree**	building	front	**tree**	**house**	**sky**	**house**	**window**	**window**	**house**
	palm	**house**	woman	balcony	**house**	**window**	**sky**	**window**	**window**	front	building	**tree**	table	street	street	**sky**
	roof	table	front	people	front	building	**house**	front	**house**	people	**window**	hill	**wall**	**sky**	**sky**	**window**
	sky	**window**	**window**	square	**wall**	**wall**	street	**door**	**sky**	**house**	**door**	landscape	front	**door**	**tree**	**palm**
	tree	square	classroom	**tree**	woman	**sky**	people	balcony	**wall**	man	**wall**	meadow	man	**tree**	**door**	**tree**
	window	woman	man	**window**	man	column	tower	entrance	**door**	mountain	flag	**roof**	**window**	**palm**	**palm**	building
	wall	**door**	building	front	square	entrance	car	**wall**	column	building	man	snow	woman	man	tile	street
	car	**sky**	**tree**	**tree**	people	**sky**	**sky**	man	**sky**	**sky**	people	fog	people	**tree**	people	**sky**
	fence	people	building	building	**sky**	front	building	sea	people	**tree**	man	**sky**	**sky**	**sky**	**tree**	**spectator**
	grandstand	**tree**	front	**sky**	**tree**	building	**tree**	**sky**	cloud	wall	**tree**	wall	**tree**	front	**sky**	**tree**
	house	man	people	front	**house**	people	people	woman	boat	front	**sky**	man	man	building	man	**fence**
	sky	woman	man	**car**	front	square	man	**tree**	man	man	**fence**	mountain	front	cloud	front	front
	palm	**house**	**sky**	**house**	man	tower	**house**	beach	sea	cloud	woman	slope	**house**	river	**house**	**car**
	spectator	**car**	**car**	**meadow**	square	**tree**	front	cloud	**tree**	mountain	bank	beach	woman	boat	building	**grandstand**
	tree	building	**fence**	**palm**	woman	man	**car**	water	beach	**house**	**car**	bed	square	people	woman	people
	bed	**wall**	**wall**	**wall**	**wall**	**bed**	**wall**	**wall**	**wall**	sky	**wall**	**bed**	**wall**	woman	**wall**	**wall**
	blanket	table	table	woman	table	**wall**	**room**	**room**	**window**	tree	**room**	**wall**	table	**wall**	woman	**bed**
	curtain	**room**	**room**	door	**room**	**room**	**bed**	**bed**	**room**	mountain	**bed**	**room**	**room**	**front**	table	**room**
	front	**window**	woman	table	**front**	**curtain**	table	table	**bed**	cloud	table	table	**front**	door	man	**window**
	room	**curtain**	**front**	man	**window**	**window**	**window**	**window**	**curtain**	**wall**	**curtain**	**wood**	man	man	**room**	**curtain**
	wall	woman	**window**	**room**	**bed**	table	**wood**	**wood**	table	rock	**window**	bedcover	woman	table	**front**	**wood**
	window	**bed**	**bed**	**bed**	woman	**wood**	**curtain**	**curtain**	**wood**	**front**	**wood**	**bedside**	**window**	house	**window**	table
	wood	door	door	building	**curtain**	lamp	lamp	door	**front**	house	door	**curtain**	**bed**	**room**	**bed**	**front**
	building	**sky**	**tree**	**tree**	**sky**	**sky**	**sky**	**sky**	**sky**	**tree**	**sky**	**sky**	**sky**	**sky**	**sky**	**sky**
	cloud	**cloud**	man	**road**	**front**	**cloud**	**tree**	**tree**	**cloud**	man	**tree**	mountain	**front**	**cloud**	**tree**	**tree**
	front	**front**	car	**front**	**cloud**	sea	mountain	**building**	**tree**	wall	cyclist	**tree**	**tree**	**front**	**cloud**	**cloud**
	hill	**tree**	cyclist	man	**tree**	man	**hill**	sea	mountain	**sky**	**front**	desert	**cloud**	**tree**	man	**building**
	meadow	man	cycling	mountain	**road**	landscape	tourist	beach	beach	woman	man	grey	**road**	car	**front**	**meadow**
	monument	**road**	short	**sky**	man	**meadow**	**front**	house	man	**front**	mountain	**hill**	man	park	mountain	**hill**
	sky	mountain	**building**	car	mountain	beach	house	**front**	house	house	people	landscape	mountain	man	**road**	mountain
	tree	car	**sky**	**cloud**	people	**tree**	landscape	city	**meadow**	mountain	**road**	snow	**hill**	shop	house	**front**

Table 3.2: Examples of annotation results generated by 14 baselines and the proposed RKML. The annotated tags are ranked based on the estimated relevance score in descending order, and the correct ones are highlighted in blue bold font. Note the ground truth annotations in the 2-nd column do not always include all relevant tags (*e.g.*, "people" for the 5-th image), and sometimes contain polysemes (*e.g.*, "palm" for the 4-th and 5-th images) and controversial tags (*e.g.*, "front").

3.7.2.5 Efficiency Evaluation

TIME	DCA	LMNN	ITML	LDML	DBoost	BoostM	RKML
IAPR TC12	1.5e4	1.4e4	4.2e4	4.2e5	1.7e4	1.1e6	4.6e2
ESP Game	2.3e4	1.7e4	5.8e4	5.5e5	4.3e4	1.2e6	1.3e3
Flickr 1M	8.1e4	6.0e4	3.0e4	5.2e5	1.2e4	3.2e5	3.4e3
TIME	KPCA	GDA	KDCA	KLFDA	KITML	MLKR	RKML
IAPR TC12	2.8e4	4.8e4	2.2e4	8.8e4	5.3e4	2.2e3	4.6e2
ESP Game	3.3e4	5.4e4	3.7e4	3.2e5	6.8e4	3.5e4	1.3e3
Flickr 1M	7.3e3	3.3e4	1.3e5	1.0e5	3.7e6	7.9e3	3.4e3

Table 3.3: Comparison of running time (s) for several different metric learning algorithms.

TIME	TP-R	TP-D	tRel	SVML	SVMK	RKML
IAPR TC12	9.1e2	4.6e2	1.0e1	2.5e3	4.0e5	4.8e2
ESP Game	2.7e2	1.5e2	1.5e1	1.6e2	8.9e4	1.3e3
Flickr 1M	1.6e5	9.9e4	5.7e3	-	-	3.4e3

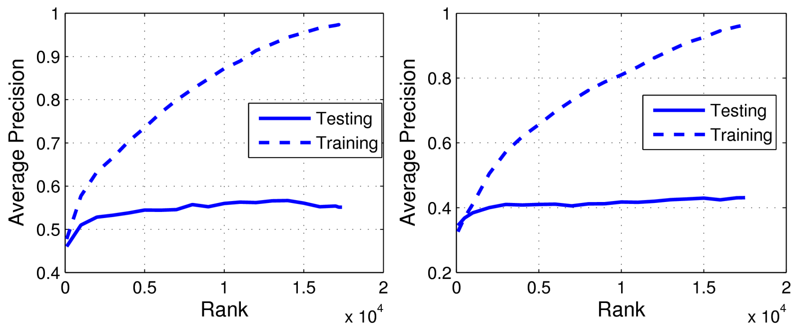
Table 3.4: Running time (s) for image annotation. SVM methods Flickr 1M are not included due to their high computational costs.

Table 3.3 summarizes the running time of different DML algorithms. We observe that RKML is significantly more efficient than any DML baseline. Table 3.4 compares the efficiency of different baselines for annotation, where the running time includes the time for both learning a distance metric and predicting image tags. We observe that compared to the other annotation methods, the proposed RKML algorithm is particularly efficient for large datasets (*i.e.*, Flickr 1M), making it suitable for large-scale image annotation.

3.7.3 Affects of Different Experimental and Parameter Setup

3.7.3.1 Sensitivity to Parameters

In this section, we analyze the sensitivity to parameters in RKML, including rank r, m', the number of retained eigenvectors when estimating the semantic similarity, and n_s, the number of sampled images used for Nyström approximation.

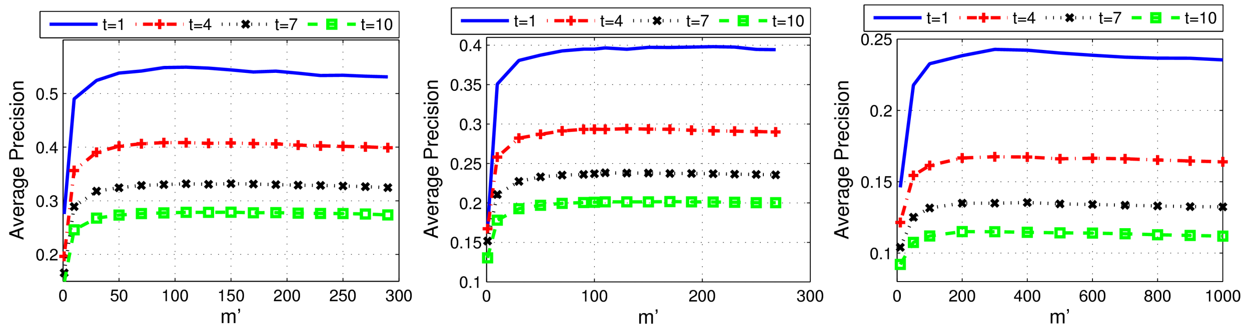


(a) *IAPR TC12* (b) *ESP Game*

Figure 3.12: Average Precision for the first tag predicted by RKML using different values of rank r. To make the overfitting effect clearer, we turn off the Nyström approximation for IAPR TC12 and ESP Game datasets. Flickr 1M dataset is not included due to its large size ($n = 999,764$). The overfitting only occurs when r approximates to the total number of images, but it is infeasible to apply such a large r in Flickr 1M dataset.

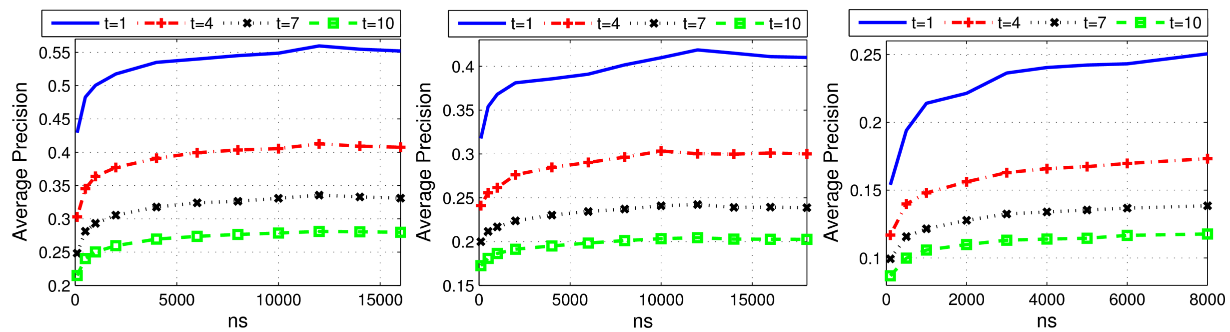
We examine the role of rank r in the proposed algorithm by evaluating the prediction accuracy with varied r on the IAPRTC 12 and ESP Game datasets for both training and testing images. To make it clear, we turn off the Nyström approximation used by RMKL in this experiment. We observe in Figure 3.12 that while the average accuracy of test images initially improves significantly with increasing rank r, it becomes saturated after certain rank. On the other hand, the prediction accuracy of training data increases almost linearly with respect to the rank, and becomes almost 1 for very large r, a clear indication of overfitting training data.

We also examine the sensitivity of the other parameters used by the proposed RKML algorithm, including m', the number of retained eigenvectors of $\tilde{Y}$, and n_s, the number of sampled images used for Nyström approximation). Figure 3.13 and 3.14 show the image annotation performance in terms of varied m' and n_s, respectively. Overall, we found that our algorithm is insensitive to the values of these parameters over a wide range, which facilitate the selection of these parameters in real-world application.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.13: Average Precision for the top t tags predicted by RKML using different values of m', the number of retained eigenvectors when estimating the semantic similarity.



(a) *IAPR TC12* (b) *ESP Game* (c) *Flickr 1M*

Figure 3.14: Average Precision for the top t tags predicted by RKML using different values of n_s, the number of sampled images used for Nyström approximation. In (c), n_s couldn't be set too large due to the dataset size.

3.7.3.2 Advantages of Kernel Trick and Nyström Approximation

Since none of the baseline algorithms, neither linear nor nonlinear DML, is able to significantly outperform the Euclidean distance, it remains unclear if kernel DML is advantageous to a linear DML. To examine this point, we implement the linear version of RKML, denoted by RLML. Table 3.5, 3.6 and 3.7 show the comparison of performance between RLML and RKML on three datasets. It is clear that RKML significantly outperforms its linear counterpart RLML, verifying the advantage of using kernel trick in distance metric learning.

However, although the kernel trick considerably improves the image annotation accuracy,

AP@t(%)	t=1	t=2	t=3	t=4	t=5	t=6	t=8	t=10
RKML	55 ± 1.2	48 ± 0.9	44 ± 0.6	41 ± 0.8	37 ± 0.6	35 ± 0.5	31 ± 0.5	28 ± 0.4
RLML	52 ± 1.3	46 ± 1.2	42 ± 1.0	38 ± 0.8	35 ± 0.7	33 ± 0.6	29 ± 0.5	26 ± 0.4
RKMLO	57 ± 0.9	51 ± 0.6	46 ± 0.7	43 ± 0.6	39 ± 0.6	37 ± 0.5	32 ± 0.5	29 ± 0.4
RKMLH	49 ± 1.1	44 ± 0.9	39 ± 0.9	36 ± 0.7	33 ± 0.7	31 ± 0.7	27 ± 0.6	24 ± 0.5

Table 3.5: Comparison of various extensions of RKML in terms of $AP@t$ on IAPR TC12 dataset. RLML is the linear version of RKML, RKMLO is the original version without Nyström approximation, and RKMLH runs RKML using binary constraints.

AP@t(%)	t=1	t=2	t=3	t=4	t=5	t=6	t=8	t=10
RKML	40 ± 1.1	35 ± 0.5	32 ± 0.4	29 ± 0.5	27 ± 0.4	25 ± 0.4	22 ± 0.4	20 ± 0.4
RLML	36 ± 0.8	31 ± 0.7	28 ± 0.7	26 ± 0.7	24 ± 0.5	22 ± 0.4	20 ± 0.4	18 ± 0.4
RKMLO	44 ± 0.8	39 ± 0.6	35 ± 0.5	32 ± 0.4	29 ± 0.4	27 ± 0.4	24 ± 0.3	21 ± 0.3
RKMLH	34 ± 1.0	30 ± 0.5	28 ± 0.5	26 ± 0.4	24 ± 0.4	22 ± 0.3	20 ± 0.3	18 ± 0.3

Table 3.6: Comparison of the extensions of RKML in terms of $AP@t$ on ESP Game dataset.

AP@t(%)	t=1	t=2	t=3	t=4	t=5	t=6	t=8	t=10
RKML	24 ± 0.1	21 ± 0.2	18 ± 0.1	17 ± 0.2	15 ± 0.2	14 ± 0.1	13 ± 0.2	12 ± 0.1
RLML	13 ± 0.3	12 ± 0.2	11 ± 0.2	11 ± 0.1	10 ± 0.06	10 ± 0.05	9.0 ± 0.05	8.0 ± 0.08
RKMLH	20 ± 0.2	18 ± 0.1	16 ± 0.2	15 ± 0.2	14 ± 0.2	13 ± 0.1	11 ± 0.1	10 ± 0.1

Table 3.7: Comparison of the extensions of RKML in terms of $AP@t$ on Flickr 1M dataset. RKMLO is excluded since the dataset is too large to do the computation on the full kernel.

it also inevitably leads to high even prohibitive computational cost. So the Nyström approximation is proposed to solve this problem, which makes a trade-off between the computational cost and annotation accuracy. To verify the effectiveness of the Nyström approximation, we implement the RKML by turning off the Nyström approximation and directly do all computation on the full kernel, and this method is denoted by RKMLO. Table 3.5, 3.6 and 3.7 compare the annotation performance of RKML and RKMLO, where we observe that RKML performs slightly worse than RKMLO. This phenomenon indicates that RKML makes a good compromise between the effectiveness and computational cost, by making tolerant sacrifice on annotation effectiveness to get rid of the great computational burden.

3.7.3.3 Analysis on Binary Constraints and Their Various Generation Ways

We observe in Section 3.7.2 that most baseline metric learning algorithms, either linear or kernel ones, perform worse than the Euclidean distance. We attribute this failure mostly to the binary constraints. As described before, all baseline distance metric learning algorithms require converting image annotations into binary constraints in image annotation tasks, which does not make full use of the annotation information. To verify this point, we run RKML with similarity measure $s_{i,j}$ computed from the binary constraints that are generated for the baseline distance metric learning algorithms, and denote this method by RKMLH. We observe in Table 3.5, 3.6 and 3.7 that RKMLH performs significantly worse than RMKL which directly uses the real-valued similarity measures, confirming the significance of using real-valued similarities for distance metric learning in automatic image annotation.

AP@t(%)	t=1	t=4	t=7	t=10
Method 1	20.7 ± 0.2	15.3 ± 0.2	12.4 ± 0.12	10.6 ± 0.10
Method 2	20.6 ± 0.3	15.2 ± 0.2	12.4 ± 0.11	10.6 ± 0.09
Method 3	20.8 ± 0.2	15.4 ± 0.1	12.5 ± 0.05	10.7 ± 0.04
Method 4	19.7 ± 0.2	14.6 ± 0.1	11.9 ± 0.06	10.2 ± 0.06
Method 5	21.3 ± 0.4	15.9 ± 0.3	12.8 ± 0.20	11.0 ± 0.14

Table 3.8: Comparison of different methods of generating binary constraints that are applied in baseline distance metric learning algorithm LMNN for the top t annotated tags on the Flickr1M dataset. Method 1 clusters the space of keywords, method 2 considers the class assignments as binary constraints, method 3 clusters the space of keywords using hierarchical clustering algorithms, method 4 clusters the space of keywords together with the visual features, and method 5 considers images sharing more than 4 keywords as similar and images sharing no keyword as dissimilar.

Since most DML algorithms were designed for binary constraints, we tried to improve the performance of standard DML algorithms by experimenting with different methods for generating binary constraints. They are listed as follows: (1) Clustering the space of keywords, (2) Generating binary constraints from classification labels[6], (3) Clustering the space of keywords using hierarchical clustering algorithms, (4) Clustering the space of keywords together with the visual features, and (5) Generating binary constraints based on the number of common keywords, *i.e.*, images sharing more than 4 keywords are considered as similar and images sharing no keywords are considered as dissimilar. Note the last one is applicable in LMNN, but not applicable in many other DML algorithms. For example, RCA [3] and DCA [78] divide image set into groups where images within a group are considered as similar and images from different groups are considered as dissimilar; but this method is not able to generate such groups. We observe that these methods yield essentially the same performance reported in our study, as shown in Table 3.8.

3.7.3.4 Comparison of the Design Choices of Semantic Similarity Measure

To obtain the numeric constraints on the annotated tag, besides log-entropy, we further explore other weighting schemes. And besides clustering using a topic model, we also experiment other binary constraint generation methods.

Binary	$l_{i,j} = 1$ if tag i exists in image j, or else 0.
Term Frequency (TF)	$l_{i,j} = tf_{i,j}$, the occurrences counts of tag j in image i.
Log	$l_{i,j} = \log(tf_{i,j} + 1)$

Table 3.9: Local weighting functions.

We examine the choice of semantic similarity by evaluating the prediction accuracy with varied definition of $\tilde{y}_{i,j}$ in Equation (5). $\tilde{y}_{i,j}$ is actually the product of a local tag weight $l_{i,j}$ that describes the relative occurrence of tag j in image i , and a global weight g_j that

[6] Flickr1M dataset also includes class assignment labels which is usually used for classification. ESP Game and IAPR TC12 do not have classification labels.

Binary	$g_j = 1$
Normal	$g_j = 1/\sqrt{\sum_i^n tf_{i,j}^2}$
Idf	$g_j = \log_2 \frac{n}{1+df_j}$
Entropy	$g_j = 1 + \sum_i^n \frac{p_{i,j} \log p_{i,j}}{\log n}$, where $p_{i,j} = \frac{tf_{i,j}}{\sum_i^n tf_{i,j}}$

Table 3.10: Global weighting functions.

describes the relative occurrence of tag j within the entire tag collection. The examined weighting functions [10] are defined as follows in Table 3.9 and 3.10.

AP@t(%)	t=1	t=4	t=7	t=10
Binary-Binary	56 ± 1.01	41 ± 0.57	33 ± 0.49	28 ± 0.45
Binary-Normal	53 ± 1.28	39 ± 0.62	32 ± 0.54	28 ± 0.44
Cosine	56 ± 1.19	41 ± 0.61	33 ± 0.52	28 ± 0.47
TF-IDF	55 ± 1.12	41 ± 0.57	33 ± 0.50	28 ± 0.44
Log-IDF	55 ± 1.12	41 ± 0.57	33 ± 0.50	28 ± 0.44
Log-Entropy	55 ± 1.10	41 ± 0.57	33 ± 0.49	28 ± 0.45

Table 3.11: Comparison of extensions of RKML with different design choices of semantic similarity for the top t annotated tags on the IAPR TC12 dataset. The leftmost column lists the different weighting methods, where the name before "-" denotes the local weights shown in Table 3.9 and the name behind "-" indicates the global weights shown in Table 3.10. "Cosine" represents the cosine similarity between tag vectors of two images.

Table 3.11 shows that different semantic similarity measures, either TF-IDF based weighting or the popular cosine similarity, provide essentially similar performances. We hence adopt the Log-Entropy weighting scheme in our experiments.

3.8 Summary

In this section, we propose a robust and efficient algorithm RKML for kernel metric learning. The proposed method addresses (i) high computational cost by avoiding the projection into PSD cone, (ii) limitation of binary constraints in tags by adopting a real-valued similarity measure, and as well as (iii) the overfitting problem by appropriately regularizing the rank of the learned kernel metric. Experiments with large-scale image annotation demonstrate the effectiveness and efficiency of the proposed RKML algorithm by comparing it to the state-of-the-art approaches for distance metric learning and image annotation. In the future, we plan to improve the annotation performance by developing a more robust semantic similarity measure.

Chapter 4

Image Tag Matrix Completion by Noisy Matrix Recovery

In this Section, we propose an *Image Tag Completion by Noisy Matrix Recovery (TCMR)* algorithm, which is able to simultaneously recover the missing tags and remove or down weight the noisy tags that are irrelevant to the visual content of images. In particular, this algorithm is designed for image tag completion, but actually it is not exclusive to image tag completion but also works pretty well for relevant image tagging tasks including image tag refinement and image tag re-ranking.

The rest of the chapter is arranged as follows. Section 4.1 motivates the problem and states main intuition behind the proposed algorithm, and as well as setups the notations. Section 4.3 introduces the detailed description of noisy matrix recovery, and extends it to the proposed TCMR algorithm. The theoretical properties and guarantee, *i.e.*, the bounds of error between the recovered tag matrix and its statistical optimal one, is given in Section 4.4, and the omitted proofs are deferred to Section 4.5. Section 4.6 presents the detailed implementation issues, and describes the proposed framework that incorporates image content consistency into the matrix completion based topic model through a graph Laplacian. Section 4.7 describes the intensive experimental setup, results and analysis. Section 4.8 concludes the chapter with future directions and Section 4.2 reviews the closely related works.

4.1 Motivation and Setup

It is apparent that different semantic tags have different biased significance in describing a topic that is determined by the image contents, and the tags associated with the same topic usually have a strong dependency on each other, which can be exploited to improve the annotation or tag completion performance.

The proposed TCMR algorithm addresses the incomplete and noisy tag problems by attempting to efficiently recover the missing tags and remove or down weight the noisy tags simultaneously. The inspiration and underlying concept behind the TCMR algorithm is the connection between the following two assumptions.

- **Idea of Language Model**. Since the tags are generated from the user's description of an image, each tag vector can be viewed as a mixture of topics and each topic follows a multinomial distribution over the vocabulary [12, 108, 206]. Note that the number of observed tags for each image is limited, while the number of parameters of the multinomial distribution to be estimated is significantly larger than the number of observed tags.

- **Low Rank Matrix Recovery**. Observed tags of any image can be assumed to sample from a mixture of a small number of multinomial distributions, which can be interpreted equivalently that the recovered tag matrix has to be of *low rank*.

With the connection of these two assumptions, the proposed TCMR algorithm enforces the recovered matrix to be low rank. Through an appropriate nuclear norm regularizer, it is able to effectively capture the interactions among different tag information, both tag keywords (column-wise) and tag vectors between different images (row-wise), which turns out to be the key in filling out missing tags and down weighting noisy ones [19, 136, 184].

Unlike in most matrix completion problems where the observed matrix entries are sampled uniformly at random from a given matrix [19, 20], each tag entry in our problem setting is sampled from an unknown multinomial distribution, making the standard least square loss and absolute loss inappropriate. Hence a maximum likelihood estimator is used in this work to ensure the learned tag probability matrix to be consistent with the observed tags, and this strategy also adds more complexity to both optimization and analysis as well.

It is noticed that although low rank matrix recovery is closely related to topic model that has been applied to many image tag related problems [108, 206, 221], most existing topic models [11] need to solve a non-convex optimization problem, which may cause the failure in finding the global optimum and turns out to be a big challenge in matrix completion problem [20, 21]. To address this limitation, TCMR proposes to solve a convex optimization problem which ensures to efficiently converge to an optimal solution.

Besides, theoretical support is provided to show that under favorable conditions, TCMR is guaranteed to recover most of the missing tags even when the user-provided tags are noisy, and that is novel among the existing studies for tag completion [126, 134, 201, 226]. Additionally, TCMR improves the tagging performance by exploiting the dependencies between image features and tags via a graph Laplacian [224, 226], which reduces the impact of incomplete and noisy tags by assigning high weights to tags that are consistent with the image visual contents, and low weights to those which are not, particularly under extreme cases. Furthermore, the empirical evaluation on tag re-ranking and tag refinement tasks demonstrates that TCMR is generally applicable and effective to other image tagging tasks.

Figure 4.1 highlights the key components of the proposed image tag completion algorithm by noisy matrix recovery. On the one hand, it ameliorates the tag confidence score by enforcing the tag matrix to be low rank. This strategy takes the advantage of both the

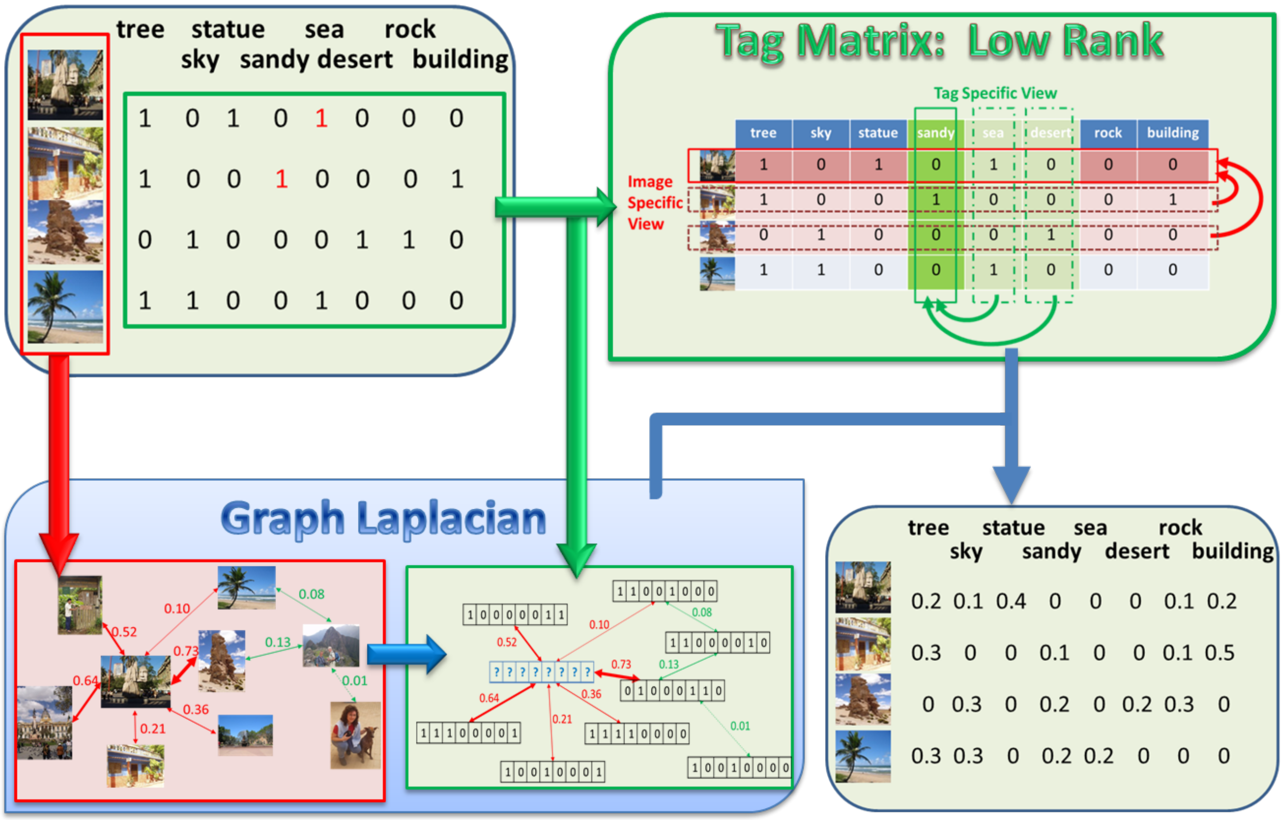


Figure 4.1: The scheme of the proposed noisy tag matrix recovery framework, *i.e.*, TCMR, for image tag completion. The low rank matrix recovery component in the upper right box exploits the tag-tag correlation, and the graph Laplacian component in the bottom left takes into account of the tag-content correlation.

row-wise and column-wise interactions within the tag matrix, *i.e.*, the dependencies among tag words and correlations of tag information between images. On the other hand, a graph Laplacian is constructed based on the visual features of images, where each node represents an image and each edge is weighted based on the distance between its connected images. Then the tag vector of an image is modified based on the weighted majority voting results among its connected neighbors. These two components conjoin together in a convex optimization framework and finally modify the tag confidence score matrix.

4.2 Related Work

There are only a few studies fitting the category of **image tag completion** with both incomplete and noisy tags. [226] proposes a data-driven framework for tag ranking that optimizes the correlation between visual cues and assigned tags. [129] removes the noisy tags based on the visual and semantic similarities, and expands the observed tags with their synonyms and hypernyms using WordNet. [201] proposes to search for the optimal tag matrix that is consistent with both observed tags and visual similarity. [134] formulates tag completion into a non-negative data factorization problem. [126] exploits sparse learning techniques to reconstruct the tag matrix. None of these studies provides any theoretical guarantee for their approaches. Matrix decomposition is adopted in [15, 149, 224] to handle both missing and noisy tags. The key limitation of these approaches is that they require a full observed matrix with a small number of errors, making it inappropriate for tag completion.

Low rank matrix recovery has been applied in many applications [19, 149], including visual recovery [136, 149], multilabel classification [15], tag refinement [224], etc. Since the function of matrix rank is non-convex, a popular approach is to replace it with the nuclear norm, the tightest convex relaxation for matrix rank [19, 20, 224]. Using the nuclear norm regularization, it is possible to accurately recover a low rank matrix from a small fraction of its entries [20] even if they are corrupted with noise [19, 57]. Various algorithms [57, 94, 149, 224] have been developed to solve the related optimization problem. Instead of the ℓ_1-norm loss [57, 224], squared loss [184] and max-margin factorization model [136] used in most studies on matrix completion/recovery, a maximum likelihood estimation is used in our work to recover the underlying tag matrix.

4.3 Tag Completion by Noisy Matrix Recovery (TCMR)

In this section, we describe a noisy matrix recovery framework for tag completion. And before presenting our algorithm and analysis, we first introduce the notations that will be used throughout this paper. We use $Q_{*,i}$ to represent the i-th column of matrix Q, $|Q|_F$, $|Q|_{tr}$ and $|Q|_*$ to represent the Frobenius norm, nuclear (trace) norm and spectral norm of matrix Q, respectively. $|Q|_1$ is used to represent the ℓ_1 norm of matrix Q, *i.e.*, $|Q|_1 = \sum_{i,j} |Q_{i,j}|$, and $|\mathbf{v}|_\infty$ is used to represent the infinity norm of vector $\mathbf{v}$, *i.e.*, $|\mathbf{v}|_\infty = \max_i |v_i|$. We also use $\mathbf{e}_i \in \{0,1\}^n$ to represent the i-th canonical basis for $\mathbb{R}^n$, and $\mathbf{1} \in \mathbb{R}^m$ to represent a m-dimensional vector with all its entries being 1.

To begin, let m be the number of unique tags, and $\mathcal{D} = \{\mathbf{d}_1, \cdots, \mathbf{d}_n\}$ be a collection of n tagged images, where $\mathbf{d}_i = (d_{i,1}, \cdots, d_{i,m})$ is the tag vector for the i-th image with $d_{i,j} = 1$ when tag j is assigned to the image and zero, otherwise. For the simplicity of analysis, in this study, we assume that all the images have the same number of assigned tags, denoted by m_* [1]. When different number of tags are observed, we can apply a simple weighting technique [150] to handle the variation in the number of tags.

Our development is based on the simple observation that the essential goal of topic model is to approximate an observed tag probability matrix by a low rank matrix (or more precisely, the product of two low rank matrices). It is this observation that motivates us to connect topic (probabilistic) model with low rank matrix completion.

[1] Note this assumption is only for the convenience of analysis, and does not affect the algorithm.

4.3.1 Noisy Matrix Recovery

Following the idea of language models [11, 12, 225], we assume that all the observed tags in each image are drawn independently from a fixed but unknown multinomial distribution. Let $\mathbf{p}_i = (p_{i,1}, \cdots, p_{i,m})$ be the multinomial distribution used to generate tags in $\mathbf{d}_i$. We use $P = (\mathbf{p}_1, \cdots, \mathbf{p}_n)$ to represent the multinomial distributions for all the images. Our goal is to accurately recover the multinomial distribution P from a limited number of observed tags in $\mathcal{D}$. In general, this is impossible since the number of parameters to be estimated is significantly larger than the number of observed tags. To address this challenge, we follow the key assumption behind most topic models [184, 224], *i.e.* tags of any image are sampled from a mixture of a small number of multinomial distributions. A direct implication of this assumption is that matrix P has to be of low rank, the foundation for the theory of low rank matrix recovery [20].

The proposed approach combines the idea of maximum likelihood estimation, a common approach for topic model, and the theory of low rank matrix recovery. It aims to recover the multinomial probability matrix P by solving the following optimization problem

$$\min_{Q \in \Delta} \quad \mathcal{L}(Q) := \underbrace{-\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{d_{i,j}}{m_*} \log Q_{i,j}}_{:=E_1} + \underbrace{\varepsilon \operatorname{rank}(Q)}_{:=E_2}, \tag{4.1}$$

where domain Δ is defined as

$$\Delta = \{Q \in (0,1)^{m \times n} \,:\, Q_{i,*}\mathbf{1}^\top = 1,\, i \in [1, n]\}, \tag{4.2}$$

and ε is a regularization parameter. We denote by $\hat{Q}$ the optimal solution to (4.1). Term

E_1 in (4.1) ensures the learned probability matrix $\hat{Q}$ to be consistent with the observed tag matrix, and term E_2 ensures that $\hat{Q}$ is of low rank, which indicates that all tags of an image are sampled from a mixture of a small number of multinomial distributions.

We note that unlike standard matrix completion theory [20, 61] where observed entries are sampled uniformly at random from a given matrix, in our model as well as other topic model, each observed tag is sampled from an unknown multinomial distribution. This difference makes the common least square loss and absolute loss inappropriate.

However, the function of matrix rank in (4.1) is non-convex and non-differentiable, which poses a problem in the optimization procedure. Therefore, we replace the rank function with the nuclear norm, the tightest convex envelope of the matrix rank function [19, 20, 21]. The nuclear norm of a matrix Q is defined as the sum of singular values of Q. With the nuclear norm regularization, it is possible to accurately recover a low rank matrix from a small fraction of its entries [20, 21, 161] even if they are corrupted with noise [19, 57, 100, 107], which exactly fits the missing and noisy tag situation. Consequently, the optimization problem in (4.1) becomes

$$\min_{Q \in \Delta} \quad \mathcal{L}(Q) = -\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{d_{i,j}}{m_*} \log Q_{i,j} + \varepsilon \, |Q|_{tr}, \tag{4.3}$$

and the domain Δ is still defined the same as in 4.2.

In (4.3) the sparsity of the recovered matrix $\hat{Q}$ is introduced by the nuclear norm. Nuclear norm regularizer enforces the matrix completion to favor the interactions between rows and columns to find a global solution [14], which is in contrast to Frobenius and ℓ_1 norm regularizers that deal with each entry in the matrix independently.

4.3.2 Incorporating Irrelevant Tags into Noisy Matrix Recovery

Regarding the fact that the initially unobserved tags are with a small probability relevant to the associated image, we also maximize the likelihood of their irrelevance, so the loss functions in both (4.1) and (4.3) are thus updated, and the optimization problem becomes

$$\min_{Q\in\Delta} \quad \mathcal{L}(Q) = -\sum_{i,j=1}^{n,m}\left[\frac{d_{i,j}}{m_*}\log Q_{i,j} + \frac{1-d_{i,j}}{m-m_*}\log(1-Q_{i,j})\right] + \varepsilon\,|Q|_{tr}, \tag{4.4}$$

where domain Δ remains to be 4.2.

4.4 Theoretical Guarantee of RKML

The following theorem bounds the difference between P, the recovered tag matrix by TCMR, and the optimal recovered probability matrix $\hat{Q}$.

Theorem 4.1. *Let r be the rank of matrix P, and N be the total number of observed tags. Let $\hat{Q}$ be the optimal solution to (4.3). Assume $N \geq \Omega(n\log(n+m))$, and denote by μ_- and μ_+ the lower and upper bounds for the probabilities in P.*

Then we have, with a high probability

$$\frac{1}{n}|\hat{Q}-P|_1 \leq O\left(\frac{rn\theta^2\log(n+m)}{N}\right), \tag{4.5}$$

where

$$\theta^2 := \frac{\mu_+|P\mathbf{1}|_\infty}{n\mu_-^2} \leq \frac{\mu_+^2}{\mu_-^2}.$$

A sketch of the proof is provided in Section 4.5.4.

It is clear that the recovery error is $O(rn\log(n+m)/N)$, implying that the tag matrix can be accurately recovered when $N \geq \Omega(rn\log(n+m))$. This is consistent with the standard results in matrix completion [107] and low rank matrix recovery [106]. The impact of low rank assumption is analyzed in Section 4.4.1. However, in stead of square loss used in standard matrix completion theory, we adopt maximum likelihood loss function in our model, which leads to additional challenges in analyzing the recovery property for our model.

4.4.1 Impact of Low Rank Assumption on Recovery Error

In order to see the impact of low rank assumption, let us consider the maximum likelihood estimation of multinomial distribution. Since tags for different images are sampled independently, we only need to consider one image at each time. Let $\mathbf{p}$ be the underlying multinomial distribution to be estimated, and let $\mathbf{d}$ be the image tag vector comprised of m_* words sampled from $\mathbf{p}$. We estimate $\mathbf{p}$ by the simple maximum likelihood estimation, *i.e.*,

$$\min_{\mathbf{p}\in[\mu_-,\mu_+]^m:\mathbf{p}^\top\mathbf{1}=1} \quad -\sum_{i=1}^{n} d_i \log p_i, \tag{4.6}$$

where m is the number of unique tags, n is the number of images, μ_- and μ_+ are the lower and upper bounds for the probabilities in matrix $P = (\mathbf{p}_1, \cdots, \mathbf{p}_n)$.

Theorem 4.2. *Define* $\mathbf{z}$ *as*

$$\mathbf{z} = \frac{\mathbf{d}}{m_*} - \mathbf{p}.$$

Let $\hat{\mathbf{q}}$ *be the optimal solution to (4.6). Then*

$$|\mathbf{p} - \hat{\mathbf{q}}|_1 \leq \frac{\mu_+^2}{\mu_-^2}|\mathbf{z}|_2^2.$$

And to bound $|\mathbf{z}|_2$, we need the following concentration inequality for vectors.

Theorem 4.3. *With a probability* $1 - 2e^{-t}$,

$$|\mathbf{z}|_2 \leq \sqrt{\frac{t + \log m}{\mu_- m_*}} |\mathbf{p}|_2.$$

Following the concentration inequality for vectors in Theorem 4.3, we bound $|\mathbf{z}|_2$. Then by combining Theorems 4.2 and 4.3, we have, with a probability $1 - 2e^{-t}$,

$$|\mathbf{p} - \hat{\mathbf{q}}|_1 \leq \frac{\mu_+^2 |\mathbf{p}|_2^2}{\mu_-^4} \frac{2(t + \log m)}{m_*}$$

By applying the above result to matrix P and taking the union bound, we have, with probability $1 - e^{-t}$,

$$\frac{1}{n}|P - \hat{Q}|_1 \leq \frac{\mu_+^2}{\mu_-^4} \max_{1 \leq i \leq n} |\mathbf{p}_i|_2^2 \frac{2n(t + \log m + \log n)}{N}. \tag{4.7}$$

We now compare the bound in (4.7) to that in (4.5). It is easy to verify that $|\mathbf{p}_i|_2^2/\mu_-^2 \geq m$ for any $\mathbf{p}_i$. Hence, the net effect of the bound in (4.5) is to replace m with r, which is exactly the impact of low rank assumption.

4.5 Proofs of Error Bounds

In this section, we give out the proofs of the main theorems proposed in Section 4.4.

4.5.1 Proof of Theorem 4.1

Proof. Define matrix M as

$$M := \sum_{i=1}^{n} \left(\frac{1}{m_*} \mathbf{d}_i - \mathbf{p}_i \right) \mathbf{e}_i^\top = \sum_{i=1}^{n} \frac{1}{m_*} \mathbf{d}_i \mathbf{e}_i^\top - P, \tag{4.8}$$

where $\mathbf{e}_i \in \{0,1\}^n$ is the canonical base for $\mathbb{R}^n$. Since the occurrence of each tag in $\mathbf{d}_i$ is sampled according to the underlying multinomial distribution $\mathbf{p}_i$, it is easy to verify that

$$\mathrm{E}[M] = 0.$$

Before presenting our analysis, we need two supporting lemmas that are important to our analysis. The detailed proofs of these lemmas are provided in Section A.2.

Lemma 4.4. *Let $P \in \Delta$ and $Q \in \Delta$ be two probability matrices. We have*

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{|P_{i,j} - Q_{i,j}|^2}{Q_{i,j}} \geq \sum_{i=1}^{n} \sum_{j=1}^{m} |P_{i,j} - Q_{i,j}| = |P - Q|_1.$$

Lemma 4.5. *([107]) Let $Z_1, \cdots, Z_n$ be independent random matrices with dimension $m_1 \times m_2$ that satisfy $\mathrm{E}[Z_i] = 0$ and $|Z_i|_* \leq U$ almost surely for some constant U, and all $i = 1, \cdots, n$. Define*

$$\sigma_Z = \max \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[Z_i Z_i^\top] \right|_*, \left| \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[Z_i^\top Z_i] \right|_* \right\}.$$

Then, for all $t > 0$, with a probability $1 - e^{-t}$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_i \right|_* \leq 2 \max \left\{ \sigma_Z \sqrt{\frac{t + \log(m_1 + m_2)}{n}}, U \frac{t + \log(m_1 + m_2)}{n} \right\}.$$

The following theorem is the key to our analysis. It shows that the estimation error $|P - Q|_1$, measured by ℓ_1 norm, will be small when P can be well approximated by a low rank matrix.

Theorem 4.6. *Let $\hat{Q}$ be the optimal solution to (4.3). If*

$$\varepsilon \geq \frac{1}{\mu_-} |M|_*,$$

where M is defined in (4.8), then

$$|\hat{Q} - P|_1 \leq \min_{Q \in \Delta} \left\{ \frac{|Q - P|_F^2}{\mu_-} + 16 \varepsilon^2 \mu_+ rank(Q) \right\}.$$

To utilize Theorem 4.6 for bounding the difference between P and $\hat{Q}$, we need to bound $|M|_*$. The theorem below bounds $|M|_*$ by using Lemma 4.5.

Theorem 4.7. *Define γ as*

$$\gamma := \frac{2}{\mu_-} \max \left\{ \frac{t + \log(m + n)}{m_*}, \sqrt{\max(1, |P\mathbf{1}|_\infty) \frac{t + \log(n + m)}{m_*}} \right\}. \tag{4.9}$$

Then with a probability $1 - e^{-t}$, we have

$$|M|_* \leq \gamma \mu_-.$$

Combining Theorems 4.6 and 4.7, we have the following result for recovering the probability matrix P.

Corollary 4.8. *Set $\varepsilon = \gamma$. With a probability at least $1 - e^{-t}$, we have*

$$|\hat{Q} - P|_1 \leq \min_{Q \in \Delta} \left\{ \frac{|Q - P|_F^2}{\mu_-} + 16\gamma^2 \mu_+ rank(Q) \right\}.$$

Furthermore, let $\hat{P}$ be the best rank-r approximation of P. We have, with a probability $1-e^{-t}$

$$|\hat{Q} - \hat{P}|_1 \leq \frac{|P - \hat{P}|_F^2}{\mu_-} + 16\gamma^2 \mu_+ r.$$

We now come to the proof of Theorem 4.1. When the rank of P is r, using Corollary 4.8, we have, with a high probability,

$$|\hat{Q} - P|_1 \leq 16\gamma^2 \mu_+ r.$$

If $|P\mathbf{1}|_\infty \geq 1$ and $m_* \geq O(\log(m+n))$, we have

$$\gamma = O\left(\frac{1}{\mu_-} \sqrt{|P\mathbf{1}|_\infty \frac{\log(n+m)}{m_*}} \right)$$

and therefore, with a high probability, we have

$$\frac{1}{n}|\hat{Q} - P|_1 \leq O\left(\frac{r\log(n+m)}{m_*} \frac{\mu_+ |P\mathbf{1}|_\infty}{\mu_-^2} \right) \leq O\left(\frac{rn\log(n+m)}{N} \frac{\mu_+ |P\mathbf{1}|_\infty}{n\mu_-^2} \right).$$

where N is the number of observed tags. This immediately implies Theorem 4.1. $\square$

4.5.2 Proof of Theorem 4.2

Proof. Following the same analysis as that for Theorem 4.6 whose proof is provided in Section 4.5.4, we have

$$\sum_{i=1}^{m} \frac{(p_i - \hat{q}_i)^2}{\hat{\mathbf{q}}_i} \leq \sum_{i=1}^{m} \frac{z_i}{\hat{q}_i}(p_i - \hat{\mathbf{q}}_i).$$

Using the fact $\hat{\mathbf{q}}_i \in [\mu_-, \mu_+]$, we have

$$|\mathbf{p}_i - \hat{\mathbf{q}}_i|_2^2 \leq \frac{\mu_+}{\mu_-}|\mathbf{z}|_2|\mathbf{p} - \hat{\mathbf{q}}|_2,$$

and therefore

$$|\mathbf{p}_i - \hat{\mathbf{q}}|_2 \leq \frac{\mu_+}{\mu_-}|\mathbf{z}|_2.$$

We finally complete the proof by using the fact

$$\sum_{i=1}^{m} \frac{(p_i - \hat{q}_i)^2}{\hat{q}_i} \geq |\mathbf{p} - \hat{\mathbf{q}}|_1.$$

□

4.5.3 Proof of Theorem 4.3

Proof. We will use the Chernoff bound, i.e. $X_1, \cdots, X_{m*}$ be independent draws from a Bernoulli distribution with $\mathbb{P}(X = 1) = \mu$. We have

$$\mathbb{P}\left(\frac{1}{m_*}\sum_{i=1}^{m_*} X_i \geq (1+\delta)\mu\right) \leq \exp\left(-\frac{\delta^2 \mu m_*}{3}\right),$$
$$\mathbb{P}\left(\frac{1}{m_*}\sum_{i=1}^{m_*} X_i \leq (1-\delta)\mu\right) \leq \exp\left(-\frac{\delta^2 \mu m_*}{2}\right).$$

Using the Chernoff bound, we have, with a probability $1 - 2\exp(-\delta^2 \mu m_*/2)$

$$|X - \mu|^2 \leq \delta^2 \mu^2.$$

By taking the union bound, we have, with a probability $1 - 2e^{-t}$

$$|\mathbf{z}|_2 \leq \sqrt{\frac{t + \log m}{\mu_- m_*}} |\mathbf{p}|_2.$$

$\square$

4.5.4 Proof of Theorem 4.6

Proof. We consider any solution $Q \in \Delta$. Since $\hat{Q}$ is the optimal solution to Eq (4.3), we have $\langle \nabla \mathcal{L}(\hat{Q}), \hat{Q} - Q\rangle \leq 0$, *i.e.*

$$-\frac{1}{m_*}\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{d_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) + \varepsilon\langle \partial|\hat{Q}|_{tr}, \hat{Q} - Q\rangle \leq 0,$$

where $\partial|\hat{Q}|_{tr}$ is a subgradient of $|\hat{Q}|_{tr}$. Using the fact that

$$\langle \partial|\hat{Q}|_{tr} - \partial|Q|_{tr}, \hat{Q} - Q\rangle \geq 0,$$

we can replace $\langle \partial|\hat{Q}|_{tr}, \hat{Q} - Q\rangle$ with $\langle \partial|Q|_{tr}, \hat{Q} - Q\rangle$, which results in the following inequality

$$-\frac{1}{m_*}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{d_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) + \varepsilon\langle \partial|Q|_{tr}, \hat{Q} - Q\rangle \leq 0.$$

Define $Z_{i,j} = (\hat{Q}_{i,j} - Q_{i,j})/\hat{Q}_{i,j}$. We have

$$-\frac{1}{m_*}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{d_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) = -\frac{1}{m_*}\sum_{i=1}^{n}\langle \mathbf{d}_i\mathbf{e}_i^\top, Z\rangle = -\langle P, Z\rangle - \langle M, Z\rangle.$$

Thus the bound in Eq (4.5) is modified as

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{P_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) + \varepsilon\langle \partial|Q|_{tr}, \hat{Q} - Q\rangle \leq \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{M_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right).$$

Since

$$-\sum_{j=1}^{m}\frac{P_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) = -\sum_{j=1}^{m}\frac{1}{\hat{Q}_{i,j}}\left(P_{i,j} - \hat{Q}_{i,j}\right)\left(\hat{Q}_{i,j} - Q_{i,j}\right).$$

we have

$$-\sum_{j=1}^{m}\frac{P_{i,j}}{\hat{Q}_{i,j}}\left(\hat{Q}_{i,j} - Q_{i,j}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(\hat{Q}_{i,j} - P_{i,j})^2}{2\hat{Q}_{i,j}} + \frac{(\hat{Q}_{i,j} - Q_{i,j})^2}{2\hat{Q}_{i,j}} - \frac{(Q_{i,j} - P_{i,j})^2}{2\hat{Q}_{i,j}}.$$

Define matrix $B \in \mathbb{R}^{n\times m}$ as $B_{i,j} = M_{i,j}/\hat{Q}_{i,j}$. Using the fact $\hat{Q}_{i,j} \in [\mu_-, \mu_+]$ and result from Lemma 4.4, we have

$$\frac{1}{2}|P-\hat{Q}|_1 + \frac{|\hat{Q}-Q|_F^2}{2\mu_+} + \varepsilon\langle \partial |Q|_{tr}, \hat{Q}-Q\rangle \le \frac{|M|_*}{\mu_-}|\hat{Q}-Q|_{tr} + \frac{|P-Q|_F^2}{2\mu_-}.$$

We write the Singular value decomposition of Q as

$$Q = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \tag{4.10}$$

where r is the rank of Q, σ_i is the i-th singular value of Q, and $(\mathbf{u}_i, \mathbf{v}_i)$ are the left and right singular vectors of Q. Let $U_\perp \in \mathbb{R}^{n\times(n-r)}$ and $V_\perp \in \mathbb{R}^{m\times(m-r)}$ be the orthogonal bases complementary to U and V, respectively. Define the linear operators $\mathcal{P}_Q$ and $\mathcal{P}_Q^\perp$ as

$$\mathcal{P}_Q(Z) = UU^\top Z + ZVV^\top - UU^\top ZVV^\top, \;\; \mathcal{P}_Q^\perp(Z) = Z - \mathcal{P}_Q(Z).$$

According to (4.10), the subgradient $\partial|Q|_{tr}$ is given by the set $\mathcal{W}$

$$\mathcal{W} = \left\{ UV^\top + U_\perp W V_\perp \;:\; W \in \mathbb{R}^{(n-r)\times(m-r)},\, |W|_* = 1 \right\}.$$

Thus by choosing an appropriate matrix W for the subgradient $\partial|Q_{tr}|$, we have

$$\langle \partial|Q|_{tr}, \hat{Q}-Q\rangle \ge -|\mathcal{P}_Q(\hat{Q}-Q)|_{tr} + |\mathcal{P}_Q^\perp(\hat{Q}-Q)|_{tr}$$

and therefore

$$\frac{1}{2}|P-\hat{Q}|_1 + \frac{|\hat{Q}-Q|_F^2}{2\mu_+} + \varepsilon|\mathcal{P}_Q^\perp(\hat{Q}-Q)|_{tr} \;\le\; \varepsilon|\mathcal{P}_Q(\hat{Q}-Q)|_{tr} + \frac{|M|_*}{\mu_-}|\hat{Q}-Q|_{tr} + \frac{|P-Q|_F^2}{2\mu_-}.$$

Using the fact

$$\varepsilon \geq \frac{1}{\mu_-}|M|_*,$$

we have

$$|P - \hat{Q}|_1 + \frac{|\hat{Q} - Q|_F^2}{\mu_+} \leq 4\varepsilon|\mathcal{P}_Q(\hat{Q} - Q)|_{tr} + \frac{|P - Q|_F^2}{\mu_-}.$$

We consider two cases. In the first case, we assume

$$|P - \hat{Q}|_1 \leq \frac{1}{\mu_-}|P - Q|_F^2,$$

in which the bound in theorem trivially holds. In the second case, we have the opposite

$$|P - \hat{Q}|_1 > \frac{1}{\mu_-}|P - Q|_F^2,$$

which implies

$$\frac{|\hat{Q} - Q|_F^2}{\mu_+} \leq 4\varepsilon|\mathcal{P}_Q(\hat{Q} - Q)|_{tr},$$

and therefore

$$|\mathcal{P}_Q(\hat{Q} - Q)|_{tr} \leq 4\varepsilon r \mu_+.$$

We complete the proof by plugging the above bound. $\square$

4.6 Implementation

In this section, we present two auxiliary techniques to improve the tag completion performance. We incorporate the visual features to improve the tag accuracy and use an extended gradient method to solve the optimization problem efficiently.

4.6.1 Incorporating Visual Features

The limitation of the noisy matrix recovery method in (4.4) is that it does not take advantage of the visual contents of the images, an important hint for accurate tag prediction. So we next modify (4.4) to incorporate the visual features. Here we introduce two common methods to integrate the visual features, including introducing the Graph Laplacian to the objective function in (4.4) and a linear combination of the recovered matrix P and the majority voting results among nearest neighbors.

4.6.1.1 Graph Laplacian Method

Let $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ include the visual features of all images, where vector $\mathbf{x}_i \in \mathbb{R}^d$ represents the visual content of the ith image. Let $W = [w_{i,j}]_{n\times n}$ be the pairwise similarity matrix, where $w_{i,j}$ is the visual similarity between images $\mathbf{x}_i$ and $\mathbf{x}_j$, *i.e.*,

$$w_{i,j} = \begin{cases} \exp\left[-\frac{d(\mathbf{x}_i,\mathbf{x}_j)^2}{\sigma^2}\right] & \text{if } j \in N_k(i) \text{ or } i \in N_k(j); \\ 0 & \text{otherwise,} \end{cases} \tag{4.11}$$

where $N_k(i)$ denotes the index set for the k nearest neighbors of the ith image, k is empirically set to $k = 0.001n$, $d(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and σ is the average distance. We adopt χ-distance if $\mathbf{x}_i$ is histogram features and Euclidean distance, otherwise.

Using matrix W, we can measure the consistency between the estimated tag probability matrix Q and visual similarities by

$$\sum_{i,j=1}^{n} W_{i,j}|Q_{*,i} - Q_{*,j}|^2 = Tr(Q^\top LQ), \tag{4.12}$$

where

$$L = \mathrm{diag}(W^\top \mathbf{1}) - W \tag{4.13}$$

and L is exactly the graph Laplacian defined in [202, 224]. By minimizing $Tr(Q^\top LQ)$, we ensure that the recovered probability matrix Q to be consistent with visual features, *i.e.*, similar images share similar tags.

Finally by combining the noisy matrix recovery component with the component of visual features, we recover the tag probability matrix Q by solving the following optimization problem

$$\min_{Q\in\Delta} \quad -\sum_{i,j=1}^{n,m}\left[\frac{d_{i,j}}{m_*}\log Q_{i,j} + \frac{1-d_{i,j}}{m-m_*}\log(1-Q_{i,j})\right] + \frac{\alpha}{n}Tr(Q^T LQ) + \beta|Q|_{tr}, \tag{4.14}$$

where Δ is defined in 4.2, and both α and β are regularization parameters.

By minimizing the objective function above, we are able to simultaneously fill out the missing tags and filter out/down weight the noisy tags. Figure 4.1 shows the framework of the whole algorithm described in 4.14, which includes the two principle components: the low rank noisy matrix recovery component reflecting the tag-tag correlation, and the graph Laplacian component exploring the image-tag dependencies.

4.6.1.2 Linear Reconstruction Approach

Although we incorporate the visual consistency in the proposed model with Graph Laplacian as explained in Section 4.6.1.1, TCMR mainly explores the statistic correlation between tags. As it is significantly obvious that visually similar images usually share similar semantic tags, we further emphasize the role of visual contents with an additional weighted linear reconstruction strategy following [126], which is simple yet empirically demonstrated to be effective

$$\Omega = \delta T + (1 - \delta)R, \tag{4.15}$$

where Ω is the expected final result, δ is a weighting parameter in $[0, 1]$, T is the normalized completion result of TCMR in (4.4), and R is the normalized tagging result generated by a majority voting strategy among the nearest neighbors of the images, or any other visual feature based image annotation results.

4.6.2 Efficient Solution of the Proposed Algorithm

We incorporate several heuristics to improve the computational efficiency. First, we adopt one projection paradigm that has been successfully applied to metric learning [39]. The key idea is to ignore the domain constraint $Q \in \Delta$ during the iteration, and only project the solution Q into Δ at the end of optimization. As a result, we only need to solve an unconstrained optimization problem. Secondly, we adopt the extended gradient method in [94]. To this end, we rewrite the objective function in (4.4) or (4.14) as $\mathcal{L}(Q) = f(Q) + \varepsilon|Q|_{tr}$. Then given the current solution Q_{k-1}, we update the solution Q_k by solving the following optimization problem

$$\arg\min_{Q} \quad P_{t_k}(Q, Q_{k-1}) = \frac{1}{2}\left|Q - \left(Q_{k-1} - \frac{1}{t_k}\nabla f(Q_{k-1})\right)\right|_F^2 + \frac{\varepsilon}{t_k}|Q|_{tr}. \tag{4.16}$$

where t_k is the step size for the kth iteration. The detailed algorithm for solving the unconstrained version of the objective functions can be found in [94].

4.6.3 Pseudo-code of TCMR

TCMR solves a semi-supervised learning problem and it modifies the tag confidence scores based on the initial binary tag matrix. Unlike the traditional image annotation algorithms, *e.g.*, RKML proposed in Chapter 3 that consists of training and testing phase, TCMR does the learning on the whole dataset and results in an updated tag matrix. Algorithm 2 summarizes the main steps of TCMR. To obtain the final tags for an image in the tag completion setting, we return the tags with top score as the final tags of an image.

4.7 Experiments

4.7.1 Datasets and Experimental Setup

Four benchmark datasets are used to evaluate our proposed algorithm. ESP Game dataset was collected for a collaborative image labeling task and consists of images including logos, drawings and personal photos. IAPR TC12 dataset consists of images of actions, landscapes, animals and many other contemporary life, and its tags are extracted from the text captions accompanying each image. Both Mir Flickr and NUS-WIDE datasets [33] include images

Algorithm 2 Image Tag Completion by Noisy Matrix Recovery

Input:

- Visual features of the whole image dataset: $X \in \mathbb{R}^{n \times d}$
- Binary tag matrix: labels $D \in \mathbb{R}^{n \times m}$
- Regularization parameters α and β
- Initial Lipschitz constant t_0, its increasing parameter γ and $k \leftarrow 0$

Compute the Laplacian matrix L based on X according to (4.12) and (4.13).
Initialize Q_0 and let all entries equal to 0.5.
while not converged **do**
  $k \leftarrow k + 1$,
  $C = Q_{k-1} - \frac{1}{t_{k-1}} \nabla f(Q_{k-1})$,
  Compute singular value decomposition: $U \Sigma V^T = C$,
  $t_k \leftarrow t_{k-1}/\gamma$.
  repeat
    $t_k = \gamma t_k$,
    $\widetilde{Q} = U \Sigma_k V^T$, where Σ_k is diagonal with $(\Sigma_k)_{ii} = \max(0, \Sigma_{ii} - \frac{\beta}{t_k})$.
  until $F(\widetilde{Q})$ (4.4 or 4.14) $<= P_{t_k}(\widetilde{Q}, Q_{k-1})$ (4.16)
  $Q_k \leftarrow \widetilde{Q}$.
end while
$Q \leftarrow Q_k$.
Output: Matrix of tag relevance score $Q \in \mathbb{R}^{n \times m}$.

crawled from Flickr [2], together with users provided tags. ImageNet [3] is an image dataset organized according to the WordNet hierarchy, which contains more than 20K concepts [4].

ESP Game and IAPR TC12 are collaboratively human labeled and thus relatively clean, while Mir Flickr and NUS-WIDE are automatically crawled from social media and hence pretty noisy. Besides, with the WordNet hierarchy, ImageNet is able to offer tens of millions of cleanly sorted images for most of the provided concepts. A bag-of-words model based on densely sampled SIFT descriptors is used to represent the visual content in Mir Flickr, ESP Game, IAPR TC12 and ImageNet datasets[5] [6]. In NUS-WIDE dataset, visual content

[2] `https://www.flickr.com/`.

[3] `http://www.image-net.org/`

[4] The list of ImageNet concepts could be referred to `http://www.image-net.org/archive/words.txt`.

[5] The features were obtained from `http://lear.inrialpes.fr/people/guillaumin/data.php`. More detailed description about Mir Flickr, ESP Game and IAPR TC12 can also be found in [67, 69].

[6] ImageNet offers a 1.2M subset of images with SIFT feature, which can be downloaded through `http://www.image-net.org/download-features.php`.

are represented by six low-level features, including color information, edge distribution and wavelet texture [33].

To evaluate the proposed approach for tag completion, we divide the original tag matrix Y into two parts: the observed tag matrix (*i.e.* training set) D and the left as evaluation ground truth (*i.e.* testing set). We create the observed tag matrix by randomly sampling a subset of tags from D for each image. To guarantee that the evaluation is meaningful, we ensure that each image has at least one evaluation tag by filtering out images with too few tags and tags associated with only a few images. Detailed statistics about the refined datasets are listed in Table 4.7.1. All the hyper parameter values used in TCMR, *e.g.* ε, α, β, and the parameter values in the baselines are determined by cross-validation.

	ESP Game	IAPR TC12	MirFlickr	NUS-WIDE	ImageNet
Number of Imgs	10,450	12,985	5,231	20,968	1,253,679
Feature dimension	1000	1000	1000	500	1000
Vocabulary size	265	291	372	420	1,625
Average tags/img	6.41	7.07	5.82	10.4	27.54
Min/max tags/img	5/15	5/23	4/43	9/15	5/125
Average imgs/tag	253.0	315.5	81.9	519.6	4751
Min/max imgs/tag	16/3,439	14/4,752	10/781	78/5,058	300/25,361
Num of observed tags*	4	4	3	4	4

Table 4.1: Statistics for the refined datasets. * indicates the number of observed tags when training the TCMR model throughout the experimental section if no specific explanation.

Following [126], we evaluate the tag completion accuracy by the *average precision @N* (*AP@N*). It measures the average percentage of the top N recovered tags that are correct. Note that a tag is correctly recovered if it is included in the original tag matrix Y but not observed in D. We also use *average recall* (*AR@N*) to measure the percentage of correct tags that are recovered by a computational algorithm out of all ground truth tags, and *coverage* (*C@N*) to measure the percentage of images with at last one correctly recovered tag. Both the mean and standard deviation of evaluation metrics over 20 experimental trials are reported in this paper.

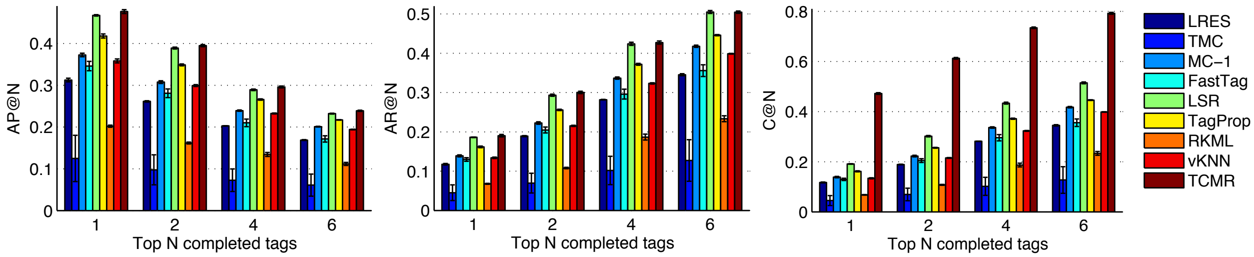
4.7.2 Comparison to state-of-the-art Tag Completion Methods

We first compare our TCMR algorithm[7] [8] proposed by (4.14) to several state-of-the-art tag completion approaches: 1) LRES [224], tag refinement towards low-rank, content-tag prior and error sparsity, 2) TMC [201] that searches for the optimal tag matrix consistent with both the observed tags and visual similarity, 3) MC-1 [15] which applies low rank matrix completion to the concatenation of visual features and assigned tags, 4) FastTag [28] that co-regularizes two simple linear mappings in a joint convex loss function, 5) LSR [126] that optimally reconstructs each image and each tag with remaining ones under constraints of sparsity. We also compare the proposed approach with three state-of-the-art image annotation algorithms that are designed for clean tags: 6) TagProp [67], 7)RKML [52], a kernel metric learning algorithm, and 8) vKNN [123], a nearest neighbor voting algorithm. Since most of them are originally designed for image annotation, we train the model using the observed tags first over the whole gallery, and then apply the models to the gallery to update the tag matrix.

Figure 4.2 shows the image tag completion results on the IAPR TC12 dataset measured by $AP@N$, $AR@N$ and $C@N$, respectively. Figure 4.3 show the tag completion performance on the left three datasets; where the rows represent different evaluation measures and the columns indicate different datasets. We observe that overall, the proposed TCMR and LSR

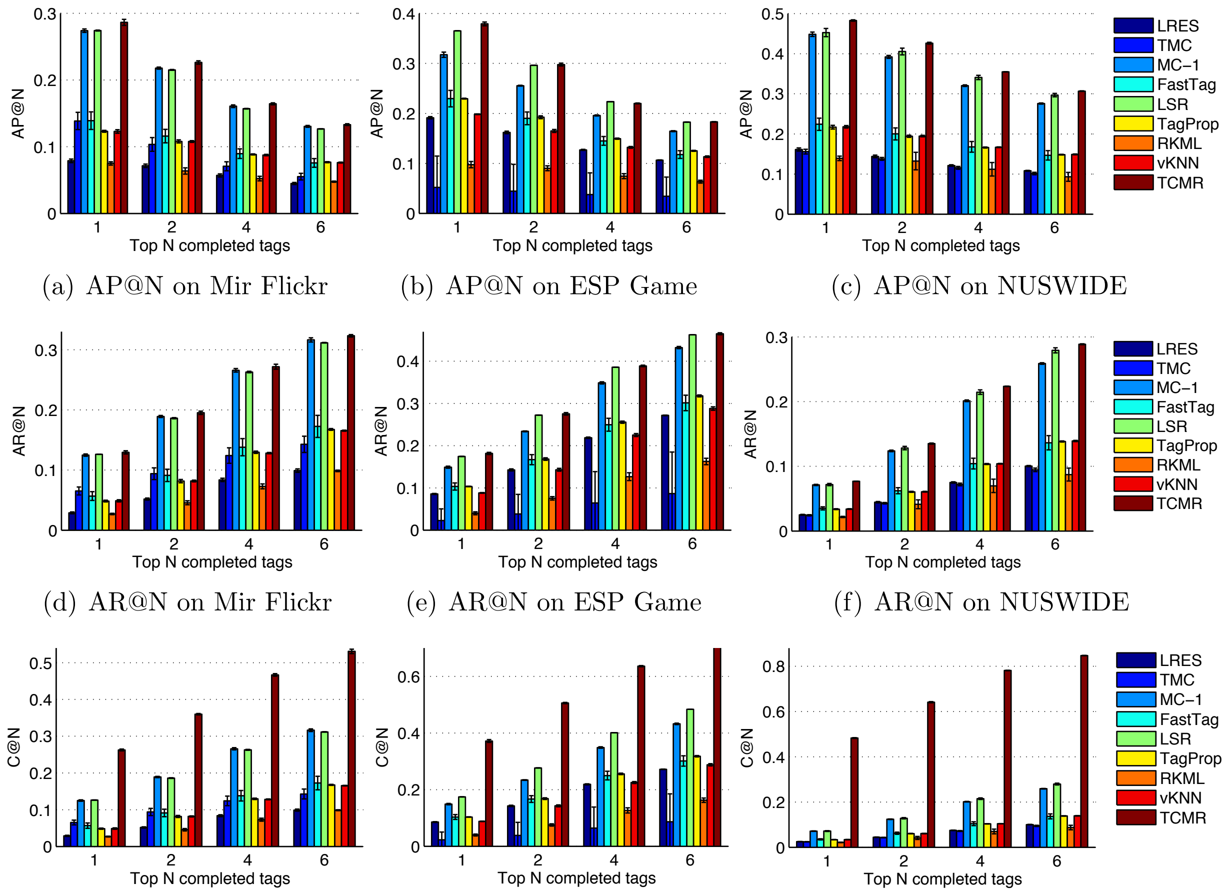
[7] Note that if without notification, TCMR stands for the algorithm proposed in (4.14), and TCMR-lr stands for the one proposed in 4.15.

[8] The source code of TCMR can be downloaded from our website `http://www.cse.msu.edu/~fengzhey/downloads/src/tcmr.zip`.



(a) AP@N (b) AR@N (c) C@N

Figure 4.2: Comparison of tag completion performance between TCMR and state-of-the-art baselines on IAPR TC12 dataset.



(a) AP@N on Mir Flickr (b) AP@N on ESP Game (c) AP@N on NUSWIDE

(d) AR@N on Mir Flickr (e) AR@N on ESP Game (f) AR@N on NUSWIDE

(g) C@N on Mir Flickr (h) C@N on ESP Game (i) C@N on NUSWIDE

Figure 4.3: Comparison of tag completion performance between TCMR and state-of-the-art baselines on other datasets including Mir Flickr, ESP Game and NUS-WIDE.

yield significantly better performance than the other approaches in comparison. TCMR performs significantly better than LSR in terms of $C@N$, as well as the other methods. In particular, TCMR recovers at least one correct tag out of the top six predicted tags for 80%

of the images while the other approaches are only able to recover at least one correct tag for less than 50% of the images, indicating that the proposed algorithm is more effective in recovering relevant tags for a wide range of images, an important property for image tag completion algorithm. We also observe that TCMR performs slightly better than LSR in terms of $AP@N$ when the number of predicted tags N is small.

4.7.2.1 Efficiency Evaluation

	LRES	TMC	MC-1	FastTag	LSR	TagProp	RKML	TCMR
MirFlickr	5.6e2	4.7e3	8.6e2	1.4e3	6.2e3	2.5e2	3.0e2	1.3e2
ESP Game	3.4e2	5.8e3	1.0e3	8.6e2	1.3e4	6.7e2	1.3e3	3.5e2
IAPR TC12	5.2e2	1.2e4	1.7e3	1.6e3	1.6e4	1.1e3	1.5e3	5.2e2
NUS-WIDE	6.8e3	2.9e4	1.8e3	2.6e3	2.8e4	1.5e3	3.8e3	1.2e3

Table 4.2: Running time (seconds) for tag completion baselines. All algorithms are run in Matlab on an AMD 4-core @2.7GHz and 64GB RAM machine.

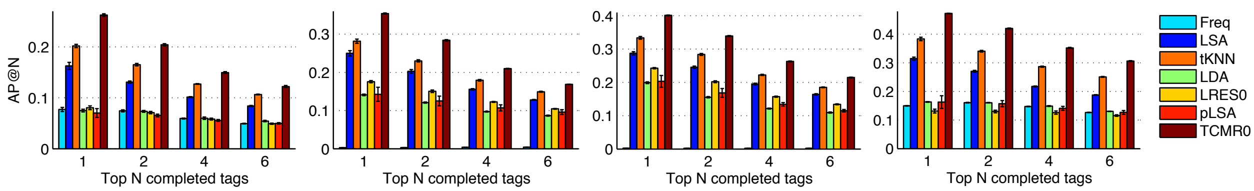
Table 4.7.2.1 summarizes the running time of all algorithms in comparison. We observe that although TCMR is not as efficient as several baselines, it is more efficient than LSR which yields similar performance as TCMR in multiple cases. The high computational cost of LSR is due to the fact that it has to train a different model for each instance, which does not scale well to large datasets.

4.7.3 Analysis of Algorithm Design

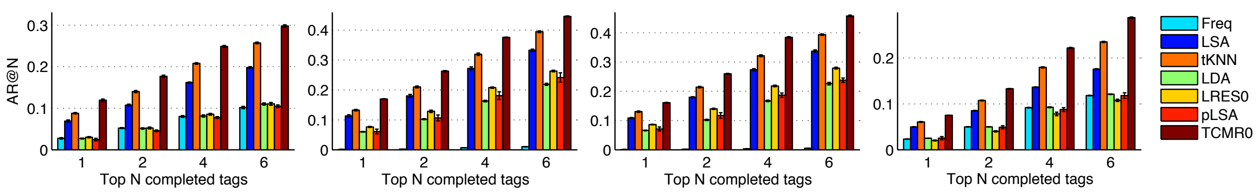
4.7.3.1 Evaluation of Noisy Matrix Recovery without Visual Features

The key component of the proposed approach is a noisy matrix recovery framework. To independently evaluate the effectiveness of noisy matrix recovery component proposed in this work, we simplify TCMR by ignoring the Graph Laplacian component according to 4.4

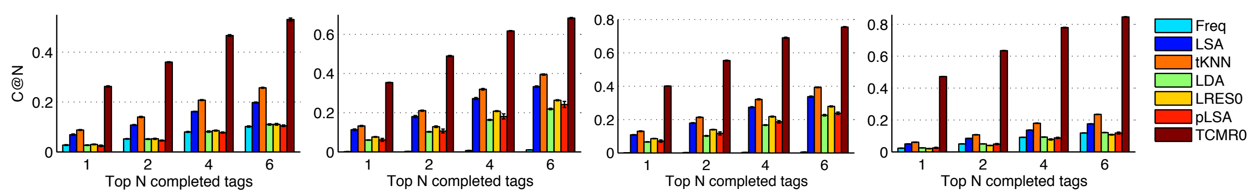
and compare it (denoted as TCMR0) to several baseline approaches for matrix completion that do not take into account visual features: (1) Freq, which assigns the most frequent tags to all the images, (2) LSA [147], Latent Semantic Analysis, (3) tKNN, majority voting among the nearest neighbors in the tag space, (4) LDA [12], (5) LRES0 [224], a version of LRES algorithm without using visual features, and (6) pLSA, probabilistic LSA.



(a) Mir Flickr (b) ESP Game (c) IAPR TC12 (d) NUSWIDE



(e) Mir Flickr (f) ESP Game (g) IAPR TC12 (h) NUSWIDE



(i) Mir Flickr (j) ESP Game (k) IAPR TC12 (l) NUSWIDE

Figure 4.4: Comparison of different topic models and matrix completion algorithms without taking into account the visual feature. The top row is evaluated by $AP@N$, the middle row is by $AR@N$, and the bottom row is by $C@N$.

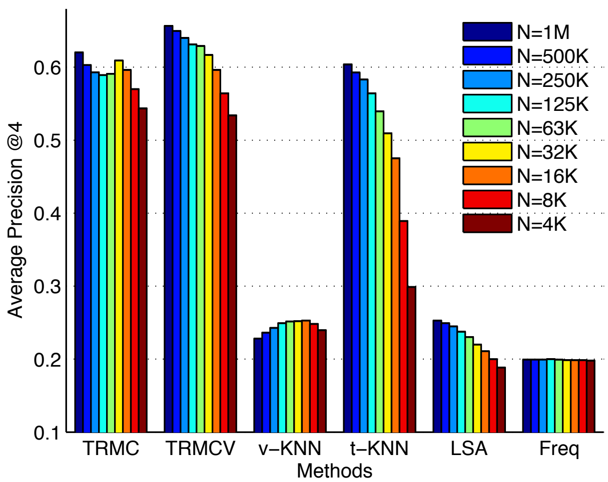
Figure 4.4 compares the tag completion performance of algorithms without visual features. We observe that the proposed noisy matrix recovery algorithm performs significantly better than the other baseline methods, implying that it can successfully capture the important dependency among tags. We also observe that a simple tKNN algorithm works better than the topical models (LSA, LDA and pLSA), suggesting that directly applying a topical model may not be appropriate for the tag completion problem.

Figures 4.2, 4.3 and 4.4 show that TMC and RKML perform much worse than the other algorithms in comparison, while LSA and tKNN perform quite well. Accordingly, we exclude TMC and RKML, and include LSA and tKNN in the following evaluation cases.

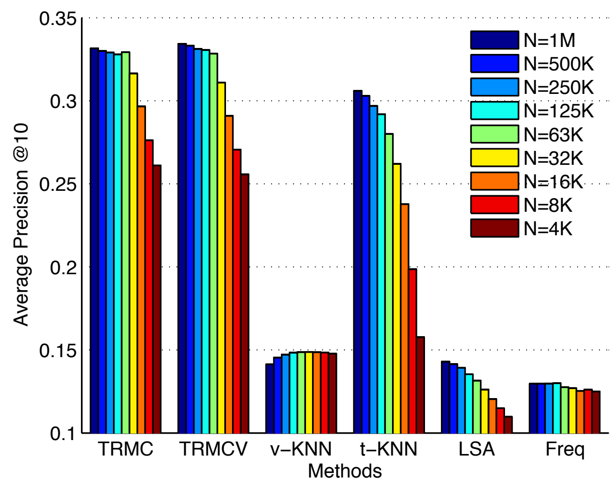
4.7.3.2 Analysis of Scalability

Scalability is a crucial problem ubiquitously presenting in Machine Learning and Computer Vision domains including tag completion, annotation, image understanding, etc. In order to identify how the proposed TCMR algorithm is sensitive to the data size, we conduct the comparison experiments on a sequence of subsets of ImageNet dataset whose scales varies from $4,000$ to $1,000,000$. Since the maximum data size is up to $1M$, some baseline algorithms compared in Section 4.7.2 are unable to implement due to the efficiency issue, so in this Section we only compare the proposed TCMR method with a few fast algorithms including (1) Freq, which assigns the most frequent tags to all the images, (2) LSA [147], Latent Semantic Analysis, (3) t-KNN, majority voting among the nearest neighbors in the tag space, and (4) v-KNN [123], a nearest neighbor voting algorithm based on the visual similarity. Since the computation of Graph Laplacian is extremely expensive over large-scale data, we replace it with more efficient strategy. We use TCMR to represent the algorithm proposed in 4.4, and TCMRV to denote its extension that incorporates the visual information by linear reconstruction following Section 4.6.1.2.

Figure 4.5 evaluates the scalability in terms of tag completion accuracy. We observe that for most methods, as the data size increases, the average precision accordingly increases. However, compared to t-KNN, TCMR based algorithms have a much impressive performance when the dataset is small (less than $63K$), indicating that TCMR is much more capable to recover the tag information with fewer samples. And when the data size exceeds $63K$, the



(a) $AP@4$



(b) $AP@10$

Figure 4.5: Scalability analysis over large-scale dataset ImageNet in terms of tagging precision. Metric $AP@N$ is used for evaluation. The size of evaluated subset N varies from $4K$ to $1M$.

accuracy curve tends to be matured and the performance moderately improves as the data size increases, implying that though there is a large redundancy as the data set enlarges, it is still helpful to explore large-scale dataset.

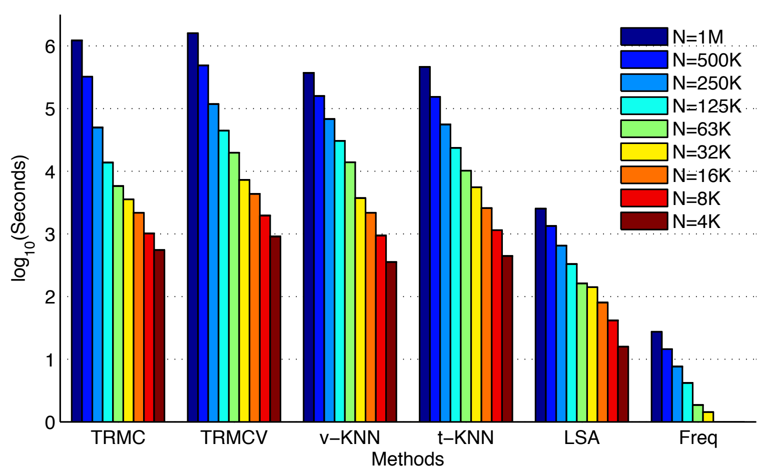


Figure 4.6: Scalability analysis over large-scale dataset ImageNet in terms of implementation time ($\log_{10}$(seconds)). The size of evaluated subset N varies from $4K$ to $1M$.

Figure 4.6 shows that all the algorithms in comparison show similar scalability in terms of the data size, *i.e.*, the time costed for implementation increases exponentially as the data size increases.

4.7.3.3 Evaluation on Various Types of Regularizer

We attribute the success of the proposed TCMR algorithm mostly to the nuclear norm regularizer that simultaneously explores the interaction between both images and tags. To verify this point, we conduct experiments that replace the nuclear norm regularizer $|Q|_{tr}$ in (4.14) with ℓ_1 norm $|Q|_1$ or Frobenius norm $|Q|_F$ regularizers. After that, since neither newly constructed optimization problem has a closed form solution, we use gradient descent method [177, 188] to solve both the non-smooth ℓ_1 and Frobenius norm optimization problem.

Regularizer	ℓ_1			Frobenius			Nuclear		
	AP@1	AP@3	time	AP@1	AP@3	time	AP@1	AP@3	time
MirFlickr	8.1 ± 0.3	5.7 ± 0.1	5.7e2	8.7 ± 0.3	6.4 ± 0.2	3.8e1	28 ± 0.6	19 ± 0.4	1.3e2
ESP Game	18 ± 0.3	13 ± 0.2	6.5e2	19 ± 0.4	14 ± 0.2	2.3e2	37 ± 0.5	25 ± 0.1	3.5e2
IAPR TC12	37 ± 0.3	27 ± 0.1	5.7e2	37 ± 0.3	27 ± 0.1	2.2e2	47 ± 0.3	33 ± 0.3	5.2e2
NUS-WIDE	17 ± 0.4	14 ± 0.2	3.9e3	17 ± 0.4	14 ± 0.2	2.1e2	48 ± 0.2	39 ± 0.2	1.2e3

Table 4.3: Comparison of tag completion performance between TCMR and its counterparts with different regularizers, evaluated by accuracy (%) and efficiency/running time (s).

Table 4.7.3.3 summarizes both the accuracy and efficiency performances of TCMR and its counterparts with the other types of regularizer. From it, we observe that ℓ_1 norm and Frobenius norm regularization give sparse estimates and greatly reduce the computation time, especially on large scale datasets. However, the nuclear norm overwhelmingly outperforms its counterparts since it enforces both the row-wise and column-wise interaction of the tag matrix, while ℓ_1 and Frobenius norms treat each entry independently.

4.7.3.4 Evaluation on Various Loss Functions

Besides, we also compare the proposed maximum likelihood loss function with a couple of popular loss functions in matrix completion work [167], including the absolute (ℓ_1 norm) loss, least square (Frobenius norm) loss, hinge loss and logistic loss.

Loss functions	[1]		[2]		[3]		[4]		[5]	
	AP@1	AP@3	AP@1	AP@3	AP@1	AP@3	AP@1	AP@3	AP@1	AP@3
MirFlickr	22.8	15.1	28.1	18.7	25.5	17.0	28.2	18.7	28.3	18.8
ESP Game	31.1	22.4	37.0	24.8	31.9	22.8	37.0	24.7	37.1	24.9
IAPR TC12	43.6	32.2	45.7	33.3	44.9	32.9	45.9	32.8	47.4	33.5
NUS-WIDE	39.1	32.8	45.9	36.3	43.3	34.8	47.0	37.4	48.3	38.6

Table 4.4: Comparison of tag completion accuracy (%) between TCMR and its counterparts with different loss functions. Standard deviation is omitted for simplicity. [1] to [5] represent absolute, least square, hinge, logistic and maximum likelihood loss functions, respectively.

Loss function	Absolute	Least square	Hinge	Logistic	Likelihood
MirFlickr	3.65e+01	6.84e+03	8.09e+02	7.53e+03	1.26e+02
ESP Game	3.52e+01	1.82e+03	5.29e+02	5.98e+03	3.50e+02
IAPR TC12	9.10e+01	5.83e+03	8.35e+03	1.47e+04	5.16e+02
NUS-WIDE	1.38e+02	2.21e+04	2.76e+03	2.06e+04	1.22e+03

Table 4.5: Comparison of tag completion efficiency (running time in second) between TCMR and its counterparts with different loss functions.

Table 4.4 and 4.5 show the tag completion performance of TCMR and its counterparts with different loss functions in terms of accuracy and efficiency. We observe that from the viewpoint of tag completion accuracy, the proposed maximum likelihood loss function significantly outperform the other loss functions, especially when the size of the dataset is large. From the viewpoint of efficiency, absolute loss and hinge loss are much faster than the other three ones, but their tag completion accuracies are significantly worse. Logistic loss function performs a bit worse than maximum likelihood loss, it however takes pretty

Algorithm	TCMR-lr			TCMR		
	AP@1	AP@3	time (s)	AP@1	AP@3	time (s)
MirFlickr	26.4 ± 0.4	17.4 ± 0.3	9.7e+1	28.3 ± 0.6	18.8 ± 0.4	1.3e+2
ESP Game	37.6 ± 0.4	25.0 ± 0.1	2.1e+2	37.1 ± 0.5	24.9 ± 0.1	3.5e+2
IAPR TC12	47.3 ± 0.5	33.6 ± 0.2	2.7e+2	47.4 ± 0.3	33.5 ± 0.3	5.2e+2
NUS-WIDE	48.3 ± 0.3	39.0 ± 0.2	1.9e+2	48.3 ± 0.2	38.6 ± 0.2	1.2e+3

Table 4.6: Performance comparison of TCMR and TCMR-lr, in terms of both accuracy (%) and running time (s).

much more computation time, which demonstrates that proposed maximum likelihood loss function is the optimal solution which makes a good compromise between the accuracy and efficiency.

From Section 4.7.3.4 and 4.7.3.3 we can easily see the reasons why we choose the combination of maximum likelihood loss and nuclear norm regularizer, which yields superior tag completion accuracy yet remains efficient in computation.

4.7.3.5 Efficient Extension of TCMR by Linear Reconstruction

We use TCMR-lr to represent the algorithm proposed in (4.15), which reconstructs the tag matrix by linearly combining the noisy matrix recovery results and the nearest neighbor voting results. The efficiency bottleneck of TCMR and TCMR-lr is to solve the optimization problems in (4.4) and (4.14) that consist of the nuclear norm. However, (4.4) is much faster because the computation of term (4.12) and its gradient is quite time consuming, which reduces the updating speed in (4.16). Table 4.6 shows that under the same experimental setup, TCMR-lr achieves similar tag completion accuracy as TCMR while it takes much less computational time.

From Table 4.6, we observe that TCMR-lr achieves almost similar tag completion performance as TCMR in terms of accuracy, under proper conditions with sufficient tag and

Noise ratio	IAPR TC12 0.7	0.9	NUS-WIDE 0.2	0.3	0.5	0.7	0.9
TCMR-lr	27 ± 0.6	6.8 ± 0.5	28 ± 0.4	25 ± 0.2	16 ± 0.1	8.2 ± 0.1	1.5 ± 0.1
TCMR	28 ± 0.7	19 ± 1.4	29 ± 0.2	26 ± 0.2	18 ± 0.1	9.7 ± 0.1	5.1 ± 1.0

Table 4.7: Performance comparison of TCMR and TCMR-lr when the observed tags are severely noisy. $AP@1$ is used for evaluation.

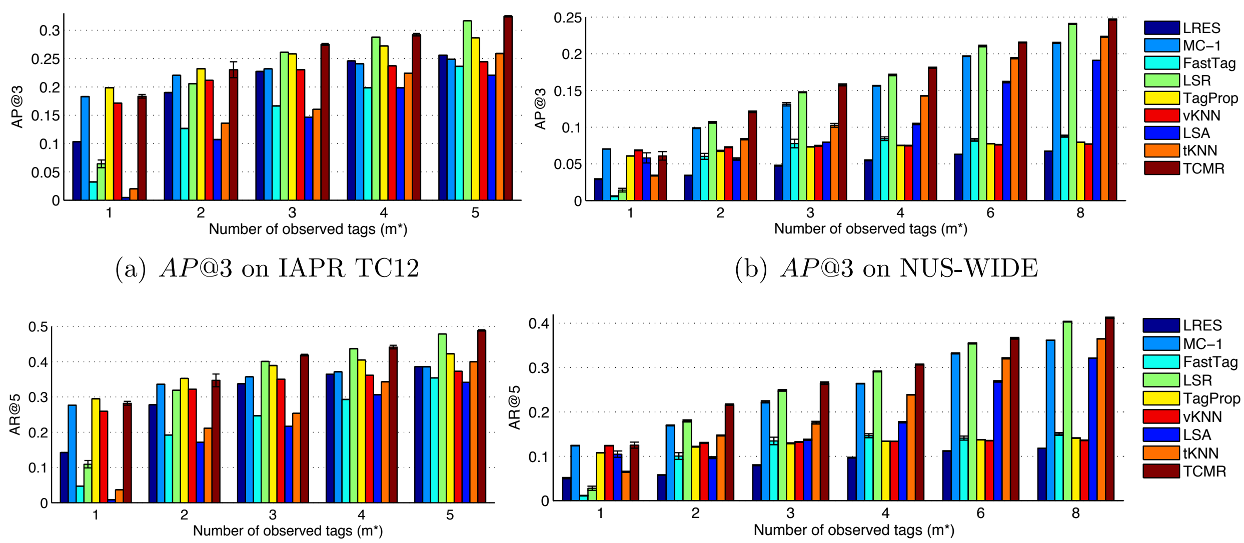
moderate noise level. The essential ideas behind these two TCMR implementations are the same, which enjoy both the topic model based noisy matrix recovery component and the visually nearest neighbor voting scheme. Besides, TCMR-lr is much faster than TCMR. Moreover, as the size of dataset increases, the gap between their accuracy reduces. So for large datasets, we can definitely use TCMR-lr to replace TCMR to speed up the optimization while do not hurt the tag completion accuracy.

Table 4.7 shows the comparison of TCMR-lr and TCMR when they perform significantly different. The experimental setup is described in Section 4.7.4.2. Since TCMR-lr takes the linear combination of two tag matrices from sub-steps, it suffers more under extreme cases when the observed tags are severely corrupted with noise. This is because the interaction/relationship of the two sub-steps are ignored, which prevents finding the global optimal solution for the whole tag completion procedure. So only under certain circumstances with moderate number of noisy observed tag entries, TCMR-lr is a good alternative of TCMR which is able to save much computation time; and when there is too much noise, TCMR is still highly recommended.

4.7.4 Effects on Missing and Noisy Tags

4.7.4.1 Sensitivity to the Number of Observed Tags

We also examine the sensitivity of the proposed TCMR to the number of initially observed tags by comparing it to the baseline algorithms on IAPR TC12 and NUS-WIDE datasets. To make a meaningful evaluation, we only keep images with 6 or more tags for IAPR TC12 dataset, and images with 9 or more tags for NUS-WIDE dataset. As before, we divide the tags into testing and training sets, and randomly sample m_* tags for each image from the training tag set to create the partially observed tag matrix, where the number of sampled tags m_* is varied. We evaluate the tag completion performance on the testing tag sets.



(a) AP@3 on IAPR TC12

(b) AP@3 on NUS-WIDE

(c) AR@5 on IAPR TC12

(d) AR@5 on NUS-WIDE

Figure 4.7: Tag completion performance with varied number of observed tags, evaluated by AP@3 (top row) and AP@5 (bottom row). IAPR TC12 is a clean and complete dataset while NUS-WIDE contains missing and noisy tags.

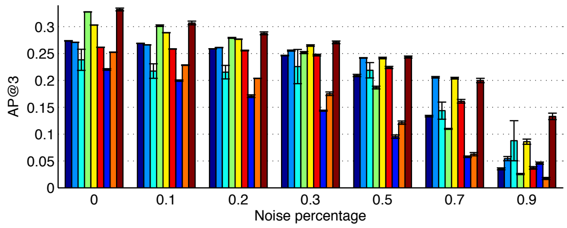
Figure 4.7 shows the influence of the number of partially observed tags to the final tag completion performance measured by AP@3 and AR@5. We observe that the performance of all algorithms improves with increasing number of observed tags. We also observe that when the number of observed tags is 3 or larger, TCMR and LSR perform significantly better than the other baseline approaches. When the number of observed tags is small (*i.e.* 1 or 2), TCMR performs significantly better than LSR, indicating that the proposed algorithm is noticeably effective even when the number of observed tags is small.

Besides, some algorithms (TCMR, LSR, MC-1, tKNN and LSA) perform similar on both IAPR TC12 and NUS-WIDE dataset, *i.e.*, the tag completion performance increases gradually as the number of observed tags increases. However, under the same experimental setting, the other algorithms (LRES, FastTag, TagProp and vKNN) improve significantly on IAPR TC12 dataset but improve slightly on NUS-WIDE. This might because IAPR TC12 is a clean dataset containing substantially complete while NUS-WIDE is a raw dataset consisting of pretty incomplete and noisy tags. This phenomenon indicates that the first group of algorithms is capable to explore the valid observed tag information even when they come with noise, , *i.e.*, they somehow explores the interaction between images or tags to dilute the impact of noisy tags.

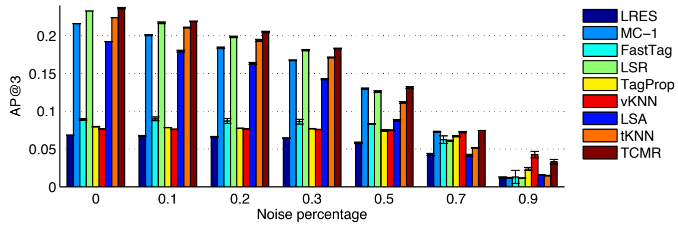
4.7.4.2 Sensitivity to Noise

To evaluate the sensitivity to noise, we conduct experiments with noisy observed tags on datasets IAPR TC12 and NUS-WIDE. To generate noisy tags, we replace some of the sampled tags with the incorrect ones that are chosen uniformly at random from the vocabulary. The percentage of noisy tags among the total observed ones in the whole gallery is varied from 0 to 0.9. To ensure there are a sufficient number of noisy tags as well as sufficient number of images, we set m_*, the number of sampled tags, to be 8 for NUS-WIDE dataset and to be 4 for IAPR TC12 dataset in this experiment.

Figure. 4.8 shows the tag completion performance for different algorithms using noisy



(a) IAPR TC12 $AP@3$



(b) NUS-WIDE $AP@3$

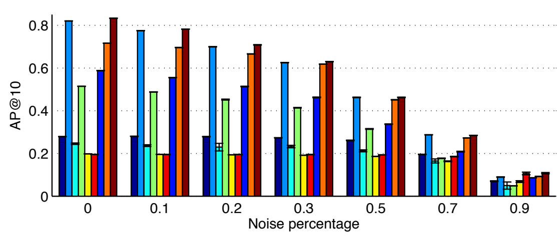
Figure 4.8: Comparison of tag completion performance ($AP@3$)using noisy observed tags.

observed tags. It is not surprising to observe that the performance of all algorithms in comparison degrades with increasing amounts of noise. We also observe that LSR seems to be significantly sensitive to the noise in the observed tags than the proposed TCMR algorithm. In particular, we find that TCMR outperforms LSR significantly when the percentage of noisy tags is large. The contrast is particularly obvious for the IAPR TC12 dataset, where LSR starts to perform worse than several other baselines when the noise level is above 50%. Besides, all algorithms reduce their performance dramatically as the noise level increases from 70% to 90%. This is not surprising because at the 90% noise level, a number of images do not have accurate observed tags for training the model, especially for the NUS-WIDE dataset whose originally assigned tags are pretty noisy. However, the proposed TCMR algorithm is overwhelmingly better in this case, especially on IAPR TC12, indicating that it is more powerful in recovering expected tags from severely noisy tagged images. Table 4.8 shows the tag completion results of exemplar images by different algorithms, where both partially true and noisy tags are observed.

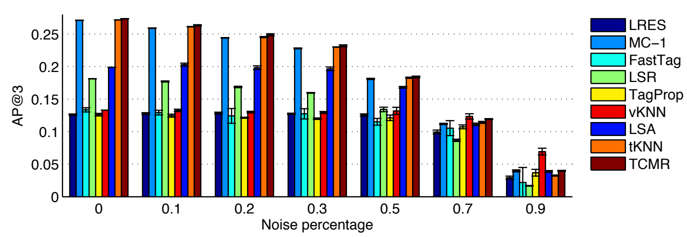
4.7.5 Effects on Other Tag-relevant Applications

To evaluate the robustness of the proposed TCMR algorithm on image tagging tasks, we compare it to the baseline algorithms on tag ranking and tag refinement tasks. Compared with tag completion, these two tasks require more initially observed tags. Besides, among the four used datasets, only NUS-WIDE has manually annotated tags, which are regarded as the true relevant tags in the evaluation phase. So in order to make the evaluation statistically meaningful, we do the evaluation on NUS-WIDE with the number of observed tags increased from $m_* = 4$ to $m_* = 8$.

We first randomly sample m_* tags for each image to create the training tag set $\mathcal{T}_{tr}$. And then the observed tag matrix is generated from $\mathcal{T}_{tr}$ by randomly adding certain number of noisy tags while removing the same number of originally associated tags for each image. The percentage of noisy tags out of total observed tags varies from 0 to 0.9. Denote the total initially assigned tags as set $\mathcal{T}_k$, and the manually labeled tag set of NUS-WIDE as $\mathcal{T}_f$. The tagging performances for tag ranking and tag refinement tasks are evaluated on testing set $\mathcal{T}_k$ and $\mathcal{T}_f$, respectively.



(a) NUS-WIDE, Tag ranking, with AP@10



(b) NUS-WIDE, Tag refinement, with AP@3

Figure 4.9: Comparison between TCMR and baseline algorithms with varied percentage of noisy tags in terms of other two tag relevant applications, including tag ranking and tag refinement. The counterpart performance of tag completion can be referred to Figure4.8(b).

Figures 4.8(b) and 4.9 show the impact of noisy tags to the final accuracy for these three tasks, including tag completion, tag ranking and tag refinement, measured by AP@3, AP@10 and AP@3, respectively. We observe that the performance of all algorithms degrades with increasing noise percentage. We also observe that when the noise percentage is low, although some baselines yield similar performance as TCMR for certain tasks($i.e$, LSR for tag completion, MC-1 for tag ranking and tKNN for tag refinement), TCMR significantly outperforms them on other tasks, which means TCMR is more robust to incomplete and noisy tags than the baseline algorithms on these three image tagging tasks.

4.8 Summary

In this section, we have proposed a robust yet efficient image tag completion algorithm (TCMR), which is capable of simultaneously fill in the missing tags and remove/down weight the noisy tags. TCMR introduces a noisy matrix recovery procedure that captures the underlying interaction among tags by enforcing the recovered tag matrix to be of low rank. Besides, a graph Laplacian based on the image visual features is also incorporated to ensure the recovered tag matrix is consistent with the visual contents of images. Experiments over five different scaled image datasets with size up to 1M demonstrate the effectiveness and efficiency of the proposed TCMR algorithm by comparing it to state-of-the-art tag completion approaches. In the future, we plan to improve the tag completion performance by incorporating the visual features more effectively, and adopting more efficient nuclear norm optimization procedure.

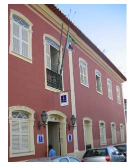
						
Ground truth	building, front, group, people, palm, lawn, tree, square, statue	boy, cap, hair, house, power, pole, roof, sky, shirt, sweater, terrain, tree	bank, bush, helmet, jacket, life, people, river, rock, tree	balcony, door, entrance, car, flag, front, lamp, house, sky, window	bed, brick, curtain, leg, man, short, sweater, wall, woman	church, llama, meadow, range, mountain, roof, tourist, tower, train
Observ-ed tags	lawn, people, square, *cloud*	cap, terrain, sky, *meadow*	life, river tree, *llama*	balcony, car, window, *water*	curtain, wall, *floor*, *team*	church, range, train, *lawn*
LSER	**people**, bike, wall, cloud, **square**, **tree**, house, **lawn**, **palm**	**terrain**, **sky**, **hair**, **sweater**, **roof**, mountain, wall, meadow, **cap**, trouser	**tree**, **river**, **life**, **helmet**, **rock**, woman llama, **jacket**, gravel, **people**	**entrance**, **car**, **front**, **balcony**, water, **window**, building, people, harbour, **sky**	**woman**, **wall**, table, room, hand, **curtain**, floor, team, person, front	wall, **tourist**, people, house, **range**, lawn, **church**, **train**, grave, child
MC-1	**people**, **square**, cloud, **lawn**, **tree**, sky, **building**, **front**, wall	**sky**, meadow, **terrain**, **cap**, wall, mountain, man, **house**, woman, **hair**	**tree**, **river**, **life**, man, llama, wall, **people**, front mountain, sky	**window**, **car**, **balcony**, water, man, **front**, building, wall, **house**, woman	**wall**, **curtain**, floor, team, window, room, **man**, table, front, **bed**	**range**, lawn, **church**, **train**, front, **mountain**, wall, people, **tourist**, man
FastTag	**tree**, tourist, footpath, shirt, river, **group**, woman, tile **people**	wall, **boy**, desk, meadow, mountain, girl, **hair**, tee-shirt, plane, fence	**life**, mountain, **people**, front, tourist, railing, **river**, llama, **tree**, wall	building, **front**, **house**, **car**, grey, **window**, rail, **balcony**, street, photo	**wall**, room, table, window, **bed**, **curtain**, hand, night, cup, towel	**tourist**, front, wall, **mountain**, classroom, house, body, fjord, square, tile
LSR	sky, **square**, **building**, **people**, **tree**, house, **lawn**, street, cloud	**house**, **sky**, hill, **boy**, grey, jacket, **tree**, **terrain**, cloud, landscape	**bank**, **jacket**, **river**, **helmet**, **bush**, tourist, boat, mountain, **tree**, **people**	**front**, building, **house**, wall, **sky**, cliff, **door**, **window**, street, man	**wall**, room, window, front, uniform, **bed**, table, jersey, **short**, round	**mountain**, view, **tower**, woman, people, **roof**, square, street, snow, park
TagProp	**people**, **tree**, **square**, house, **front**, wall, tourist, man, woman	wall, woman, man, **sky**, front, **sweater**, **hair**, mountain, table, desert	**people**, **tree**, woman, front, man, **rock**, wall, **river**, sky, mountain	wall, **front**, man, building, woman, table, people, **house**, **sky**, **entrance**	front, **woman**, **wall**, table, **man**, house, room, people, tree, window	people, wall, **tourist**, **mountain**, front, man, table, woman, tree, square
vKNN	**tree**, wall, house, **people**, sky, woman, bike, **front**, **square**	**sweater**, desert, **sky**, landscape, **terrain**, **hair**, mountain, wall, cloud, front	**people**, **tree**, **helmet**, front, **river**, **bush**, woman, **life**, sky, man	**front**, building, people, **house**, **entrance**, **sky**, wall, **balcony**, tree, **window**	room, **woman**, table, front, house, **wall**, **man**, chair, window, child	**tourist**, people, wall, table, house, square, **mountain**, tree, hill, lawn
LSA	**people**, cloud, **square**, roof, **group**, meadow, **building**, tower, landscape	**sky**, meadow, cloud, **hair**, **roof**, road, short, **tree**, woman, **boy**	**tree**, **bush**, lake, palm, meadow, **river**, tourist, slope, building, grass	**car**, **window**, street, **house**, building, room, **lamp**, **front**, bed, bush	**wall**, room, table, **bed**, window, hair, girl, wood, boy, **curtain**	**mountain**, building, **range**, people, snow, tree, house, street, city, wall
tKNN	**people**, **square**, cloud, **lawn**, sky, **tree**, mountain, street, **building**	**sky**, meadow, **terrain**, **cap**, people, cloud, hill, mountain, road, **tree**	**tree**, **river**, **life**, **bush**, house, sky, building, man, **people**, **bank**	**window**, **car**, **balcony**, wall, **house**, **front**, building, bed, room, curtain	**wall**, floor, **curtain**, room, **bed**, front, window, girl, team, **brick**	**range**, **church**, **mountain**, view, lawn, **train**, front, snow, landscape, column
TCMR	**people**, **square**, **lawn**, sky, **building**, **tree**, cloud, street, **palm**	**sky**, **terrain**, **cap**, **boy**, hill, **house**, **hair**, landscape, **sweater**, cloud	**tree**, **river**, **life**, boat, **jacket**, **bank**, llama, **helmet**, **rock**, mountain	**car**, **window**, **balcony**, **door**, building, wall, **front**, **house**, water, **sky**	**wall**, floor, **curtain**, **bed**, **brick**, room, window, front, table, team,	**range**, lawn, **church**, **train**, **mountain**, people, **tower**, square, view, street

Table 4.8: Examples of tag completion results generated by some baseline algorithms and the proposed TCMR. The observed tags in red italic font are noisy tags, and others are randomly sampled from the ground truth tags. The completed tags are ranked based on the recovered scores in descending order, and the correct ones are highlighted in blue bold font.

Chapter 5

Summary and Conclusions

In this dissertation, we designed two algorithms for image tag completion on large scale datasets where the observed tags might be incomplete and corrupted with noise. The proposed algorithms, namely, RKML and TCMR, achieve the ultimate task around two questions including

- How can we find better neighbors (visually similar images) for a given image?
- How can we maximally exploit the hints behind the given tags?

5.1 Contributions

This dissertation mainly answers the two questions raised above by proposing specific algorithms as follows, giving theoretical guarantees and providing empirical comparisons with state-of-the-art baseline algorithms.

5.1.1 Image Annotation by Kernel Metric Learning

The RKML (short for **Regression based Kernel Metric Learning**) algorithm presented in Chapter 3 is a distance metric learning algorithm designed for search based image annotation. It answers the first question and addresses a couple of challenges commonly existing in kernel metric learning, in terms of both theory and real-world application. The main contributions of the proposed RKML can be concluded as follows.

- **Provide a kernel metric learning with theoretical guarantee**. We demonstrate the robustness of RKML in the high dimensional kernel space by proving the theoretical guarantees of the learned kernel metric for the first time.

- **Efficient metric computation**. The PSD property is automatically guaranteed by the special property of regression and thus no need to take extra projections, and Nyström approximation [43] is applied to avoid the direct computation based on the full kernel. Those actions greatly improve the metric computational efficiency.

- **Effective metric for image annotation**. The notorious overfitting risk is alleviated by a rank regularizer of the learned kernel metric. Besides, image tags are directly utilized to compute numeric semantic similarities, which make better use of the tag information and substantially promote the image annotation performance.

5.1.2 Image Tag Completion by Noisy Matrix Recovery

The TCMR (short for **tag completion by noisy matrix recovery**) algorithm presented in Chapter 4 is a noisy matrix completion based algorithm designed for image tag completion problem. It answers the second question raised at the beginning of this chapter and addresses the challenges of applying noisy matrix completion theory to practical image tag completion tasks. The main contributions of the proposed TCMR are summarized as follows.

- **Incorporate noisy matrix recovery theory to image tag completion with theoretical guarantees**. Low rank noisy matrix recovery is achieved by nuclear norm with minimization, leading to the success in filling out missing tags while down-weighting noisy ones. Both theoretical support and empirical evaluation are provided.

- **Propose a convex optimization problem based on topic model**. Although inspired by the idea of topic models, unlike them the proposed TCMR solves a convex optimization problem, leading to a more efficient optimization procedure and avoiding the estimation of a bunch of hyper-parameters.

- **Exploit fully the image visual contents**. TCMR improves the tag completion performance by exploiting the statistical dependence between image features and tags via a graph Laplacian [224, 226], which reduces the impact of incomplete and noisy tags by keeping the recovered tag matrix consistent with image visual features.

- **Apply to multiple tag relevant tasks**. TCMR has been successfully applied to multiple tag relevant tasks including tag completion, tag ranking and tag refinement under the defective scenario. Extensive experiments on benchmark datasets show that TCMR is more robust to incomplete and noisy tags than the baseline algorithms on these three image tagging tasks.

5.2 Future Work

The studies presented in the thesis lead to several important research questions, which we plan to investigate in the coming months and will appear in the final version of the dissertation.

- **Improve the efficiency of TCMR on large scale datasets**. Although TCMR has a quite good tag completion performance in terms of both effectiveness and efficiency, it is currently not able to well deal with large scale datasets due to the computational cost. The bottleneck is the nuclear norm optimization. The state-of-the-art solution to optimize the nuclear norm is iterative schemes, where each iteration involves a SVD decomposition and it usually takes more than 100 iterations to converge for a $1M \times 2K$ matrix. Through our experiments, we found that when the data size is sufficient large (larger than 60K), as the data size increases, the tag completion performance improves marginally and insignificantly, which indicates there is much redundant information. By taking advantage of this point, it is possible to integrate the underlying idea of RKML, random sampling, to optimize the nuclear norm solution in large-scale problem.

- **Apply numeric tag information**. Currently, both RKML and TCMR are using binary tag information. However there actually are much numeric tag information that reflects the confidence score when assigning a tag to an image, where missing tags are usually with small scores and noisy tags are mainly with the ambiguous ones. So we next plan to explore this more specific information to improve our image tag completion algorithms.

- **Explore a distance metric learning adaptive to incomplete and noisy tags**. Distance metrics are usually learned from supervised information that is assumed to be perfect in traditional problems like classification and clustering. However, in image tagging problem, this assumption is no longer true. Each image can be associated with multiple tags, and among them some are incorrect or irrelevant to the visual content. And also some tag associations might be missing for some reasons. In this situation, how obtain a valid and effective distance metric turns to be meaningful and profitable. So we plan to extend our work of RKML and make it adaptive to images with missing and noisy tags. Matrix completion technique will be used to supplementary capture the tag-tag correlation.

- **Introduce deep convolutional neural network**. As currently most hottest topic in Machine Learning and Computer Vision, deep convolutional neural network has been proved to be efficient and significantly effective by vast literature [34, 35, 109, 121]. In the following years, I plan to explore more about the deep learning and try to apply it to image tagging problems

5.3 Conclusions

This dissertation answers the two questions raised in the beginning of this chapter by presenting two new image tag completion algorithms for large scale datasets where the observed tags might be incomplete and noisy. To assign appropriate tags to each image, two effective and efficient image tagging models are embedded into the proposed algorithms, including kernel metric learning among images and image tag completion by noisy matrix recovery.

The concluded research makes significant contributions to (i) the theoretical foundations of exploring the image-image correlation and tag-tag interaction, (ii) the challenges of kernel metric learning, (iii) the difficulty of coupling topic model and noisy matrix recovery, and (iv) the empirical applications and implementations to large scale image data. Algorithms presented here advance the state-of-the-art of image annotation and image tag completion works. Several future research directions of new image tagging relevant algorithms are also identified.

APPENDIX

Appendix

Technical Backgrounds

A.1 Low Rank Matrix Approximation

In this appendix, we give the proofs of the detailed supporting theorems of low rank matrix approximation for Section 3.5.

A.1.1 Proof of Theorem 3.2

Proof. Let $(\lambda_i, \mathbf{u}_i), i = 1, \ldots, n$ be the eigenvalues and eigenvectors of K. Define $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$. According to [175], the eigenfunctions of L_n is given by

$$\widehat{\varphi}_i(\cdot) = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{n} U_{j,i} \kappa(\mathbf{x}_j, \cdot).$$

We therefore have

$$\begin{aligned}
& \sum_{i=1}^{r} \widehat{\varphi}_i(\cdot) \langle \widehat{\varphi}_i(\cdot), g_k(\cdot) \rangle_{\mathcal{H}_\kappa} \\
= & \sum_{i=1}^{r} \sum_{a,b=1}^{n} \frac{1}{\lambda_i} \kappa(\mathbf{x}_a, \cdot) \langle \kappa(\mathbf{x}_b, \cdot), g_k(\cdot) \rangle_{\mathcal{H}_\kappa} U_{a,i} U_{b,i} \\
= & \sum_{i=1}^{r} \sum_{a=1} \kappa(\mathbf{x}_a, \cdot) \frac{1}{\lambda_i} U_{a,i} U_{b,i} Y_{b,k} = \sum_{i=1}^{r} \sum_{a=1} \kappa(\mathbf{x}_a, \cdot) \frac{1}{\lambda_i} U_{a,i} U_{*,i}^{\top} \mathbf{y}^k \\
= & \sum_{a=1}^{n} \kappa(\mathbf{x}_a, \cdot) [U_r \Sigma_r^{-1} U_r \mathbf{y}^k]_i = \sum_{a=1}^{n} \kappa(\mathbf{x}_a, \cdot) [K_r^{-1} \mathbf{y}^k]_i.
\end{aligned}$$

$\square$

A.1.2 Proof of Theorem 3.3

Proof. Define a linear operator G as

$$G[f] = \sum_{k=1}^{m} g_k(\cdot)\langle g_k, f\rangle_{\mathcal{H}_K}.$$

Define two projection operator $\widehat{P}$ and P as

$$\widehat{P}[f] = \sum_{i=1}^{r} \widehat{\varphi}_i(\cdot)\langle \widehat{\varphi}_i(\cdot), f(\cdot)\rangle_{\mathcal{H}_K}, \quad P[f] = \sum_{i=1}^{r} \varphi_i(\cdot)\langle \varphi_i(\cdot), f(\cdot)\rangle_{\mathcal{H}_K}.$$

Using G, $\widehat{P}$ and P, we write $\widehat{T}$ and T_* as

$$\widehat{T} = \widehat{P}G\widehat{P}, \quad T_* = PGP.$$

Using the $\sin\Theta$ theorem [181], we have

$$|\widehat{P} - P| \leq \frac{|L - L_n|_2}{\lambda_r(L_n) - \lambda_{r+1}(L)}.$$

Since $\lambda_r(L_n) = \lambda_r/n$, and $\lambda_{r+1}(L) \leq \lambda_{r+1}(L_n) + |L - L_n|_2$, we have

$$|\widehat{P} - P| \leq \frac{|L - L_n|_2}{(\lambda_r - \lambda_{r+1})/n - |L - L_n|_2}.$$

We complete the proof by using the fact

$$|(\widehat{T}-T)[f]|_{\mathcal{H}_\kappa} \leq |(\widehat{P}-P)G\widehat{P}[f]|_{\mathcal{H}_\kappa} + |PG(\widehat{P}-P)[f]|_{\mathcal{H}_\kappa}.$$

□

A.2 Matrix Completion

In this appendix, we give the proofs of the two supporting lemmas that are used to bound the matrix recovery error for Section 4.5.

A.2.1 Proof of Lemma 4.4

Proof. We have

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|P_{i,j}-Q_{i,j}|^2}{Q_{i,j}} &= \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{|P_{i,j}-Q_{i,j}|^2}{Q_{i,j}}\right)\left(\sum_{i=1}^{j}Q_{i,j}\right)\\ &\geq \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|P_{i,j}-Q_{i,j}|}{\sqrt{Q_{i,j}}}\sqrt{Q_{i,j}} = |P-Q|_1.\end{aligned}$$

□

A.2.2 Proof of Lemma 4.5

Proof. To facilitate our analysis, we rewrite each $\mathbf{d}_i$ as

$$\mathbf{d}_i = \sum_{j=1}^{m_*}\mathbf{d}_i^j,$$

where $\mathbf{d}_i^j$ is the image tag vector corresponding to the j-th word sampling for the tag vector of the i-th image. To utilize Lemma 4.5, we define $Z_{i,j}$ as

$$Z_i = \left(\mathbf{d}_i^j - \mathbf{p}_i\right) \mathbf{e}_i^\top,$$

and therefore

$$M = \frac{1}{m_*} \sum_{i=1}^{n} \sum_{j=1}^{m_*} Z_{i,j}.$$

To bound U in Lemma 4.5, we have

$$|Z_{i,j}|_* \leq \left|\mathbf{d}_i^j - \mathbf{p}_i\right|_2 \leq |\mathbf{d}_i^j|_2 \leq 1.$$

To bound σ_Z, we compute

$$\begin{aligned} &\left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[Z_{i,j} Z_{i,j}^\top\right]\right|_* = \left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[\left(\mathbf{d}_i^j - \mathbf{p}_i\right)\left(\mathbf{d}_i^j - \mathbf{p}_i\right)^\top\right]\right|_* \\ = \quad &\left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[\mathbf{d}_i^j (\mathbf{d}_i^j)^\top\right] - \mathbf{p}_i \mathbf{p}_i^\top\right|_* \leq \max_{1 \leq j \leq m} \tfrac{1}{n} \textstyle\sum_{i=1}^{n} p_{i,j}(i - p_{i,j}^2) = \tfrac{|P\mathbf{1}|_\infty}{n}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} &\left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[Z_i^\top Z_i\right]\right|_* = \left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[\left(\mathbf{d}_i^j - \mathbf{p}_i\right)^\top \left(\mathbf{d}_i^j - \mathbf{p}_i\right) \mathbf{e}_i \mathbf{e}_i^\top\right]\right|_* \\ = \quad &\left|\tfrac{1}{nm_*} \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m_*} \mathrm{E}\left[\left(\mathbf{d}_i^\top \mathbf{d}_i - \mathbf{p}_i^\top \mathbf{p}_i\right) \mathbf{e}_i \mathbf{e}_i^\top\right]\right|_* \leq \tfrac{1}{n}. \end{aligned}$$

We complete the proof by plugging the bounds for U and σ_Z. □

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24. ACM, 2007.

[2] Mahdieh Soleymani Baghshah and Saeed Bagheri Shouraki. Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data. *Pattern Recognition*, 43(8):2982–2992, 2010.

[3] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *JMLR*, 6:937–965, 2005.

[4] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[5] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 408–415. IEEE, 2001.

[6] G. Baudat and Fatiha Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[8] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Similarity learning for provably accurate sparse linear classification. *arXiv preprint arXiv:1206.6476*, 2012.

[9] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2013.

[10] Michael W. Berry and Murray Browne. *Understanding search engines: mathematical modeling and text retrieval*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1999.

[11] David M. Blei. Probabilistic topic models. *Communications of the ACM*, 2012.

[12] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[13] Serhat Selcuk Bucak, Rong Jin, and Anil K. Jain. Multi-label learning with incomplete class assignments. In *CVPR*, pages 2801–2808, 2011.

[14] Ricardo S. Cabral, Fernando De la Torre, João P. Costeira, and Alexandre Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Transactions Pattern Analysis and Machine Intelligence (PAMI)*, 2014.

[15] Ricardo S. Cabral, Fernando D. De la Torre, Joao P. Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. In *NIPS*, pages 190–198. 2011.

[16] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[17] T. Tony Cai and Wenxin Zhou. Matrix completion via max-norm constrained optimization. *CoRR*, abs/1303.0341, 2013.

[18] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[19] Emmanuel J. Candès and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[20] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[21] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

[22] Gustavo Carneiro, Antoni B Chan, Pedro J Moreno, and Nuno Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *PAMI*, 29(3):394–410, 2007.

[23] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[24] Ratthachat Chatpatanasiri, Teesid Korsrilabutr, Pasakorn Tangchanachaianan, and Boonserm Kijsirikul. A new kernelization framework for mahalanobis distance learning algorithms. *Neurocomputing*, 73(10):1570–1579, 2010.

[25] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 11:1109–1135, 2010.

[26] Jianhui Chen, Zheng Zhao, Jieping Ye, and Huan Liu. Nonlinear adaptive distance metric learning for clustering. In *KDD*, 2007.

[27] Lin Chen, Dong Xu, Ivor W. Tsang, and Jiebo Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, 2010.

[28] Minmin Chen, Alice Zheng, and Kilian Q. Weinberger. Fast image tagging. In *ICML*, 2013.

[29] Yudong Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust matrix completion and corrupted columns. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 873–880, 2011.

[30] Li Cheng. Riemannian similarity learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 540–548. JMLR Workshop and Conference Proceedings, May 2013.

[31] Alexander L Chistov and D Yu Grigor'ev. Complexity of quantifier elimination in the theory of algebraically closed fields. In *Mathematical Foundations of Computer Science 1984*, pages 17–31. Springer, 1984.

[32] Radha Chitta, Rong Jin, and Anil K. Jain. Efficient kernel clustering using random fourier features. In *ICDM*, 2012.

[33] Tat-S. Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, 2009.

[34] Dan Ciresan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE, 2012.

[35] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[36] Claudio Cusano, Gianluigi Ciocca, and Raimondo Schettini. Image annotation using svm. In *Electronic Imaging 2004*, pages 330–338. International Society for Optics and Photonics, 2003.

[37] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[38] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[39] Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007.

[40] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pages 48–64, 2014.

[41] Yining Deng, BS Manjunath, Charles Kenney, Michael S Moore, and Hyundoo Shin. An efficient color representation for image retrieval. *Image Processing, IEEE Transactions on*, 10(1):140–147, 2001.

[42] Zhi-Hong Deng, Hongliang Yu, and Yunlun Yang. Image tagging via cross-modal semantic mapping. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1143–1146, New York, NY, USA, 2015. ACM.

[43] Petros Drineas and Michael W Mahoney. On the nyström method for approximating a gram matrix for improved kernel-based learning. *JMLR*, 6:2153–2175, 2005.

[44] Kun Duan, Devi Parikh, David Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[45] Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230, 2004.

[46] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112, London, UK, UK, 2002. Springer-Verlag.

[47] Jianping Fan, Yuli Gao, and Hangzai Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*, 2004.

[48] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645, 2010.

[49] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[50] Songhe Feng, Zheyun Feng, and Rong Jin. Learning to rank image tags with limited training examples. *Image Processing, IEEE Transactions on*, 24(4):1223–1234, 2015.

[51] Zheyun Feng, Songhe Feng, Rong Jin, and Anil K Jain. Image tag completion by noisy matrix recovery. In *Computer Vision–ECCV 2014*, pages 424–438. Springer, 2014.

[52] Zheyun Feng, Rong Jin, and Anil Jain. Large-scale image annotation by efficient and robust kernel metric learning. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1609–1616. IEEE, 2013.

[53] Rina Foygel, Ohad Shamir, Nati Srebro, and Ruslan Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, pages 2133–2141, 2011.

[54] Rina Foygel and Nathan Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In Sham M. Kakade and Ulrike von Luxburg, editors, *COLT*, volume 19 of *JMLR Proceedings*, pages 315–340. JMLR.org, 2011.

[55] Fabio Del Frate, Fabio Pacifici, Giovanni Schiavon, and Chiara Solimini. Use of neural networks for automatic classification from high-resolution images. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(4):800–809, 2007.

[56] George W. Furnas, Caterina Fake, Luis von Ahn, Joshua Schachter, Scott Golder, Kevin Fox, Marc Davis, Cameron Marlow, and Mor Naaman. Why do tagging systems work? In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 36–39, New York, NY, USA, 2006. ACM.

[57] Arvind Ganesh, John Wright, Xiaodong Li, Emmanuel J. Candès, and Yi Ma. Dense error correction for low-rank matrices via principal component pursuit. In *ISIT*, 2010.

[58] Shenghua Gao, Liang-Tien Chia, and I.W. Tsang. Multi-layer group sparse coding – for concurrent image classification and annotation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2809–2816, June 2011.

[59] Shenghua Gao, Zhengxiang Wang, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Automatic image tagging via category label and web data. In *ACM Multimedia*, 2010.

[60] Amir Globerson and Sam T. Roweis. Metric learning by collapsing classes. In *NIPS*, 2005.

[61] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. Transduction with matrix completion: Three birds with one stone. In *Advances in neural information processing systems*, pages 757–765, 2010.

[62] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[63] Jacob Goldberger, Sam T. Roweis, Geoffrey E. Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *NIPS*, 2004.

[64] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[65] Rafael C Gonzalez. *Digital image processing*. Pearson Education India, 2009.

[66] David Gross. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

[67] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image annotation. In *ICCV*, 2009.

[68] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[69] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 902 – 909, jun 2010.

[70] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220. ACM, 2007.

[71] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3386–3393. IEEE, 2012.

[72] Bharath Hariharan, S. V. N. Vishwanathan, and Manik Varma. Large scale Max-Margin Multi-Label classification with prior knowledge about densely correlated labels. In *Proceedings of International Conference on Machine Learning*, 2010.

[73] Tomer Hertz, Aharon-Bar Hillel, and Daphna Weinshall. Boosting margin based distance functions for clustering. In *ICML*, 2004.

[74] Tomer Hertz, Aharon-Bar Hillel, and Daphna Weinshall. Learning a kernel function for classification with small training samples. In *ICML*, 2006.

[75] PS Hiremath and Jagadeesh Pujari. Content based image retrieval using color, texture and shape features. In *Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on*, pages 780–784. IEEE, 2007.

[76] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[77] Steven C. H. Hoi, Wei Liu, and Shih-Fu Chang. Semi-supervised distance metric learning for collaborative image retrieval. In *CVPR*, 2008.

[78] Steven CH Hoi, Wei Liu, Michael R Lyu, and Wei-Ying Ma. Learning distance metrics with contextual constraints for image retrieval. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2072–2078. IEEE, 2006.

[79] Richang Hong, Meng Wang, Yue Gao, Dacheng Tao, Xuelong Li, and Xindong Wu. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *Cybernetics, IEEE Transactions on*, 44(5):669–680, 2014.

[80] Cho-Jui Hsieh and Peder Olsen. Nuclear norm minimization via active subspace selection. In *Proceedings of The 31st International Conference on Machine Learning*, pages 575–583, 2014.

[81] Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

[82] Kaizhu Huang, Rong Jin, Zenglin Xu, and Cheng-Lin Liu. Robust metric learning by smooth optimization. In *UAI 2010, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, July 8-11, 2010*, pages 244–251, 2010.

[83] Kaizhu Huang, Yiming Ying, and Colin Campbell. Gsml: A unified framework for sparse metric learning. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 189–198. IEEE, 2009.

[84] Kaizhu Huang, Yiming Ying, and Colin Campbell. Generalized sparse metric learning with relative comparisons. *Knowledge and Information Systems*, 28(1):25–45, 2011.

[85] Ke Huang and Selin Aviyente. Wavelet feature selection for image classification. *Image Processing, IEEE Transactions on*, 17(9):1709–1720, 2008.

[86] Yinjie Huang, Cong Li, Michael Georgiopoulos, and Georgios C Anagnostopoulos. Reduced-rank local distance metric learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 224–239. Springer, 2013.

[87] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 835–843. Curran Associates, Inc., 2009.

[88] Martin Jaggi, Marek Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 471–478, 2010.

[89] Prateek Jain, Brian Kulis, Jason V Davis, and Inderjit S Dhillon. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research*, 13(1):519–547, 2012.

[90] Prateek Jain, Brian Kulis, and Inderjit S Dhillon. Inductive regularized learning of kernel functions. In *Advances in Neural Information Processing Systems*, pages 946–954, 2010.

[91] Prateek Jain, Brian Kulis, Inderjit S. Dhillon, and Kristen Grauman. Online metric learning and fast similarity search. In *NIPS*, pages 761–768, 2008.

[92] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 119–126, New York, NY, USA, 2003. ACM.

[93] Chuanjun Ji, Xiangdong Zhou, Lan Lin, and Weidong Yang. Labeling images by integrating sparse multiple distance learning and semantic context modeling. In *ECCV*, 2012.

[94] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *ICML*, pages 457–464. ACM, 2009.

[95] Rong Jin, Shijun Wang, and Yang Zhou. Regularized distance metric learning:theory and algorithm. In *NIPS*. 2009.

[96] Charles R Johnson. Matrix completion problems: a survey. In *Proceedings of Symposia in Applied Mathematics*, volume 40, pages 171–198, 1990.

[97] Zoltan Kato and Ting-Chuen Pong. A markov random field image segmentation model for color textured images. *Image and Vision Computing*, 24(10):1103–1114, 2006.

[98] Dor Kedem, Stephen Tyree, Kilian Q. Weinberger, Fei Sha, and Gert Lanckriet. Non-linear metric learning. In *Advances in Neural Information Processing Systems 25*, pages 2582–2590. 2012.

[99] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[100] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, August 2010.

[101] Margaret EI Kipp and D Grant Campbell. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–18, 2006.

[102] Olga Klopp. High dimensional matrix estimation with unknown variance of the noise. February 2012.

[103] Olga Klopp et al. Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5:1161–1183, 2011.

[104] Olga Klopp et al. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[105] Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *arXiv preprint arXiv:1412.8132*, 2014.

[106] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Ecole d'été de probabilités de Saint-Flour. Springer, 2011.

[107] Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[108] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *ACM conference on Recommender systems*, pages 61–68. ACM, 2009.

[109] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[110] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

[111] Gautam Kunapuli and Jude W. Shavlik. Mirror descent for metric learning: A unified approach. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *ECML/PKDD (1)*, volume 7523 of *Lecture Notes in Computer Science*, pages 859–874. Springer, 2012.

[112] Kazuhiro Kuroda and Masafumi Hagiwara. An image retrieval system by impression words and specific object names–iris. *Neurocomputing*, 43(1):259–276, 2002.

[113] Yung kyun Noh, Byoung tak Zhang, and Daniel D. Lee. Generative local metric learning for nearest neighbor classification. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1822–1830. Curran Associates, Inc., 2010.

[114] Marco La Cascia, Saratendu Sethi, and Stan Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *Content-Based Access of Image and Video Libraries, 1998. Proceedings. IEEE Workshop on*, pages 24–28. IEEE, 1998.

[115] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *In CVPR*, 2009.

[116] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[117] Landauer. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007.

[118] Monique Laurent. Matrix completion problems matrix completion problems. In *Encyclopedia of Optimization*, pages 1967–1975. Springer, 2009.

[119] Victor Lavrenko, R Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *Advances in neural information processing systems*, page None, 2003.

[120] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[121] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[122] Kuen-Long Lee and Ling-Hwei Chen. An efficient computation method for the texture browsing descriptor of mpeg-7. *Image and Vision Computing*, 23(5):479–489, 2005.

[123] Xirong Li, Cees G.M. Snoek, and Marcel Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, pages 180–187, New York, NY, USA, 2008. ACM.

[124] Daryl Lim, Gert Lanckriet, and Brian McFee. Robust structural metric learning. In *Proceedings of The 30th International Conference on Machine Learning*, pages 615–623, 2013.

[125] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[126] Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Xiaojun Ye. Image tag completion via image-specific and tag-specific linear sparse reconstructions. *CVPR*, 2013.

[127] Nati Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman. Complexity measures of sign matrices. *Combinatorica*, 27(4):439–463, 2007.

[128] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11(4):467–476, 2002.

[129] Dong Liu, Xian-Sheng Hua, Meng Wang, and Hong-Jiang Zhang. Image retagging. In *Proceedings of the International Conference on Multimedia*, pages 491–500. ACM, 2010.

[130] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *World Wide Web*, pages 351–360, 2009.

[131] Dong Liu, Shuicheng Yan, Xian-Sheng Hua, and Hong-Jiang Zhang. Image retagging using collaborative tag propagation. *IEEE Transactions on Multimedia*, 13(4):702–712, 2011.

[132] Jiakai Liu, Rong Hu, Meihong Wang, Yi Wang, and Edward Y Chang. Web-scale image annotation. In *Advances in Multimedia Information Processing-PCM 2008*, pages 663–674. Springer, 2008.

[133] Weifeng Liu, Dacheng Tao, Jun Cheng, and Yuanyan Tang. Multiview hessian discriminative sparse coding for image annotation. *Computer Vision and Image Understanding*, 118:50–60, 2014.

[134] Xiaobai Liu, Shuicheng Yan, Tat-Seng Chua, and Hai Jin. Image label completion by pursuing contextual decomposability. *TOMCCAP*, pages 21:1–21:20, 2012.

[135] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[136] Nicolas Loeff and Ali Farhadi. Scene discovery by matrix factorization. In *ECCV*, 2008.

[137] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[138] Shiqian Ma, Donald Goldfarb, and Lifeng Chen. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353, 2011.

[139] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A new baseline for image annotation. In *Proceedings of the 10th European Conference on Computer Vision*, ECCV, 2008.

[140] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[141] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

[142] Brian McFee and Gert R Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 775–782, 2010.

[143] Mary Meeker. Internet Trends 2014 - Code Conference. Technical report, Kleiner Perkins Caufield Byers (KPCB), May 14, 2014.

[144] Thomas Mensink, Jakob J. Verbeek, and Gabriela Csurka. Learning structured prediction models for interactive image labeling. In *CVPR*, 2011.

[145] Donald Metzler and R Manmatha. An inference network approach to image retrieval. In *Image and video retrieval*, pages 42–50. Springer, 2004.

[146] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *JMLR*, 6:2651–2667, 2006.

[147] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. In *Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pages 275–278, New York, NY, USA, 2003. ACM.

[148] Florent Monay and Daniel Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. In *Proceedings of the 12th Annual ACM International*

Conference on Multimedia, MULTIMEDIA '04, pages 348–351, New York, NY, USA, 2004. ACM.

[149] Yadong Mu, Jian Dong, Xiaotong Yuan, and Shuicheng Yan. Accelerated low-rank visual recovery by random projection. In *CVPR*, pages 2609–2616. IEEE Computer Society, 2011.

[150] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

[151] Zhenxing Niu, Gang Hua, Xinbo Gao, and Qi Tian. Semi-supervised relational topic model for weakly annotated image recognition in social media. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 4233–4240, 2014.

[152] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[153] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[154] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[155] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[156] Shibin Parameswaran and Kilian Q Weinberger. Large margin multi-task metric learning. In *Advances in neural information processing systems*, pages 1867–1875, 2010.

[157] James Petterson and Tibério S. Caetano. Submodular multi-label learning. In *NIPS*, pages 1512–1520, 2011.

[158] Duangmanee Putthividhya, Hagai Thomas Attias, and Srikantan S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010.

[159] Ali Mustafa Qamar, Eric Gaussier, J-P Chevallet, and Joo Hwee Lim. Similarity learning for nearest neighbor classification. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 983–988. IEEE, 2008.

[160] Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua, and Hong-Jiang Zhang. An efficient sparse metric learning in high-dimensional space via l1-penalized log-determinant regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 841–848, New York, NY, USA, 2009. ACM.

[161] Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

[162] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[163] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[164] Jonathan Robinson and Vojislav Kecman. Combining support vector machine learning with the discrete cosine transform in image compression. *Neural Networks, IEEE Transactions on*, 14(4):950–958, 2003.

[165] Angelika Rohde, Alexandre B Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[166] Rómer Rosales and Glenn Fung. Learning sparse metrics via linear programming. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–373. ACM, 2006.

[167] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Comput.*, 16(5):1063–1076, May 2004.

[168] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

[169] Bernhard Schölkopf, Alex J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[170] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization and beyond*. MIT Press, 2002.

[171] Shai Shalev-Shwartz, Yoram Singer, and Andrew Y. Ng. Online and batch learning of pseudo-metrics. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 94–, New York, NY, USA, 2004. ACM.

[172] Chunhua Shen, Junae Kim, Lei Wang, and Anton van den Hengel. Positive semidefinite metric learning with boosting. In *Advances in Neural Information Processing Systems*, pages 1651–1659. 2009.

[173] Chunhua Shen, Junae Kim, Lei Wang, and Anton Van Den Hengel. Positive semidefinite metric learning using boosting-like algorithms. *The Journal of Machine Learning Research*, 98888(1):1007–1036, 2012.

[174] Amit Singer and Mihai Cucuringu. Uniqueness of low-rank matrix completion by rigidity theory. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1621–1641, 2010.

[175] Steve Smale and Ding-Xuan Zhou. Geometry on probability spaces. *Constructive Approximation*, 30(3):311–323, 2009.

[176] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[177] Jan A. Snyman. *Practical mathematical optimization : an introduction to basic optimization theory and classical and new gradient-based algorithms*. Applied optimization. Springer, New York, 2005.

[178] Nathan Srebro and Ruslan Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064, 2010.

[179] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *Learning Theory*, pages 545–560. Springer, 2005.

[180] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In *International Conference on Machine Learning Workshop*, 2012.

[181] G. W. Stewart and Ji-Guang Sun. Matrix Perturbation Theory (Computer Science and Scientific Computing). 1990.

[182] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *JMLR*, 8:1027–1061, 2007.

[183] Ishrat Jahan Sumana, Md Monirul Islam, Dengsheng Zhang, and Guojun Lu. Content based image retrieval using curvelet transform. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 11–16. IEEE, 2008.

[184] Qi Tian, C. Aggarwal, Guo-Jun Qi, Heng Ji, and Thomas S. Huang. Exploring context and content links in social media: A latent space method. *PAMI*, 34(5):850–862, 2012.

[185] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6(615-640):15, 2010.

[186] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[187] Lorenzo Torresani and Kuang-chih Lee. Large margin component analysis. In *Advances in neural information processing systems*, pages 1385–1392, 2006.

[188] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 477–485, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[189] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[190] Jingyan Wang, Yihua Zhou, Haoxiang Wang, Xiaohong Yang, Feng Yang, and Austin Peterson. Image tag completion by local learning. In *Advances in Neural Networks–ISNN 2015*, pages 232–239. Springer, 2015.

[191] Jun Wang, Huyen T Do, Adam Woznica, and Alexandros Kalousis. Metric learning with multiple kernels. In *Advances in Neural Information Processing Systems*, pages 1170–1178, 2011.

[192] Jun Wang, Alexandros Kalousis, and Adam Woznica. Parametric local metric learning for nearest neighbor classification. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1601–1609. Curran Associates, Inc., 2012.

[193] Qifan Wang, Bin Shen, Shumiao Wang, Liang Li, and Luo Si. Binary codes embedding for fast image tagging with incomplete labels. In *Computer Vision–ECCV 2014*, pages 425–439. Springer, 2014.

[194] X-J Wang, Lei Zhang, Feng Jing, and Wei-Ying Ma. Annosearch: Image auto-annotation by search. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1483–1490. IEEE, 2006.

[195] Zheng Wang, Jiashi Feng, Changshui Zhang, and Shuicheng Yan. Learning to rank tags. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '10, pages 42–49, New York, NY, USA, 2010. ACM.

[196] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

[197] Kilian Q Weinberger and Lawrence K Saul. Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1160–1167. ACM, 2008.

[198] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[199] K.Q. Weinberger and G. Tesauro. Metric learning for kernel regression. In *Artificial Intelligence and Statistics*, 2007.

[200] Lei Wu, Steven C. H. Hoi, Rong Jin, Jianke Zhu, and Nenghai Yu. Distance metric learning from uncertain side information with application to automated photo tagging. In *ACM Multimedia*, 2009.

[201] Lei Wu, Rong Jin, and Anil K Jain. Tag completion for image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):716–727, 2013.

[202] Pengcheng Wu, Steven Chu-Hong Hoi, Peilin Zhao, and Ying He. Mining social images with distance metric learning for automated image tagging. In *Proceedings of the Fourth*

ACM International Conference on Web Search and Data Mining, WSDM, pages 197–206, New York, NY, USA, 2011. ACM.

[203] Yu Xiang, Xiangdong Zhou, Zuotao Liu, Tat-Seng Chua, and Chong-Wah Ngo. Semantic context modeling with maximal margin conditional random fields for automatic image annotation. In *CVPR*, 2010.

[204] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492, 2010.

[205] Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 505–512, 2002.

[206] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. Tag refinement by regularized lda. In *ACM International Conference on Multimedia*, pages 573–576. ACM, 2009.

[207] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

[208] Xing Xu, Akira Shimada, and Rin-ichiro Taniguchi. Tag completion with defective tag assignments via image-tag re-weighting. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6. IEEE, 2014.

[209] Zhixiang Xu, Kilian Q Weinberger, and Olivier Chapelle. Distance metric learning for kernel machines. *arXiv preprint arXiv:1208.3422*, 2012.

[210] Liu Yang and Rong Jin. Distance metric learning: A comprehensive survey. Technical report, Michigan State Univ., 2009.

[211] Liu Yang, Rong Jin, Rahul Sukthankar, and Yi Liu. An efficient algorithm for local distance metric learning. In *AAAI*, volume 2, 2006.

[212] Peipei Yang, Kaizhu Huang, and Cheng-Lin Liu. Multi-task low-rank metric learning based on common subspace. In *Neural Information Processing*, pages 151–159. Springer, 2011.

[213] Peipei Yang, Kaizhu Huang, and Cheng-Lin Liu. Geometry preserving multi-task metric learning. *Machine learning*, 92(1):133–175, 2013.

[214] Yang Yang, Wensheng Zhang, and Yuan Xie. Image automatic annotation via multi-view deep representation. *Journal of Visual Communication and Image Representation*, 33:368–377, 2015.

[215] Yi Yang, Fei Wu, Feiping Nie, Heng Tao Shen, Yueting Zhuang, and Alexander G Hauptmann. Web and personal image annotation by mining label correlation with relaxed visual graph embedding. *Image Processing, IEEE Transactions on*, 21(3):1339–1351, 2012.

[216] Dit-Yan Yeung and Hong Chang. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints. *Pattern Recognition*, 39(5):1007–1010, 2006.

[217] Yiming Ying, Kaizhu Huang, and Colin Campbell. Sparse metric learning via smooth optimization. In *Advances in neural information processing systems*, pages 2214–2222, 2009.

[218] Joaquin Zepeda, Ewa Kijak, and Christine Guillemot. Sift-based local image description using sparse representations. In *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*, pages 1–6. IEEE, 2009.

[219] Changshui Zhang, Feiping Nie, and Shiming Xiang. A general kernelization framework for learning algorithms based on kernel pca. *Neurocomputing*, 73(4):959–967, 2010.

[220] Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[221] Xianxing Zhang and Lawrence Carin. Joint modeling of a matrix with associated text via latent binary features. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1556–1564. Curran Associates, Inc., 2012.

[222] Yu Zhang and Dit-Yan Yeung. Transfer metric learning by learning task relationships. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1199–1208. ACM, 2010.

[223] Ning Zhou, William K. Cheung, Guoping Qiu, and Xiangyang Xue. A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1281–1294, 2011.

[224] Guangyu Zhu, Shuicheng Yan, and Yi Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *International Conference on Multimedia*. ACM, 2010.

[225] Xiaojin Zhu, David M. Blei, and John Lafferty. Taglda: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison, 2006.

[226] Jinfeng Zhuang and Steven C. H. Hoi. A two-view learning approach for image tag ranking. In *WSDM*, pages 625–634, 2011.

[227] Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM computing surveys (CSUR)*, 38(2):6, 2006.