Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions

Zijia Lin[†,‡] Guiguang Ding[‡] Mingqing Hu[§] Jianmin Wang[‡] Xiaojun Ye[‡]
[†]Department of Computer Science and Technology, Tsinghua University, Beijing, P.R.China
[‡]School of Software, Tsinghua University, Beijing, P.R.China
[§]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R.China

Abstract

*Though widely utilized for facilitating image management, user-provided image tags are usually incomplete and insufficient to describe the whole semantic content of corresponding images, resulting in performance degradations in tag-dependent applications and thus necessitating effective tag completion methods. In this paper, we propose a novel scheme denoted as LSR for automatic image tag completion via image-specific and tag-specific **L**inear **S**parse **R**econstructions. Given an incomplete initial tagging matrix with each row representing an image and each column representing a tag, LSR optimally reconstructs each image (*i.e. *row) and each tag (*i.e. *column) with remaining ones under constraints of sparsity, considering image-image similarity, image-tag association and tag-tag concurrence. Then both image-specific and tag-specific reconstruction values are normalized and merged for selecting missing related tags. Extensive experiments conducted on both benchmark dataset and web images well demonstrate the effectiveness of the proposed LSR.*

1. Introduction

With the prevalence of social network and digital photography in recent years, numberless images have been posted to various photo sharing communities, *e.g*. Flickr. Generally, such large-scale social images are associated with user-provided textual tags for describing their semantic content, which are widely utilized for facilitating image management. However, due to the time-consuming tagging process and the arbitrariness of user tagging behaviours, the user-provided tags probably contain imprecise ones, and they are usually incomplete, as also revealed in [1, 11]. The imprecision and incompleteness of user-provided tags probably leads to performance degradations of various tag-dependent applications like tag based image retrieval, *etc*. Therefore, recently tag refinement, including denoising and completion, has become an attractive subject of many ongoing researches. However, previous work on tag refinement, as referred to in related work, focused more on denoising but less on completion, even though incompleteness can also introduce serious negative effects.

Given an incomplete initial tagging matrix, tag completion is to fill it up by identifying more correct associations between images and tags. In this paper we propose a novel tag completion scheme denoted as LSR and tackle the problem from both perspectives of image (*i.e*. row) and tag (*i.e*. column). Specifically, each image and tag is optimally reconstructed with remaining ones under constraints of sparsity, and then the reconstruction values from both perspectives are normalized and merged for predicting the relevance of unlabelled tags. Regarding the image-specific reconstruction, both low-level image features and high-level tagging row vectors are considered. As for the tag-specific reconstruction, we mainly consider their corresponding column vectors in the initial tagging matrix, which essentially mines their concurrence for seeking unlabelled high-confidence tags with initially labelled ones within an image. Therefore, the proposed LSR is a unified framework merging image-image similarity, image-tag association and tag-tag concurrence for tag completion.

The main contributions of our research are summarized as follows.

- We propose an effective tag completion scheme via image-specific and tag-specific linear sparse reconstructions, considering and merging various contextual information.
- We propose to perform tag completion for each row and column separately, instead of performing global refinement for the tagging matrix, enabling LSR to complete both existing datasets (*i.e*. transductive method) and unseen images (*i.e*. inductive method).

The remainder of this paper is organized as follows. Section 2 gives an overview of related work. Section 3 presents

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.212

the formulation details of the proposed LSR. Then detailed description of experiments is given in Section 4. And finally we conclude the paper in Section 5.

2. Related Work

As tag completion is to add high-confidence candidate tags to a given image, it is intuitive to compare it with image auto-annotation and tag recommendation. Image auto-annotation [2, 10, 3] is to automatically associate unlabelled images with semantically related tags. A. Makadia *et al.* [10] proposed a widely-used annotation baseline denoted as JEC, which is a straightforward but sophisticated greedy algorithm propagating labels from nearest visual neighbours to the target image. M. Guillaumin *et al.* [3] put forward another auto-annotation model named TagProp, which adopts discriminative metric learning methods in nearest neighbour models and maintains the state-of-the-art performance. Tag recommendation [11, 14, 5] is a trade-off between auto-annotation and manual tagging, which is to recommend semantically related tags to a user while he is annotating an image online. B. Sigurbjörnsson and R. v. Zwol [11] introduced a generic tag recommendation method that deploys the collective knowledge residing in images. And S. Lee *et al.* [5] formulated tag recommendation as a maximum a posteriori (MAP) problem using a visual folksonomy.

Though with similar goals, image auto-annotation and tag completion are still different, since the majority of existing auto-annotation methods are founded on the assumption that images in training set are *completely* annotated with *appropriate* tags. Hence the scenario of auto-annotation generally consists of a perfect training set and an unlabelled or partially labelled test set. But what tag completion faces is just a dataset made up of partially annotated images, and it is supposed to add missing related tags to each image. In this sense, tag completion seems more difficult, as no extra perfect training set is available. As for tag recommendation methods, they are generally designed to work online and prefer to incorporating feedback from labellers, while tag completion can be automatically done offline with looser requirements of real-time performance.

As mentioned previously, tag completion is included in tag refinement framework, which has recently become an attractive subject of many ongoing researches [12, 4, 6, 16, 9, 7]. S. Lee *et al.* [4] utilized neighbour voting to learn the relevance of each tag, and then differentiated noisy tags from correct ones. D. Liu *et al.* [6] performed tag denoising according to the consistency between “visual similarity” and “semantic similarity” in images, and then enriched denoised tags with their synonyms and hypernyms in WordNet. In [16] G. Zhu *et al.* formulated the tag refinement problem as a decomposition of the initial tagging matrix into a low-rank refined matrix and a sparse error matrix, with the optimization objective of low-rank, content consistency, tag correlation and error sparsity. Y. Liu *et al.* [9] constructed semantic unities with pairs of associated tag and image, and further proposed a hyper-graph model for tag clustering and refinement.

By reviewing previous researches on tag refinement, we realize that they focused more on tag denoising but less on tag completion. Although a few of them (*i.e.* [16, 9, 7]) are declared to be unified frameworks, unifying with tag denoising can probably introduce risks to the performance of tag completion, due to the difficulty in controlling the degree of denoising. Therefore, recently researchers began to pay more attention to tag completion and even treat it as an independent problem. L. Wu *et al.* [13] proposed to address the problem by searching for the optimal tagging matrix consistent with both observed tags and visual similarities. X. Liu *et al.* [8] formulated the tag completion procedure as a non-negative data factorization problem and tackled it by embedding various contextual information like within-image and cross-image relations, *etc*.

3. Proposed LSR

As illustrated in Fig. 1, the proposed LSR consists of two parts, *i.e.* image-specific and tag-specific linear sparse reconstructions, from which corresponding tag completion results based on reconstruction values are output for further normalization and integration. Then according to the merged completion result, unlabelled tags with higher reconstruction values are selected.

Specifically, we formulate each of the image-specific and tag-specific reconstructions as a convex optimization problem under constraints of sparsity. The sparsity constraints are attributed to the observation that generally an image contains a few objects and a tag connotes a few levels of meaning, and usually corresponding objects or levels of meaning are redundantly contained or implied in the context. In image-specific reconstruction, both low-level image features and high-level tagging row vectors are considered and integrated into a unified objective function. In tag-specific reconstruction, we perform linear sparse reconstruction for each column vector of the tagging matrix with remaining ones. It should be noticed that the correlations of tagging vectors utilized in image-specific and tag-specific reconstructions are totally different. The former is to represent the semantic similarity between images while the latter is to represent the concurrence between tags. Hence image-specific reconstruction mainly utilizes the visual similarity and semantic similarity between images, while tag-specific reconstruction mines the concurrence between tags.

With both separate linear sparse reconstructions, LSR further normalizes and merges their respective tag completion results. Here we adopt a weighted linear combination strategy as follows, which is well demonstrated by experi-

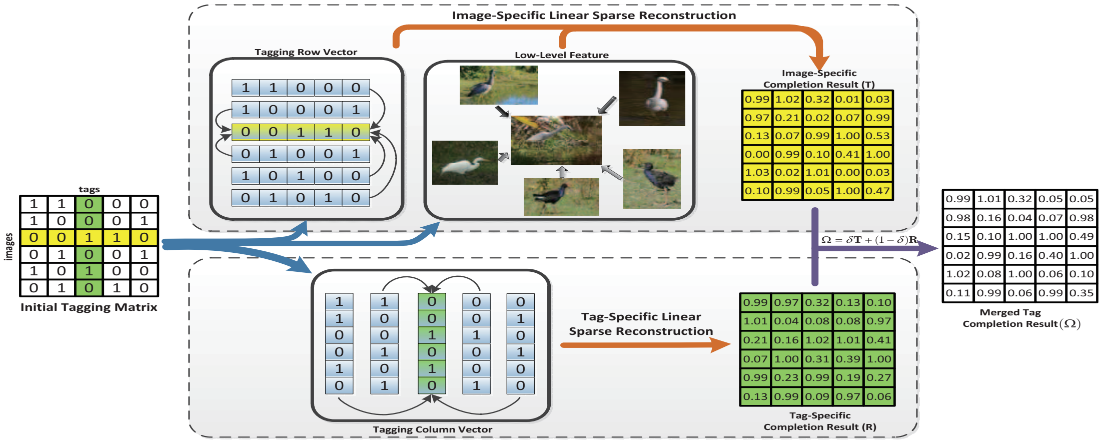
Image-Specific Linear Sparse Reconstruction
Tagging Row Vector
Low-Level Feature
Image-Specific Completion Result (T)
tags
images
Initial Tagging Matrix
$\Omega = \delta \mathbf{T} + (1-\delta)\mathbf{R}$
Merged Tag Completion Result (Ω)
Tag-Specific Linear Sparse Reconstruction
Tagging Column Vector
Tag-Specific Completion Result (R)

Figure 1. Framework of LSR, illustrated with toy data. Given an incomplete initial tagging matrix, LSR separately performs tag completion from both perspectives of image (upper dotted square) and tag (lower dotted square), and then normalizes and merges corresponding results.

ments to be effective though straightforward.

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

where Ω is the expected final result, T and R are respectively the normalized completion results from image-specific and tag-specific reconstructions, and δ is a weighting parameter in $(0,1)$. Then based on Ω, corresponding unlabelled tags with higher reconstruction values are added to each incompletely annotated image.

3.1. Image-Specific Reconstruction

Given the incomplete initial tagging matrix $D_{m\times n}$, where m and n respectively denote the number of images and tags, image-specific linear sparse reconstruction is to perform tag completion from the perspective of row. As mentioned previously, both low-level features and high-level tagging vectors are considered.

Linear sparse reconstruction w.r.t. low-level image features is to reconstruct an image with others using their corresponding feature vectors. Assuming the feature vector of a to-be-reconstructed image is $f_{l\times 1}$, where l is the dimensionality of the feature vector, the image-specific reconstruction w.r.t. low-level features can be formulated as follows.

$$\Theta_1 = min_{\alpha} \quad \|f - F\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{2}$$

where $\|\cdot\|_2$ and $\|\cdot\|_1$ are respectively $L2$ norm and $L1$ norm, $F_{l\times(m-1)}$ is a dictionary matrix consisting of feature vectors of other images, $\alpha_{(m-1)\times 1}$ is the objective weighting vector with each element representing the weight of corresponding image in the linear sparse reconstruction of f, and λ is a tuning factor for penalizing the non-sparsity of α.

Regarding the linear sparse reconstruction w.r.t. high-level tagging vectors, we introduce a group sparse structure for the reconstruction weights as [15], which is attributed to the observation that images associated with an identical tag probably share more common semantic content and thus form a group. Here we denote the ith group of reconstruction weights as $g_i = \{\beta_{\kappa(i,1)}, \beta_{\kappa(i,2)}, \cdots, \beta_{\kappa(i,|g_i|)}\}$, where $\kappa(i,j)$ is the index of the jth weight of the ith group in the weighting vector β. Note that each tag corresponds to a group of reconstruction weights, *i.e.* the weights of images containing the tag, and thus the groups can be overlapped since images are usually labelled with several tags. Here the objective function is formulated as follows.

$$\Theta_2 = min_{\beta} \quad \|W\left(t - \widehat{T}\beta\right)\|_2^2 + \omega\sum_{i=1}^{n}\|g_i\|_2 \tag{3}$$

where $t_{n\times 1}$ is the tagging vector of a to-be-reconstructed image, $\widehat{T}_{n\times(m-1)}$ is the dictionary matrix containing tagging vectors of other images, $\beta_{(m-1)\times 1}$ is the objective weighting vector denoting the weights of other images in the linear sparse reconstruction of t, and ω is a tuning factor for balancing the group sparsity. Here the group sparsity $\sum_{i=1}^{n}\|g_i\|_2$ separately uses $L2$ norm for smoothing intra-group weights and $L1$ norm for emphasizing inter-group sparsity. It is reasonable since only a few tags are associated with the target image (*i.e.* inter-group sparsity), and images in the same group are all supposed to contribute to the reconstruction if the corresponding tag would be associated (*i.e.* intra-group smoothing). Additionally, W is a diagonal matrix for weighting the reconstruction residual of each entry in t, defined as $W_{i,i} = \exp(t_i)$. It can be seen that W assigns higher weights to the non-zero entries (*i.e.* labelled tags) of the initial tagging vector, since they are ensured while the zero ones (*i.e.* unlabelled tags) are not.

Furthermore, for image-specific reconstruction we integrate both objective functions above into a unified optimization framework as follows.

$$\begin{aligned}\Theta = min_{\alpha,\beta} \quad & \|f - F\alpha\|_2^2 + \lambda\|\alpha\|_1 + \\ \mu & \left(\|W\left(t - \widehat{T}\beta\right)\|_2^2 + \omega\sum_{i=1}^{n}\|g_i\|_2 \right) + \nu\|\alpha - \beta\|_2^2\end{aligned} \tag{4}$$

where μ is a weighting parameter for balancing the reconstructions w.r.t. low-level features and high-level tagging vectors, and ν is a tuning factor for penalizing the difference between α and β. Note that there is one-to-one correspondence between α and β. Though the reconstructions w.r.t. low-level features and high-level tagging vectors are supposed to be consistent (*i.e.* $\alpha = \beta$), the insurmountable semantic gap between both still makes them slightly different. And thus it is more reasonable to introduce a soft penalization for the difference between α and β rather than use a hard equality constraint to replace β with α, as validated in our experiments. The integrated objective function can be demonstrated to be convex, meaning that there exists a global optimal solution. Then the optimal α and β can be merged for obtaining a reconstructed tagging vector t' for the target image, as shown in formula (5).

$$t' = \widehat{T}\left(\rho\alpha + (1 - \rho)\beta\right) \tag{5}$$

where ρ is a weighting factor in $(0, 1)$.

By performing linear sparse reconstructions for all to-be-completed images, an image-specific reconstructed tagging matrix $T_{m\times n}$ consisting of all t'^T can be output for further normalization and integration. Specifically, in our experiments the maximal value of $T_{m\times n}$ is normalized as 1.

3.2. Tag-Specific Reconstruction

Tag-specific linear sparse reconstruction is to perform tag completion for the incomplete initial tagging matrix $D_{m\times n}$ from the perspective of column. Here we denote the tagging column vector in D of a to-be-completed tag as $\tau_{m\times 1}$, and the dictionary matrix consisting of other tagging column vectors as $\widehat{R}_{m\times(n-1)}$. Then the process of tag-specific reconstruction can be formulated as follows.

$$\Psi = min_{\gamma} \quad \|W'\left(\tau - \widehat{R}\gamma\right)\|_2^2 + \xi\|\gamma\|_1 \tag{6}$$

where $\gamma_{(n-1)\times 1}$ is the objective weighting vector with each element representing the weight of corresponding tag in the reconstruction, and ξ is a tuning factor for penalizing the non-sparsity of γ. Additionally, W' is a diagonal weighting matrix for the reconstruction residuals of all entries in τ, which is defined in the same way as W in formula (3). The tag-specific objective function can also be demonstrated to be convex and thus there exists a global optimal γ, which can then be utilized to obtain a reconstructed tagging column vector $r' = \widehat{R}\gamma$ for the target tag.

With all tags reconstructed, a tag-specific reconstructed tagging matrix $R_{m\times n}$ consisting of all r' can be output for further normalization, which in our experiments is to normalize the maximal value as 1. Then the normalized image-specific and tag-specific tagging matrix, *i.e.* T and R, are merged as formula (1) for performance enhancement.

3.3. Solution and Implementation Issues

Since the objective functions for image-specific and tag-specific reconstructions (*i.e.* formula (4) and (6)) are both unconstrained convex optimization problems, effective (sub-)gradient descent based methods can be adopted.

Regarding the objective function of image-specific linear sparse reconstruction (*i.e.* formula (4)), respective partial derivatives for α and β can be separately calculated as follows and then concatenated as the gradient of η, subject to $\eta = \left(\alpha^T, \beta^T\right)^T$.

$$\frac{\partial\Theta}{\partial\alpha} = -2F^Tf + 2F^TF\alpha + \lambda\, I\left(\alpha\right) + 2\nu\left(\alpha - \beta\right) \tag{7}$$

$$\begin{aligned}\frac{\partial\Theta}{\partial\beta} = & \mu\left(-2\widehat{T}^TW^TWt + 2\widehat{T}^TW^TW\widehat{T}\beta + G\beta\right) \\ & + 2\nu\left(-\alpha + \beta\right)\end{aligned} \tag{8}$$

where $I\left(\alpha\right)$ is an indicator function for all entries in α, defined as $I\left(\alpha\right)_i = \frac{a_i}{|a_i|}$ and assigned as 0 when $|a_i| = 0$. Though $\|\alpha\|_1$ is not differentiable at the specific point of zero, here we follow previous work and simply use 0 as the partial derivatives of zero entries in α. Additionally, $G\beta$ is the partial derivative regarding the group sparsity part in Θ, and G is a diagonal matrix with its diagonal elements defined as $G_{i,i} = \sum\limits_{g_k\ s.t.\ \beta_i \in g_k} \dfrac{\omega}{\|g_k\|_2}$.

The objective function of tag-specific linear sparse reconstruction (*i.e.* formula (6)) can also be optimized with the gradient w.r.t. γ given as follows.

$$\frac{\partial\Psi}{\partial\gamma} = -2\widehat{R}^TW'^TW'\tau + 2\widehat{R}^TW'^TW'\widehat{R}\gamma + \xi I\left(\gamma\right) \tag{9}$$

where $I\left(\gamma\right)$ is an indicator function for all entries in γ.

Instead of performing global optimization for the tagging matrix as most previous work, LSR performs tag completion for each image and tag separately, enabling itself to complete both existing datasets (*i.e.* transductive method) and unseen images (*i.e.* inductive method). However, the computational cost of LSR may still be high if the to-be-reconstructed vector (*i.e.* f, t or τ) is high-dimensional, or the dictionary matrix (*i.e.* F, $\widehat{T}$ or $\widehat{R}$) is large. Hence we propose that dimensionality reduction methods or sampling strategies like kNN (*i.e.* k Nearest Neighbours) be adopted for shrinking vectors or building smaller dictionary matrices while keeping acceptable performance.

	Corel5k	Flickr30Concepts
Vocabulary Size	260	2,513
Nr. of Images	4,918	27,838
Tags per Image	3.4 / 5	8.3 / 70
Del. Tags per Image	1.4 (40%)	3.3 (40%)
Test Set	492	2,807

Table 1. Statistics of Corel5k and Flickr30Concepts. Counts of tags are given in the format "mean / maximum".

4. Experiments

4.1. Datasets and Measurements

In our experiments, we use the well-known benchmark dataset Corel5k and a new-built web image dataset named Flickr30Concepts for evaluating the proposed LSR. Some statistics of both datasets are given in Table 1. With accurate manual annotations, the labelled tags of each image in Corel5k are relatively complete and contain little noise, making it an ideal evaluation benchmark for tag completion. The new-built dataset, Flickr30Concepts, is collected from Flickr by submitting 30 non-abstract concepts[1] as queries, and for each query the top 1,000 of the retrieved images are gathered. Then we utilize WordNet for stemming and filtering all raw tags, and finally obtain a vocabulary containing 2,513 distinct words, which is larger and more challenging than most vocabularies ever used in experiments of previous related work. Flickr30Concepts will be published soon.

To perform tag completion, we randomly delete 40% of the associated tags for all images in both datasets, ensuring that each image has at least one tag deleted and finally contains at least one tag. Therefore, we strike out images originally associated with only one tag and finally obtain two refinement datasets with statistics shown in table 1. Furthermore, we split Corel5k and Flickr30Concepts into test set (around $1/10$) and training set. Note that here we use the standard Corel5k split for experiments. Furthermore, we take around $1/9$ of the training set of Corel5k as a validate set for parameter tuning. Due to the high cost of manual judgements for tag completion results, we take the originally labelled tags of each image as ground truth for both datasets, and measure the completion results regarding the missing ground-truth tags (*i.e*. deleted tags). Though in most experiments only the images in training set are utilized to build dictionary matrices for a test image (*i.e*. inductive method), the proposed LSR itself can also utilize all images in both training and test sets for reconstruction (*i.e*. transductive method), in which case we denote it as πLSR.

The experimental results of tag completion are measured with $average\ precision@N$ (*i.e.* $AP@N$), $average\ recall@N$ (*i.e.* $AR@N$) and $coverage@N$ (*i.e.* $C@N$). In the top N completed tags, $precision@N$ is to measure the ratio of correct tags, and $recall@N$ is to measure the ratio of missing ground-truth tags, which are both averaged over all test images. $Coverage@N$ is to measure the ratio of test images with at least one correct completed tag. All these measurements are respectively defined as: $AP@N = \frac{1}{m}\sum_{i=1}^{m}\frac{N_c(i)}{N}$, $AR@N = \frac{1}{m}\sum_{i=1}^{m}\frac{N_c(i)}{N_{mg}}$ and $C@N = \frac{1}{m}\sum_{i=1}^{m} I\left(N_c\left(i\right) > 0\right)$, where m is the number of test images, $N_c\left(i\right)$ is the quantity of correctly recovered tags from missing ones for the ith image, $N_{mg}\left(i\right)$ is the number of missing ground-truth tags, and $I\left(\cdot\right)$ is a condition function that returns 1 when the condition is satisfied and 0 otherwise.

In our experiments, we extract ten kinds of features[2] for each image and adopt PCA to separately perform dimensionality reduction for all features of an image, which are then concatenated to be a 400-dimensional merged feature vector. When measuring visual distance between images, we empirically utilize $L2\,norm$ for Edge Histogram and FCTH, χ^2 for Color Layout and JCD, and $L1\,norm$ for remaining features. Then all feature distances are normalized and merged with equal weights as a final visual distance.

4.2. Parameter Settings

Before applying LSR on both datasets, we use the validate set of Corel5k for tuning parameters and analysing their corresponding influences on tag completion result.

Regarding image-specific reconstruction, to reduce computational cost, we adopt a kNN strategy and take 200 nearest visual neighbours from training set for any to-be-completed image in the validate set to build the dictionary matrix. The merging parameter ρ in formula (5) is empirically set as 0.5 for equally weighting the roles of feature vectors and tagging vectors. Then we utilize the control variable method for analysing the influences of λ, μ, ω and ν in formula (4). Specifically, we initialize λ, μ, ω and ν as 1, and tune each parameter with others fixed. Each parameter is tuned with values in $\{0, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$, and their influences are illustrated in Fig. 2 (sub-figure $[a1]$ to $[a4]$). Note that here we utilize $AP@2$, $AR@2$ and $C@2$ as measurements since on average 1.4 tags are deleted for each image in Corel5k. It can be seen that the optimal parameter settings for λ, μ, ω and ν are respectively around 2, 0.2, 2 and 1. And the observation that all the performance curves w.r.t. precision, recall and coverage are convex on the whole reflects the significance of all parts included in the objective function of image-specific reconstruction. Moreover, in sub-figure $[a4]$ the tag com-

[1] The 30 non-abstract concepts are: aircraft, ball, beach, bike, bird, book, bridge, car, chair, child, clock, countryside, dog, door, fire, fish, flower, house, kite, lamp, mountain, mushroom, pen, rabbit, river, sky, sun, tower, train, tree.

[2] The features include: Color Correlogram, Color Layout, CEDD, Edge Histogram, FCTH, JCD, Jpeg Coefficient Histogram, RGB Color Histogram, Scalable Color, SURF with Bag-of-Words model.

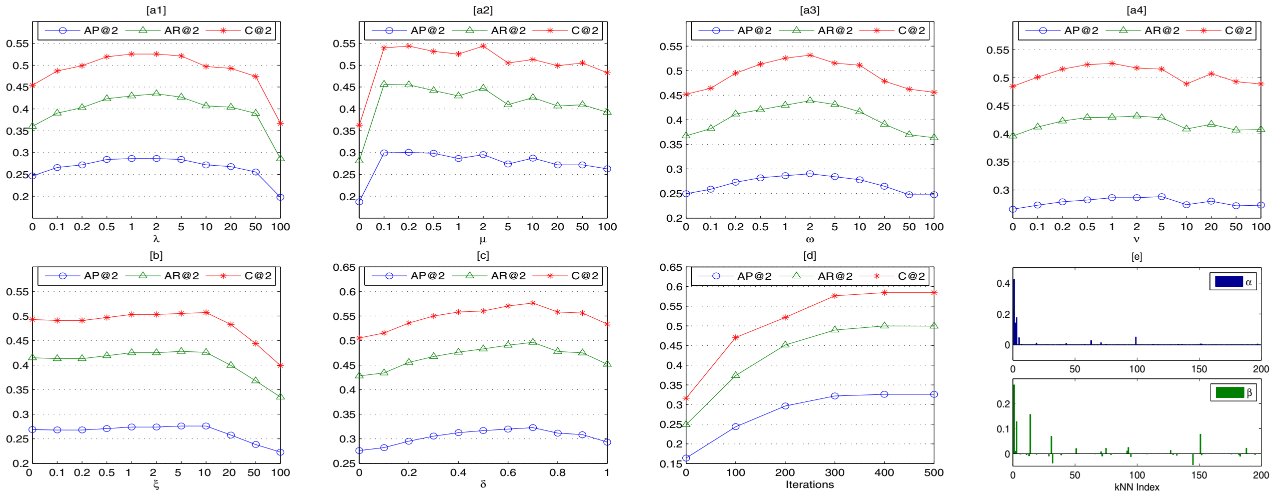

Figure 2. Influences of λ, μ , ω and ν in objective function of image-specific reconstruction (sub-figure $[a1]$ to $[a4]$) , ξ in objective function of tag-specific reconstruction (sub-figure $[b]$) , δ for merging completion results of both perspectives (sub-figure $[c]$) , and $iterations$ in the solution processes (sub-figure $[d]$), in terms of $AP@2$, $AR@2$ and $C@2$ on the validate set of Corel5k, with sub-figure $[e]$ giving an illustration of the optimized α (blue) and β (green) in an image-specific reconstruction process.

pletion performance tends to decrease as ν increases to a large value, which indirectly validates our proposal that α and β should be consistent but slightly different, since a large ν will force α to be closer and even equal to β. Similarly, we perform parameter tuning experiments for ξ in the tag-specific reconstruction, with values in $\{0, 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100\}$, as illustrated in Fig. 2 (sub-figure $[b]$), from which we can find that the optimal setting for ξ is around 5 or 10.

We further investigate the influences of the merging parameter δ in formula (1) , by varying δ from 0 to 1 with a step of 0.1. As shown in Fig. 2 (sub-figure $[c]$), the optimal setting for δ is around 0.7. Moreover, it can be concluded that the optimal merged completion result outperforms that of mere image-specific (*i.e.* $\delta = 1$) or mere tag-specific (*i.e.* $\delta = 0$) reconstruction, which demonstrates the effectiveness of performing tag completion from both perspectives.

Furthermore, we investigate how the tag completion performance varies with the iterations in the solution processes for optimizing image-specific and tag-specific objective functions, as illustrated in Fig. 2 (sub-figure $[d]$). For ease of presentation, we show the merged completion result. It can be seen that with the iterations increasing, the tag completion performance achieves continuous improvement and tends to be stable after about 400 iterations, which is adopted as the upper bound for iterations in following experiments. To provide inside details regarding the optimization process, we further give an illustration of the optimized α and β in an image-specific reconstruction process. As shown in sub-figure $[e]$, the optimized α and β are both sparse, and similar but not exactly equal to each other, which well supports our former proposal. Though not illustrated here due to the space limit, the optimized γ in the tag-specific reconstruction is also sparse.

4.3. Tag Completion Results

To evaluate the performance of the proposed LSR, we adopt remarkable image auto-annotation methods (*i.e.* JEC [10] and TagProp [3]), tag recommendation approaches (*i.e.* Vote+ [11] and Folksonomy [5]) and recently proposed unified tag refinement frameworks of denoising and completion (*i.e.* LR [16] and SUG [9]) as baselines for comparison. Note that the parameters of adopted baselines are also carefully tuned on the validate set of Corel5k with corresponding proposed tuning strategy for achieving optimal performance and making fair comparisons (*e.g.* kNN = 200 for JEC and σML of TagProp, $[m, k_s, k_d, k_r] = [35, 3, 4, 2]$ for Vote+, *etc.*), which actually proves to perform better than just adopting published parameter settings. Regarding LSR, we extensively evaluate several variants. LSR_img_feat and LSR_img_tvec separately represent two variants that perform tag completion utilizing only low-level features or high-level tagging vectors for image-specific reconstruction, *i.e.* formula (2) and (3). LSR_img and LSR_tag are variants that respectively utilize mere image-specific or tag-specific linear sparse reconstruction, *i.e.* formula (4) and (6), of which the former is the integration of LSR_img_feat and LSR_img_tvec. LSR is the proposed integrated scheme merging LSR_img and LSR_tag (*i.e.* inductive version), and πLSR is a transductive version that utilizes all other images in both training set and test set to build dictionary matrices for any test image. Note that all the completion results are measured on the same test sets.

For experiments on Corel5k, we measure all the algo-

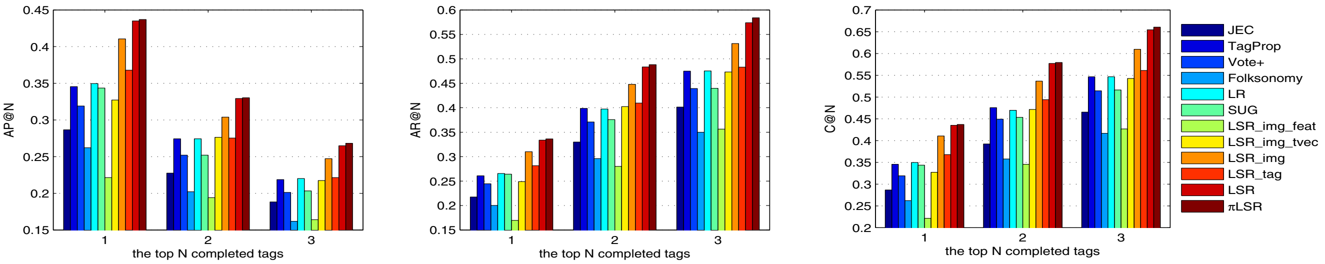

Figure 3. Tag completion results on Corel5k in terms of $AP@N$, $AR@N$ and $C@N$ with N in $\{1, 2, 3\}$. Among the baselines, JEC and TagProp are image auto-annotation methods, Vote+ and Folksonomy are tag recommendation approaches, while LR and SUG are unified tag refinement frameworks of denoising and completion. Others are variants of the proposed LSR.

rithms in terms of $AP@N$, $AR@N$ and $C@N$, with N varying from 1 to 3, as illustrated in Fig. 3. From the experimental results we can draw the following conclusions. 1) Most variants of LSR consistently outperform the image auto-annotation, tag recommendation and tag refinement baselines, providing a demonstration for their effectiveness. 2) LSR_img outperforms both LSR_img_feat and LSR_img_tvec, which validates the necessity of considering both low-level features and high-level tagging vectors in image-specific linear sparse reconstruction, and using their semantic consistency for performance enhancement. 3) LSR outperforms both LSR_img and LSR_tag, which provides further evidence for the effectiveness of performing tag completion from both perspectives of image and tag. 4) πLSR slightly outperforms LSR, since πLSR utilizes more partially labelled images in test set for building dictionary matrices while LSR does not. 5) Initially labelled tags are important for tag completion, as validated by that LSR_img_feat achieves the worst performance while other tag-dependent variants perform much better.

Regarding experiments on the much larger real-world Flickr30Concepts, we take $AP@4$, $AR@4$ and $C@4$ as performance measurements, since the mean number of missing tags for each image in Flickr30Concepts is 3.3. Note that we cannot obtain the completion results of SUG on Flickr30Concepts due to its high computational cost to calculate the eigenvalues of the normalized Laplacian matrix of a large hypergraph. From Table 2 we can draw nearly the same conclusions as those on Corel5k, which further demonstrate the effectiveness of the proposed scheme. Moreover, since here the variants of LSR use the same parameter settings as Corel5k, such results in some way provide an evidence for the robustness of LSR.

	$AP@4$	$AR@4$	$C@4$
JEC	0.25	0.30	0.49
TagProp	0.23	0.29	0.50
Vote+	0.23	0.27	0.48
Folksonomy	0.21	0.26	0.47
LR	**0.27**	**0.34**	**0.51**
SUG	-	-	-
LSR_img_feat	0.16	0.21	0.39
LSR_img_tvec	0.25	0.28	0.43
LSR_img	0.33	0.40	0.60
LSR_tag	0.29	0.35	0.59
LSR	**0.37**	**0.45**	**0.67**
πLSR	**0.38**	**0.46**	**0.69**

Table 2. Experimental results on real-world Flickr30Concepts, in terms of $AP@4$, $AR@4$ and $C@4$.

	Corel5k $(N = 2)$			Flickr30Concepts $(N = 4)$		
	AP	AR	C	AP	AR	C
TMC	**0.23**	**0.33**	**0.40**	**0.19**	**0.21**	**0.37**
DLC	0.09	0.13	0.18	0.07	0.09	0.23
LSR_img_feat	0.08	0.12	0.15	0.04	0.05	0.14
LSR_img_tvec	0.16	0.23	0.28	0.09	0.10	0.18
LSR_img	0.19	0.28	0.33	0.11	0.13	0.24
LSR_tag	0.28	0.41	0.49	0.29	0.35	0.59
LSR	**0.28**	**0.42**	**0.50**	**0.30**	**0.36**	**0.60**
πLSR	**0.29**	**0.43**	**0.51**	**0.31**	**0.37**	**0.62**

Table 3. Experimental results of TMC, DLC and LSR on Corel5k and Flickr30Concepts with only SIFT BoW feature, in terms of $AP@N$, $AR@N$ and $C@N$.

To compare with the most recently published tag completion methods (*i.e.* TMC [13] and DLC [8]), we further conducted experiments on both datasets with new image feature representations, since DLC requires each image feature vector to be non-negative and TMC prefers the dot product of feature vectors to be non-negative. Here we utilize the SIFT feature with Bag-of-Words (BoW) model to represent each image with a 1,000-dimensional vector, which is natively non-negative while many other features are not. SIFT is also the only common feature used by both baselines and even the main feature in experiments of TMC. The experimental results are shown in Table 3, which further demonstrate the effectiveness of the proposed LSR.

	Corel5k			Flickr30Concepts		
	$(N = 2)$			$(N = 4)$		
	AP	AR	C	AP	AR	C
JEC	0.28	0.41	0.47	0.35	0.44	0.54
TagProp	**0.32**	**0.47**	**0.54**	**0.36**	**0.46**	0.61
Vote+	0.27	0.40	0.49	0.31	0.39	**0.62**
Folksonomy	0.23	0.34	0.40	0.27	0.35	0.50
LR	0.30	0.44	0.51	0.31	0.40	0.55
SUG	0.27	0.40	0.48	-	-	-
LSR	**0.37**	**0.55**	**0.63**	**0.48**	**0.61**	**0.79**
πLSR	**0.38**	**0.55**	**0.63**	**0.48**	**0.61**	**0.78**

Table 4. Experimental results on benchmark Corel5k and real-world Flickr30Concepts with the training sets completely labelled, in terms of $AP@N$, $AR@N$ and $C@N$.

Both baselines yield inferior results here, especially DLC, which can be due to: 1) both are non-convex and may converge to a local optimum, 2) DLC depends heavily on image features for matrix factorization and calculating image similarities. Regarding the proposed scheme, with only SIFT feature many retrieved visual neighbours for building dictionary matrices are unrelated, causing a substantial decline in the performance of image-specific reconstruction. Yet boosted by tag-specific reconstruction, the merged completion results (*i.e*. LSR and πLSR) are still acceptable, which verifies the robustness of LSR and the necessity of performing tag completion from both image and tag perspectives.

4.4. Further Evaluation with Completely Labelled Training Set

To further evaluate the proposed LSR, we conduct experiments on Corel5k and Flickr30Concepts with completely labelled training sets to see whether it can still yield better performance. Specifically, we recover all the missing tags in training sets of both datasets. Then all the algorithms are applied to the same partially labelled test sets as former experiments with corresponding completely labelled training sets and measured in terms of $AP@N$, $AR@N$ and $C@N$, as shown in Table 4. It can be concluded that: 1) the proposed LSR and πLSR significantly outperform other baselines, which further demonstrates their effectiveness, 2) all the algorithms consistently achieve much better performance with a completely labelled training set on both Corel5k and Flickr30Concepts, 3) the advantage of πLSR over LSR tends to be weakened, as the utilization of more partially labelled images in test set may cause negative effects when the training set is already completely labelled.

5. Conclusions

In this paper we propose an effective scheme denoted as LSR for automatic image tag completion, using image-specific and tag-specific linear sparse reconstructions. The proposed LSR respectively fits both reconstructions into convex optimization frameworks that utilize various contextual information. And it achieves the state-of-the-art performance in extensive experiments conducted on both benchmark dataset and web images for tag completion.

6. Acknowledgments

This research was supported by the National Basic Research Project of China (Grant No. 2011CB70700), the National Natural Science Foundation of China (Grant No. 61271394, 61005045). And the authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *SIGCHI '07*.

[2] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR '04*.

[3] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV '09*.

[4] S. Lee, W. Neve, and Y. Ro. Image tag refinement along the 'what' dimension using tag categorization and neighbor voting. In *ICME '10*.

[5] S. Lee, W. D. Neve, K. N. Plataniotis, and Y. M. Ro. Map-based image tag recommendation using a visual folksonomy. *Pattern Recogn. Letters*, 31(9), July 2010.

[6] D. Liu, X. Hua, M. Wang, and H. Zhang. Image retagging. In *MM '10*.

[7] D. Liu, S. Yan, X. Hua, and H. Zhang. Image retagging using collaborative tag propagation. *TMM*, 13(4):702 –712, 2011.

[8] X. Liu, S. Yan, T. Chua, and H. Jin. Image label completion by pursuing contextual decomposability. *ACM TOMCCAP*, 8(2), May 2012.

[9] Y. Liu, F. Wu, Y. Zhang, J. Shao, and Y. Zhuang. Tag clustering and refinement on semantic unity graph. In *ICDM '11*.

[10] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV '08*.

[11] B. Sigurbjörnsson and R. v. Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*.

[12] C. Wang, F. Jing, L. Zhang, and H. Zhang. Content-based image annotation refinement. In *CVPR '07*.

[13] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *TPAMI*, 35(3):716–727, 2013.

[14] L. Wu, L. Yang, N. Yu, and X. Hua. Learning to tag. In *WWW '09*.

[15] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. Metaxas. Automatic image annotation using group sparsity. In *CVPR '10*.

[16] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *MM '10*.