AUTOMATIC IMAGE TAGGING AND RECOMMENDATION VIA PARAFAC2

Evangelia Pantraki and Constantine Kotropoulos

Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
email: pantraki@csd.auth.gr, costas@aiia.csd.auth.gr

ABSTRACT

An important aspect when sharing images in social networks is the tags the images are annotated with. Another closely related problem is the ability to successfully recommend images to users. An automatic image annotation and recommendation system is proposed based on Parallel Factor Analysis 2 (PARAFAC2). Here, PARAFAC2 is applied to a collection of three matrices, namely the image-feature matrix, whose columns are representations capturing the visual appearance of images, the image-tag matrix, whose columns indicate the tags associated with each image, and the image-user matrix, whose columns identify who has uploaded or is associated to each image. PARAFAC2 is able to harness the multi-tag and the multi-user information for reducing the dimensionality of the feature vectors extracted from the images. That is, by projecting the feature vector onto the semantic space derived via PARAFAC2, a sketch (i.e., a coefficient vector of reduced dimensions) is obtained. To predict the tags to be assigned to a test image, the test image sketch is multiplied by the left singular vectors of the image-tag matrix, yielding a tag vector. Similarly, to recommend users who might be interested to a test image, the sketch is multiplied by the left singular vectors of the image-user matrix, yielding a recommendation vector. Promising results are demonstrated when the aforementioned framework is applied to an image dataset of Greek popular tourist landmarks extracted from Flickr, using a 10-fold cross-validation experimental protocol.

***Index Terms*—** Automatic Image Tagging, Image Recommendation, Multi-label Classification, PARAFAC2.

1. INTRODUCTION

The explosive growth of digital technologies, the emergence of social networks, and the deployment of popular photo-sharing web services (e.g., Flickr, Instagram, PhotoBucket) has lead to an exponential growth of the number of images hosted and shared on the Web as well as the creation of large image databases. To leverage the asset of such voluminous information, the images should be annotated with tags, describing their content. Most photo-sharing websites ask the users to define the textual content of the images or the videos they upload. Image annotation or image tagging is the process of adding metadata to the digital content in the form of captioning or keywords [1].

Many research efforts in image annotation focus on content-based image retrieval, where images are indexed and retrieved by resorting to low-level features, such as color, shape, or texture. A regression model with a regularized penalty is developed to annotate images with multiple labels in [2]. For each label, different groups of heterogenous features are selected, employing structural group sparsity. A multi-label sparse coding framework for feature extraction and classification within the context of automatic image annotation is also proposed in [3]. In [4], the authors argue that image regions should be annotated with tags in order to cope with the diversity of web image content. That is, instead of annotating a whole image, tags are assigned to image regions thanks to spatial group sparse coding. Many works on automatic image tagging use graphs and hypergraphs. In [5], a unified graph and a random walk based framework is proposed to bridge the semantic gap between the image content and the tags. In [6], a graph based automatic image annotation and semantic image retrieval approach is also proposed. The authors develop a bi-relational graph model that comprises an image similarity graph and a graph depicting label correlations and connect them by an additional bipartite graph induced from label assignments. [7] deals with the image annotation problem within the social media environment. The authors propose a graph based tagging reinforcement method, where the relations among the tags, the image features and users' friends are taken into consideration. The recommendation problem is addressed in [8], where a sparse linear model is created in order to make recommendations to users that are based solely on the user's profile. The problem of image tag imprecision is addressed in [9] as a convex optimization problem, which simultaneously minimizes the image-tag matrix rank and priors as well as error sparsity.

In this paper, a novel framework for joint automatic multi-label image annotation and image recommendation is pro-

978-1-4673-7454-5/15/$31.00 ©2015 IEEE

posed, extending the previous works [10, 11]. The starting point is to extract feature vectors capturing the visual appearance of each image. Next, an irregular third-order tensor (or more precisely hypermatrix) is formed having three slices. The first slice is the image-feature matrix. It contains typical feature vectors, i.e., the GIST descriptors extracted from the training images [12]. The second slice is the image-tag matrix. Its columns are the multi-label vectors (i.e., tags) associated to the training images. The third slice is the image-user matrix whose columns identify the user who has uploaded the image or the users who are interested in it. Parallel Factor Analysis 2 (PARAFAC 2) [13] is applied to the aforementioned irregular third-order tensor so that the semantic similarities between the label sets associated to images (or the common preferences between users) drive the extraction of meaningful feature vectors of reduced dimensions referred to as *sketches* hereafter. The reasoning behind this approach is that PARAFAC2 represents the feature vector and the associated label and user vectors as linear combinations of basis vectors with coefficients taken from the same vector space. The left singular vectors of the image-feature matrix span a lower dimensional semantic space dominated by the label and user information. Any feature vector extracted from a test image is projected onto this semantic space first in order to obtain a test sketch. Then, the tag annotation vector is obtained by multiplying the test sketch by the left singular vectors of the second slice. Similarly, the user recommendation vector is obtained by multiplying the test sketch by the left singular vectors of the third slice.

The performance of the proposed automatic image annotation and user recommendation framework is assessed by conducting experiments on an image dataset of Greek popular tourist landmarks retrieved from the social network Flickr [14]. The dataset contains information for the user who has uploaded each image and any tags each image is annotated with. Furthermore, we are aware of the friendship relations among the users. To efficiently recommend images to users, we rely on this information assuming that users who are declared as friends have common interests. Accordingly, each user is interested in the images his or her friends have uploaded. Promising results are reported when the aforementioned framework is applied to the aforementioned image dataset, using a 10-fold cross-validation experimental protocol.

The paper is organized as follows. In Section 2, basic concepts from multilinear algebra and notations are listed. The proposed joint multi-label image annotation and multi-user image recommendation framework, which is based on the PARAFAC2, is detailed in Section 3. Experimental results are demonstrated in Section 4, and conclusions are drawn in Section 5.

2. NOTATION AND MULTILINEAR ALGEBRA BASICS

Tensors are considered as the multidimensional equivalent of matrices (i.e., second-order tensors) and vectors (i.e., first-order tensors) [15]. Throughout the paper, tensors are denoted by boldface Euler script calligraphic letters (e.g. $\boldsymbol{\mathcal{X}}$), matrices are denoted by uppercase boldface letters (e.g., $\mathbf{U}$), vectors are denoted by lowercase boldface letters (e.g., $\mathbf{u}$), and scalars are denoted by lowercase letters (e.g., u). $\|.\|_F$ denotes the Frobenius matrix norm, while $\mathbf{B}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{B}$. Let $\mathbb{Z}$ and $\mathbb{R}$ denote the set of integer and real numbers, respectively. A third-order real-valued tensor $\boldsymbol{\mathcal{X}}$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times I_3}$, where $I_n \in \mathbb{Z}$ and $n = 1, 2, 3$. Each element of $\boldsymbol{\mathcal{X}}$ is addressed by 3 indices, i.e., $x_{i_1 i_2 i_3}$. Mode-n unfolding of tensor $\boldsymbol{\mathcal{X}}$ yields the matrix $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \ldots I_{n-1} I_{n+1} \ldots I_3)}$. Hereafter, the operations on tensors are expressed in matricized form [15].

3. JOINT MULTI-LABEL IMAGE ANNOTATION AND MULTI-USER IMAGE RECOMMENDATION

Supervised subspace learning algorithms, such as Linear Discriminant Analysis, assume that the data points annotated with the same label lie close to each other in the feature space, while data bearing different labels are far away. This assumption does not hold in a multi-label framework, rendering subspace learning algorithms useless.

PARAFAC is a multi-way generalization of the singular value decomposition (SVD) [16]. PARAFAC2 [13] is a variant of PARAFAC, which relaxes some of PARAFAC constraints. That is, while PARAFAC applies the same factors across a set of matrices, PARAFAC2 applies the same factor along one mode. The aforementioned relaxation allows the other factor matrices to vary, enabling the application of PARAFAC2 to a collection of matrices having the same number of columns, but different number of rows [15]. Such a collection forms the slices of an irregular third-order tensor. Another important characteristic of PARAFAC2 is its ability to overcome the weakness of conventional supervised subspace learning algorithms to handle multi-labelled data. Due to these characteristics, PARAFAC2 has emerged as an appealing method for multi-label classification. It has been applied successfully to feature extraction and multi-label classification of documents [11] and music tagging [10]. Here, our goal is not confined to the multi-label annotation of images via PARAFAC2. We are interested to extend the potential of PARAFAC2 to image recommendation by exploiting user preferences on the top of image tags.

In order to jointly annotate images with multiple tags and recommend images to multiple users, we train a PARAFAC2 model on an irregular third-order tensor $\boldsymbol{\mathcal{X}}$ having three slices (i.e., matrices). Let $\mathbf{X}^{(1)} \in \mathbb{R}_+^{F \times I}$ be the training image-feature matrix, where F denotes the number of features and I

is the number of images. The image-tag matrix holds the tags associated with each image and is denoted as $\mathbf{X}^{(2)} \in \mathbb{R}_+^{V \times I}$, where V indicates the cardinality of the tag vocabulary. Its ki element $x_{ki}^{(2)} = 1$ if the ith image is labeled with the kth tag in the vocabulary and 0 otherwise. The third matrix holds the user-image relations, i.e., it identifies the users who are interested in each image. Let us denote the third matrix as $\mathbf{X}^{(3)} \in \mathbb{R}_+^{U \times I}$, where U indicates the cardinality of the set of users. Its li element $x_{li}^{(3)} = 1$ if the lth user is associated with the ith image and 0 otherwise.

Since $\mathfrak{X}$ has three slices, the PARAFAC2 seeks a decomposition of the form:

$$\mathbf{X}^{(n)} = \mathbf{U}^{(n)} \, \mathbf{H} \, \mathbf{S}^{(n)} \, \mathbf{W}^T, \quad n = 1, 2, 3 \tag{1}$$

where $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times k}$, $n = 1, 2, 3$ is an orthogonal matrix for each slice, $\mathbf{H} \in \mathbb{R}^{k \times k}$, $\mathbf{S}^{(n)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix of weights for the nth slice of $\mathfrak{X}$, and $\mathbf{W} \in \mathbb{R}^{I \times k}$ is a coefficient matrix. Clearly, $I_1 = F =$ the number of features, $I_2 = V =$ vocabulary size, and $I_3 = R =$ number of users. The value of k specifies the number of latent variables to be extracted from each image.

The decomposition (1) can be obtained by solving the optimization problem:

$$\underset{\mathbf{U}^{(n)}, \, \mathbf{H}, \, \mathbf{S}^{(n)}, \, \mathbf{W}}{\operatorname{argmin}} \sum_{n=1}^{3} \| \mathbf{X}^{(n)} - \mathbf{U}^{(n)} \, \mathbf{H} \, \mathbf{S}^{(n)} \, \mathbf{W}^T \|_F^2. \tag{2}$$

The optimization problem of equation (2) can be effectively solved with the algorithm described in [11]. Having solved the optimization problem (2), one computes the matrix $\mathbf{B} \triangleq \mathbf{U}^{(1)} \, \mathbf{H} \, \mathbf{S}^{(1)} \in \mathbb{R}_+^{F \times k}$. $\mathbf{B}$ spans a feature space of reduced dimensions k, where the semantic relations between the feature vectors and their associations with users are retained. Indeed, the semantic relations between the label vectors as well as the user vectors are propagated to the feature space through the common right singular vectors.

As long as the reduced dimensions feature space spanned by $\mathbf{B}$ is created, a test sketch is derived by pre-multiplying the feature vector extracted from a test image $\mathbf{x} \in \mathbb{R}_+^{F \times 1}$ with $\mathbf{B}^\dagger$, i.e., $\tilde{\mathbf{x}} = \mathbf{B}^\dagger \, \mathbf{x} \in \mathbb{R}^{k \times 1}$. To predict the tags of the test image, one has to compute the tag vector $\mathbf{a} \in \mathbb{R}_+^{V \times 1}$ by

$$\mathbf{a} = \mathbf{U}^{(2)} \, \mathbf{H} \, \mathbf{S}^{(2)} \, \tilde{\mathbf{x}}. \tag{3}$$

The tags associated with the largest values in $\mathbf{a}$ annotate the test image. To recommend the test image to users who would be interested in it, one should compute the recommendation vector $\mathbf{r} \in \mathbb{R}_+^{R \times 1}$ given by

$$\mathbf{r} = \mathbf{U}^{(3)} \, \mathbf{H} \, \mathbf{S}^{(3)} \, \tilde{\mathbf{x}}. \tag{4}$$

The test image is recommended to the users who are associated with the largest values in $\mathbf{r}$.

4. EXPERIMENTAL EVALUATION

4.1. Dataset

In order to assess the performance of the proposed framework in joint image tagging and recommendation, we conducted experiments on the dataset of images used in [14]. The dataset has been retrieved from Flickr and consists of 1292 images of Greek popular tourist landmarks. Each image in the dataset is annotated with a set of tags, describing its semantic content. The just mentioned annotation was performed manually by the users who uploaded the image.

The vocabulary consists of a set of $V = 2366$ tags. The images were uploaded by a set of $R = 440$ users. The visual appearance of images is captured by GIST descriptors of size $F = 512$, as in [14]. In addition, friendship relations among the users as well as participation of users to groups are available. Table 1 summarizes the dimensions of the dataset entities that are of interest in our study, e.g., the numbers of images, features, tags, and users.

Table 1. Cardinalities of dataset entities.

Images	1264
Features	512
Tags	2366
Users	440

4.2. Dataset Preprocessing

Several preprocessing steps were needed in order to train effectively the PARAFAC2 model described in Section 3. The first step was to exclude any images that were not annotated at all, yielding finally $I = 1264$ images.

Since the images were manually annotated by the users of the social network, the image-tag matrix was sparse. In order to reduce the sparsity of the image-tag matrix, we applied the Nearest Neighbour algorithm to the columns of the image-feature matrix $\mathbf{X}^{(1)}$ in order to find the 10 nearest neighbors to the feature vector of each image in the training set. Accordingly, each image inherits the tags of its 10 nearest neighbours.

The initial image-user matrix was also sparse, since each image was uploaded by one user, as happens to any social network. To obtain a more dense image-user matrix, we exploited the friendship relations between the users. More specifically, if a relation between an image and a user exists, this relation is inherited by his or her friends. This inheritance is justified on the grounds that friends share the same interests, thus they upload similar images, and more likely would like to see such images. The latter assumption is quite important for effective image recommendation.

4.3. Evaluation protocol and metrics

In order to evaluate the proposed tagging method, the dataset was randomly split into a training and test set at a ratio 60% and 40%, respectively. During the experimental evaluation, 10-fold cross validation was employed. The length of the tag vector returned by the system was 10, i.e., each test image was automatically annotated with a set of 10 tags. The length of the user vector returned by the system was 3, i.e., each test image was recommended to 3 users. The number of returned tags and users was chosen based on the characteristics of the specific dataset.

The mean per word precision, the mean per word recall, and the mean F_1 measure, i.e., the averaged harmonic mean of per word precision and recall were employed as metrics to evaluate the proposed image tagging. Their definitions can be found in [17, 18], but they are repeated briefly given next for completeness. For each word/tag t in the vocabulary of size V, let us denote by $|t_{GT}|$ the number of test images that have been annotated with the word t by the users of the social network. Also, let us denote by $|t_M|$ the number of test images that have been annotated with the word t by the proposed method. If we denote by $|t_C|$ the number of images that are correctly annotated with t, e.g., the images for which the users' annotation and the proposed method's annotation are identical, then the per word recall (RLC) is $\frac{|t_C|}{|t_{GT}|}$, while the per word precision (PRC) is $\frac{|t_C|}{|t_M|}$. If the method never annotates an image with the word t, then the per word precision is undetermined. In such a case, the per word precision can be estimated by the appearance frequency of the word t in the training images' annotations. The F_1 measure is defined as

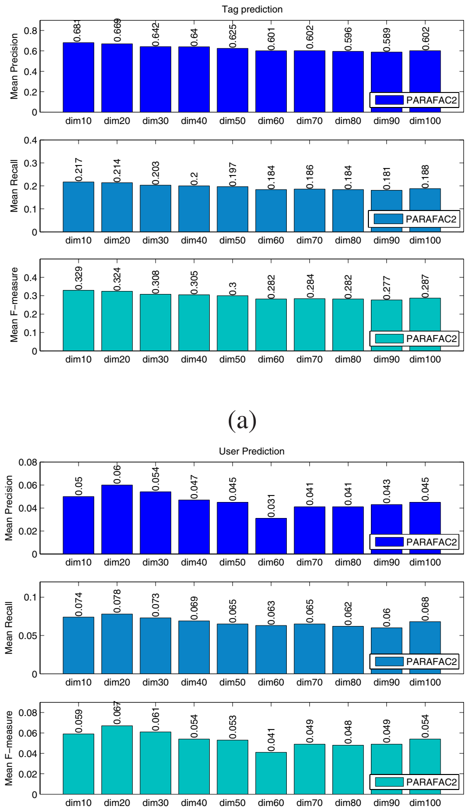
$$F_1 = \frac{2\,PRC\;RLC}{PRC + RLC}. \tag{5}$$

The F_1 measure takes values between 0 and 1. The higher F_1 measure values the more effective a tagging method is. In any fold, the mean per word recall and precision are calculated across all vocabulary words and (5) is computed. The mean F_1 measure is obtained by averaging the F_1 measures across the 10 folds. The former definitions can be easily adapted to image recommendation.

4.4. Results

We applied PARAFAC2 with different numbers of latent variable dimensions k. In Figure 1, the mean per word (user) precision, the mean per word (user) recall, and the mean F_1 measure for image tagging (user prediction) are plotted for various values of k. PARAFAC2 is shown to be a computationally friendly method, since it allows for great dimensionality reduction without any deterioration of its efficiency.

Even though the range of precision and recall is $[0, 1]$, the aforementioned metrics may be upper-bounded by a value less than 1, if the number of tags appearing in the ground truth annotation is either greater or less than the number of tags that are returned by the automatic image annotation system. The same applies to the predicted user vector. In Figure 2, the mean values of evaluation metrics are presented for $k = 20$. The performance of the PARAFAC2 model is compared to that of two baseline models, the Random and the UpperBnd model [19]. These models give a sense of the actual range for each metric.



(a)

(b)

Fig. 1. PARAFAC2 model prediction metrics for different dimensions of latent variables k on the dataset used in [14]: (a) Mean per word precision, mean per word recall, and mean F_1 measure for image tagging; (b) Mean per user precision, mean per user recall, and mean F_1 measure for user prediction.

The Random model sets the lower limit of the values admitted by the metrics on a given dataset. It samples words (without replacement) from a multinomial distribution parameterized by the word prior distribution, $P(i), i = 1, 2, \ldots, V$ estimated using the observed word counts in the training set [19]. Therefore, the tag selection according to the Random model relies on the tag appearance frequency, such that the most common tags are more likely to be chosen to annotate an image. Apparently, the Random model for the user prediction follows the same procedure.

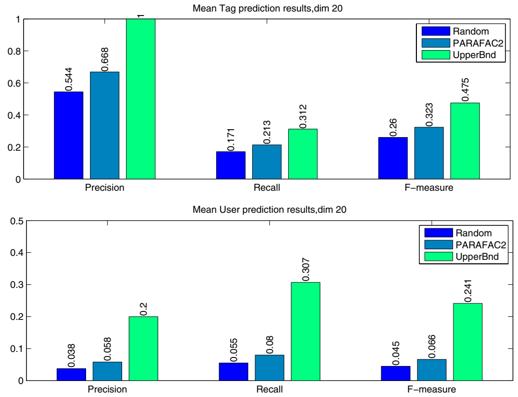


Fig. 2. Mean per tag (per user) metrics for the PARAFAC2 model (k=20) against the same metrics for the Random and the UpperBnd models [19] measured on the database used in [14].

On the other hand, the UpperBnd model sets the upper limit admitted by each metric on a given dataset and indicates the best possible performance [19]. The UpperBnd model uses the ground truth tags of annotated images. As stated before, the model predicts a fixed number of tags. If the desired number of predicted tags is smaller than the ground truth tags of the image, a subset of the ground truth tags is randomly selected. Similarly, if the ground truth annotation contains too few tags, tags are randomly added to the annotation from the rest of the vocabulary. The model is easily adapted to image recommendation as well.

In Table 2, quantitative results of the PARAFAC2 performance on automatic image tagging and recommendation are summarized. The PARAFAC2 performance is compared with that of the Random and UpperBnd models. The reported performance metrics are means and standard errors inside parentheses. Standard error is evaluated by dividing the sample standard deviation with the sample size. The performance metrics were evaluated using 10-fold cross-validation for latent variable dimension $k = 20$.

Table 2. Mean prediction results on the dataset used in [14].

		Tag prediction		
System	**Protocol**	**Precision**	**Recall**	**F_1 measure**
PARAFAC2	10FCV, V =2366, U=440	**0.668 (0.002)**	**0.213 (0.0008)**	**0.323 (0.001)**
Random [19]	10FCV, V =2366, U=440	0.544 (0.0007)	0.171 (0.0002)	0.260 (0.0003)
UpperBnd [19]	10FCV, V =2366, U=440	1 (0)	0.312 (0.0003)	0.475 (0.0003)
		User prediction		
System	**Protocol**	**Precision**	**Recall**	**F_1 measure**
PARAFAC2	10FCV, V =2366, U=440	**0.058 (0.002)**	**0.08 (0.0008)**	**0.066 (0.0003)**
Random [19]	10FCV, V =2366, U=440	0.038 (0.0006)	0.055 (0.0006)	0.045 (0.0006)
UpperBnd [19]	10FCV, V =2366, U=440	0.2 (0.001)	0.307 (0.002)	0.241 (0.0006)

By inspecting Table 2 and Figure 2, PARAFAC2 clearly exhibits better performance than the Random model with respect to the per-word (per-user) precision, per-word (per-user) recall, and F_1 measure using 10-fold cross-validation.

To test whether the evaluation metrics differences between the PARAFAC2 and the baseline models are statistically significant, we apply the approximate analysis in [20]. For the tag prediction, the per word recall and per word precision differences between the PARAFAC2 and the Random models as well as the PARAFAC2 and the UpperBnd models are found to be significant at the 95% level of significance. For the user prediction (image recommendation), the per user recall and per user precision differences between the PARAFAC2 and the UpperBnd models are found to be significant at the 95% level of significance. Furthermore, the per user recall differences between the PARAFAC2 and the Random models are found to be significant at the 95% level, while the per user precision differences are found to be significant at the 90% level of significance.

Although for user prediction (i.e., image recommendation), PARAFAC2 almost doubles the values admitted by the metrics when the Random model is applied, there is plenty of room for improvement. The low values admitted by the metrics are attributed to the dataset structure. As is previously said, each image was uploaded by only one user and we artificially increased the image-users relations by exploiting the friendship relations among the users. As a fact, in the specific dataset, few users had a wide circle of friends, while most users had declared only a few friendship connections. If we take into account the PARAFAC2 performance on tag prediction, we are quite certain that had the friendship relationships been more dense, higher absolute values would have been achieved by the PARAFAC2 model.

To this end, it would be very interesting to examine the performance of the proposed image annotation and recommendation system on different dense datasets. Also, better performance may be obtained by adding more matrices to the PARAFAC2 model, which can capture, say the users' preferences. For example, one such matrix could be the matrix associating each user with the images the user has marked as favorites.

5. CONCLUSIONS

An appealing automatic image tagging and image recommendation system has been proposed. PARAFAC2 has been employed for semantically oriented feature extraction, multilabel image annotation, and image recommendation to multiple users. The inclusion of three slices in the PARAFAC2 model enables capturing the latent relations between the images features, the tags, and the user interests. Promising results have been reported.

ACKNOWLEDGMENT

This research has been co-financed by the European Union (European Regional Development Fund - ERDF) and Greek national funds through the Operation Program Competitiveness-Cooperation 2011 - Research Funding Program: 11SYN-10-1730-ATLAS.

6. REFERENCES

[1] S. Lindstaedt, R. Mörzinger, R. Sorschag, and V. Pammer, "Automatic image annotation using visual content and folksonomies," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 97–113, March 2009.

[2] F. Wu, Y. Han, Q. Tian, and Y. Zhuang, "Multi-label boosting for image annotation by structural grouping sparsity," in *Proc. ACM Int. Conf. Multimedia*, Florence, Italy, October 2010, pp. 15–24.

[3] C. Wang, S. Yan, L. Zhang, and H. J. Zhang, "Multi-label sparse coding for automatic image annotation," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Miami, FL, USA, June 2009, pp. 1643–1650.

[4] Y. Yang, Y. Yang, Z. Huang, H. T. Shen, and F. Nie, "Tag localization with spatial correlations and joint group sparsity," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Providence, RI, USA, June 2011, pp. 881–888.

[5] H. Ma, J. Zhu, M. R. T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, August 2010.

[6] H. Wang, H. Heng, and C. Ding, "Image annotation using bi-relational graph of images and semantic labels," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Providence, RI, USA, June 2011, pp. 793–800.

[7] X. Zhang, X. Zhao, Z. Li, J. Xia, R. Jain, and W. Chao, "Social image tagging using graph-based reinforcement on multi-type interrelated objects," *Signal Processing*, vol. 93, no. 8, pp. 2178–2189, August 2013.

[8] X. Ning and G. Karypis, "SLIM: Sparse linear methods for top-n recommender systems," in *Proc. 11th IEEE Int. Conf. Data Mining*, Vancouver, BC, Canada, December 2011, pp. 497–506.

[9] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. ACM Int. Conf. Multimedia*, Florence, Italy, October 2010, pp. 461–470.

[10] Y. Panagakis and C. Kotropoulos, "Automatic music tagging via PARAFAC2," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011, pp. 481–484.

[11] P. Chew, B. Bader, T. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *Proc. 13th ACM Int. Conf. Knowledge Discovery and Data Mining*, San Jose, CA, USA, August 2007, pp. 143–152.

[12] A. Oliva and A. Torralba, "Building the GIST of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[13] R. A. Harshman, "PARAFAC2: Mathematical and technical notes," *UCLA Working Papers in Phonetics*, vol. 22, pp. 30–47, 1972.

[14] K. Pliakos and C. Kotropoulos, "Simultaneous image tagging and geo-location prediction within hypergraph ranking framework," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 6894–6898.

[15] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, September 2009.

[16] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an explanatory multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[17] M. Hoffman, D. Blei, and P. Cook, "Easy as CBA: A simple probabilistic model for tagging music," in *Proc. 10th Int. Conf. Music Information Retrieval*, Kobe, Japan, October 2009, pp. 369–374.

[18] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Sparse multi-label linear embedding within nonnegative tensor factorization applied to music tagging," in *Proc. 11th Int. Conf. Music Information Retrieval*, Utrecht, The Netherlands, August 2010, pp. 393–398.

[19] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, February 2008.

[20] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik, "What size test set gives good error rate estimates?," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 52–64, January 1998.