Simultaneous Image Clustering, Classification and Annotation for Tourism Recommendation

Konstantinos Pliakos and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki, 54124, Greece
{kpliakos,costas}@aiia.csd.auth.gr
http://www.aiia.csd.auth.gr

Abstract. The exponential increase in the amount of data uploaded to the web has led to a surge of interest in multimedia recommendation and annotation. Due to the vast volume of data, efficient algorithms for recommendation and annotation are needed. Here, a novel two-step approach is proposed, which annotates an image received as input and recommends several tourist destinations strongly related to the image. It is based on probabilistic latent semantic analysis and hypergraph ranking enhanced with the visual attributes of the images. The proposed method is tested on a dataset of 30000 images bearing text information (e.g., title, tags) collected from *Flickr*. The experimental results are very promising, as they achieve a top rank precision of 80% for tourism recommendation.

Keywords: Probabilistic Latent Semantic Analysis (PLSA), Clustering, Image Classification, Image Annotation, Recommendation systems, Hypergraph

1 Introduction

Nowadays, the continuously rising popularity of photo-sharing web applications leads to a huge amount of uploaded images. Browsing through this volume resorts to search engines, which exploit mainly the text information in tags, titles, etc. Image tags are keywords, which are added to an image by a user of a social media platform, describing the image content from this user's point of view. The aforementioned image annotation is a very critical procedure, as it is responsible for search engine retrieval accuracy and contributes to the organization of the images uploaded to the web. It aims at bridging the gap between the semantic and visual content of an image. However, it suffers from several limitations, such as spam, lack of uniformity, and noise. Several times, the tags given to an image by a user are far from being accurate, containing much redundancy, or even false information. Therefore, an automated annotation system is of paramount importance. Recently, besides annotation, much progress has been made toward developing new recommendation systems. However, achieving satisfactory efficiency or accuracy remains still an open problem.

Tourism is a vital economic sector for Greece and many other countries. Nowadays, the way people decide their tourist destination differs from the past.

It is no longer solely based on brochures or simple search on the web. The sectors of e-tourism and marketing are thriving and the need for developing efficient tourism recommendation systems is indisputable. Here, a tourism related recommendation system is presented and experimental results are disclosed, demonstrating its great potential.

Our work was motivated by [1] where the problem of vast amount of images was handled by building an Internet landmark recognition engine, resorting to efficient object recognition and unsupervised clustering techniques. In [2], a cluster-based landmark and event detection scheme was presented that was based on clustering performed on both visual and tag similarity graphs. More relevant to our approach are the methods presented in [3] and [4]. A worldwide tourism recommendation system was implemented based only on visual matching and minimal user input in [3] and a probabilistic model was developed in [4] that was based on Latent Dirichlet Analysis (LDA) for simultaneous image classification and annotation.

The novel contribution of this paper is in the development of a complete image annotation and tourism recommendation system. In particular, the GPS coordinates (latitude, longitude) of a dataset of 30000 geo-tagged images crawled from *Flickr* were clustered by means of hierarchical clustering to form 2993 clusters. Hereafter, these clusters are referred to as geo-clusters. From them, the 100 most dense geo-clusters were selected as places of interest (POI). Indeed, popular tourist destinations attract more visitors, who upload more geo-tagged images on social media sharing platforms. The text information (e.g., titles, tags) of the images that belong to each geo-cluster was concatenated, forming a geo-cluster derived document. Next, probabilistic latent semantic analysis (PLSA) [5], [6], [7] properly initialized was employed to build an image annotation sub-model. PLSA performs a probabilistic mixture decomposition, which associates an unobserved class variable to co-occurrence of terms and documents. That is, the PLSA is used to represent documents as probability distributions of topics treated as unobserved class variables. By applying PLSA to a term-document matrix the relations between the terms and the documents are captured by observing the probability distribution between the documents and the generated topics and between the topics and terms. Here, the PLSA is applied in a term-document (geo-cluster) matrix and the annotation is performed by assigning the geo-cluster derived document and all the images belonging to the geo-cluster the most strongly related terms to it. The annotation sub-model of the proposed system is also enhanced by exploiting the visual attributes of images. Such an approach is more complete than that in existing methods, such as [3] and [4]. Next, a hypergraph was constructed, capturing the relations computed by the PLSA between the geo-cluster derived documents and the topics as well as the vocabulary terms. A hypergraph is defined as a set of vertices made by concatenating different kind of objects (e.g., documents, topics, terms) and hyperedges linking these vertices. In contrast to simple graphs, multi-link relations between the vertices are captured in hypergraphs. Tourism recommendation is treated as a hypergraph ranking problem and the 5 top ranked geo-clusters are

recommended as tourist destinations. For evaluation purposes, 200 images were removed from the dataset along with their text information to be used for recommendation assessment. The experiments demonstrate the merits of the proposed system. Both classification and annotation results are very promising. A top rank precision of 80% is disclosed for tourism recommendation.

The remainder of this paper is organized as follows. In Section 2, the image annotation is detailed. The hypergraph ranking model is analyzed in Section 3. The dataset and the term-document matrix are described in Section 4. Hypergraph construction is explained in Section 5. The outline of the proposed system is presented in Section 6. In Section 7, experimental results are presented demonstrating the effectiveness of the proposed method. Conclusions are drawn and topics for future research are indicated in Section 8.

2 Image Annotation

2.1 Image Annotation Using Semantic Topics

In text processing, PLSA models each term in a document as a sample from a mixture model. The mixture components are multinomial random variables that can be interpreted as topic representations. The data generation process can be described as follows: 1) select a document d with probability $P(d)$, 2) pick a latent topic z with probability $P(z|d)$ and 3) generate a term t with probability $P(t|z)$. The joint probability model is defined by the mixture:

$$\left.\begin{aligned} P(t,d) &= P(d)P(t|d) \\ P(t|d) &= \sum_{z\in Z} P(t|z)P(z|d) \end{aligned}\right\} \tag{1}$$

where $t \in T = \{t_1, t_2, \cdots, t_k\}$ and $d \in D = \{d_1, d_2, \cdots, d_m\}$ represent the vocabulary terms and documents, respectively, while $z \in Z = \{z_1, z_2, \cdots, z_n\}$ is an unobserved class variable representing the topics. As it is indicated in (1), the document specific term distribution $P(t|d)$ is a convex combination of the n topic conditional distributions $P(t|z)$. The annotation procedure is performed as follows:

1 PLSA is applied to a term-document matrix $\mathbf{A} \in \mathbb{R}^{k\times m}$.
2 For each document to be annotated, the most related topic is chosen, that with the highest probability.
3 The 30 most related terms to that topic are employed to annotate the document.

Terms providing geographical information are identified using geo-gazetteers[1]. Thus, a complete annotation model, which provides, geographic and semantic information is obtained.

[1] http://www.geonames.org

2.2 PLSA Initialization

PLSA depends on proper initialization method. In addition to the common random initialization, there are many other schemes, e.g., the Random C (RC) [6]. A variant of RC is the Dense Random C (DRC) summarized in Algorithm 1. The DRC treats the columns of $\mathbf{A}$ unequally. Only the densest columns are chosen, as they provide more valuable information. The reduction of the number of the columns makes the method less time consuming. The DRC was found to be more effective than the RC in the experiments conducted.

Algorithm 1 Dense Random C Initialization

Input: matrix $\mathbf{A} \in \mathbb{R}^{k \times m}$ with $A(i, j,) \geq 0$.
Output: matrix $\mathbf{S} \in \mathbb{R}^{k \times n}$, containing the conditional probabilities $P(t|z)$.

1 Count the non-zero elements of each column of $\mathbf{A}$.
2 Compute the mean document vector $\bar{\mu}$.
3 Find the c columns of $\mathbf{A}$, having more non-zero elements than $\bar{\mu}$.
4 Average x randomly chosen columns out of the c and set the average column vector as a column of $\mathbf{S}$. Repeat 3-4 for all columns of $\mathbf{S}$.

2.3 Classification-based Visual Content Annotation

The visual features of an image provide valuable, complementary, information about its content. Image annotation is strongly related to image classification, considering the class label as a global description of the image, while the tags are treated as local description of the individual image parts. Here, 13 seed images, which represent 13 different visual topics were chosen manually from the dataset. For each of them, the 5 nearest neighbor images in the dataset were located by means of the k-Nearest Neighbor algorithm (k-NN), resorting to the distance between GIST descriptors [8] of the seed image and any image in the dataset. This way, 13 classes were formed, each of them containing 6 images. The average GIST descriptor among the 6 images that belong to each class defines a template for each class. Each visual topic (class) was assigned manually one label and a few representative tags, e.g., clouds, sky, sea, sunset, defining the image visual content. Each test image is classified into one of the 13 classes by finding the minimum distance between its GIST descriptor and the templates of the 13 classes. Representative images assigned to different classes are shown in Fig. 1.

3 Tourism Recommendation

The second part of the proposed system consists of a hypergraph model representing the multi-link relations between terms of the vocabulary, documents (geo-clusters), and topics as they were computed in Sec. 2.1.

Fig. 1. A sample of 12 images, representing different classes.

Hereafter, set cardinality is denoted by $|\cdot|$, the ℓ_2 norm of a vector appears as $\|.\|_2$ and $\mathbf{I}$ is the identity matrix of compatible dimensions. A hypergraph $\mathbf{H}$ is a generalization of a graph with edges connecting more than two vertices. $\Psi(V, E, w)$ denotes a hypergraph with set of vertices V and set of hyperedges E to which a weight function $w : E \to \mathbb{R}$ is assigned. V consists of sets of objects of different type (documents, topics, terms). A $|V| \times |E|$ incidence matrix is formed having elements $H(v, e) = 1$ if $v \in e$ and 0 otherwise. Based on $\mathbf{H}$, the vertex and hyperedge degrees are defined as:

$$\left.\begin{array}{l} \delta(v) = \sum_{e \in E} w(e) H(v, e) \\ \delta(e) = \sum_{v \in V} H(v, e) \end{array}\right\}. \tag{2}$$

The following diagonal matrices are defined: the vertex degree matrix $\mathbf{D}_u$ of size $|V| \times |V|$, the hyperedge degree matrix $\mathbf{D}_e$ of size $|E| \times |E|$, and the $|E| \times |E|$ matrix $\mathbf{W}$ containing the hyperedge weights.

Let $\mathbf{\Theta} = \mathbf{D}_u^{-1/2}\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T\mathbf{D}_u^{-1/2}$, then $\mathbf{L} = \mathbf{I} - \mathbf{\Theta}$ is the positive semi-definite Laplacian matrix of the hypergraph. The elements of $\mathbf{\Theta}$, $\Theta(u, v)$, indicate the relatedness between the objects u and v. In order to compute a real valued ranking vector $\mathbf{f} \in \mathbb{R}^{|V|}$, one minimizes

$$\Omega(\mathbf{f}) = \frac{1}{2}\mathbf{f}^T\mathbf{L}\mathbf{f}, \tag{3}$$

requiring all vertices with the same value in the ranking vector $\mathbf{f}$ to be strongly connected [9]. The aforementioned optimization problem was extended by including the ℓ_2 regularization norm between the ranking vector $\mathbf{f}$ and the query vector $\mathbf{y} \in \mathbb{R}^{|V|}$ in music recommendation [10]. The function to be minimized is expressed as

$$\tilde{Q}(\mathbf{f}) = \Omega(\mathbf{f}) + \vartheta \, ||\mathbf{f} - \mathbf{y}||_2^2 \tag{4}$$

where ϑ is a regularizing parameter. The best ranking vector $\mathbf{f}^* = \arg\min_{\mathbf{f}} \tilde{Q}(\mathbf{f})$ is [10]:

$$\mathbf{f}^* = \frac{\vartheta}{1+\vartheta}\left(\mathbf{I} - \frac{1}{1+\vartheta}\mathbf{\Theta}\right)^{-1}\mathbf{y}. \tag{5}$$

4 Dataset and Term-Document Matrix

Popular tourist destinations attract more visitors, who upload more geo-tagged images on social media sharing platforms. To properly organize such geo-tagged images into geographical clusters, an hierarchical clustering algorithm, based on geographical distances computed with the "Haversine formula"[2] was applied. Thus, from 30000 geo-tagged randomly selected images related to Greece, that were collected from *Flickr*, 2993 geo-clusters were formed. The 100 most dense geo-clusters were considered as places of interest. Next, a document was created of each geo-cluster comprising the concatenation of all text information (e.g., title, tags) available for all the images assigned to the geo-cluster.

A vocabulary was defined by processing the text information contained in a dataset of 150000 images, in order to properly capture the context of the tourism application. Prior to vocabulary extraction, all characters were converted to lower case and unreadable or redundant information was removed. A vocabulary of unique words was generated along with their frequencies and terms with frequency less than 100 were removed from the vocabulary. By doing so, a vocabulary of 1901 terms was finally retained.

Next, having created 100 documents and having set the vocabulary of terms, a term-document matrix $\mathbf{A}$ was formed with size 1901×100. Each element $A(i,j)$ corresponds to the number of occurrences of a term i in a document (geo-cluster) j. In order to proceed to the image annotation, the PLSA was applied to $\mathbf{A}$.

5 Hypergraph Construction

The vertex set is defined as $V = Doc \cup Top \cup Ter$, where Doc, Top, Ter correspond to documents (geo-clusters), topics, and vocabulary terms, respectively. The hypergraph $\mathbf{H}$ is formed by concatenating the hyperedge set. It has a size of 2201×100 elements. It is formed by the concatenation of 100 documents, 200 topics, and 1901 vocabulary terms capturing the multi-link relations among the 100 geo-cluster derived documents. The weights of the hyperedge set are set equal to one.

As was mentioned in Sec. 2.1, each document is represented by a probability distribution on a set of topics. For accuracy and simplicity reasons, although each document can be related to more than one topics, only the relation between the document and the topic corresponding to the highest probability $P(z|d)$ is represented in the hypergraph. Then, the relations between any topic and the 15 most strongly related terms to that topic are retained.

[2] http://www.movable-type.co.uk/scripts/latlong.html

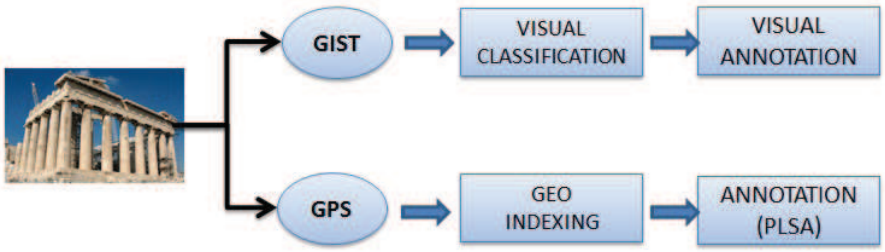

Fig. 2. Annotation system.

The query vector $\mathbf{y}$ is initialized by setting the entry corresponding to the target document (geo-cluster) g, where the input image was assigned, to 1 and all other objects v connected to the specific document to $\Theta(g, v)$. It is underlined, that $\Theta(i, j)$ is the element of $\mathbf{\Theta}$ which corresponds to the objects i and j and it is a relatedness measure of the 2 connected objects. The query vector $\mathbf{y}$ has a length of 2201 elements.

The ranking vector $\mathbf{f}^*$ is derived by solving (5), after setting the values of the query vector $\mathbf{y}$. It has the same size and structure as $\mathbf{y}$. The values corresponding to documents (geo-clusters) are used for tourist destination recommendation with the top ranked geo-clusters being recommended as tourist destinations to the user, who has imported the input image.

6 System Outline

Fig. 2 demonstrates the proposed system annotation. Given an image as input, the distances between image geo-location captured using GPS technology and the geo-cluster centers are computed. The input image is then assigned to the nearest geo-cluster. Simultaneously, the image visual content is classified by means of a nearest neighbor algorithm fed by the image GIST descriptor [8]. Next, the class label and the predefined representative tags offer a visual content annotation. Simultaneously, the vocabulary terms assigned to the closest geo-cluster derived document by the PLSA, offer geographic and semantic annotation for this image, as was demonstrated in Sec. 2.1. Proceeding to tourism recommendation, the query vector $\mathbf{y}$ is initialized, as was mentioned in Sec. 5. Hypergraph ranking is applied and the 5 top ranked geo-clusters are recommended as tourist destinations.

7 Experimental Results

The averaged Recall-Precision curve is used as figure of merit. Precision is defined as the number of correctly recommended objects divided by the number of all recommended objects. Recall is defined as the number of correctly recommended objects divided by the number of all objects.

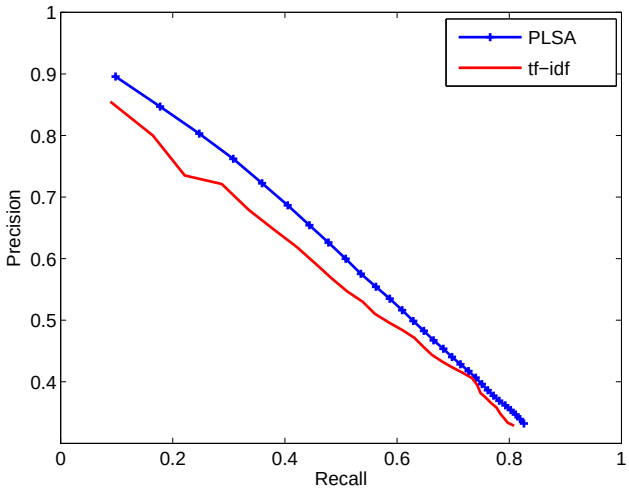

Fig. 3. Recall-precision of PLSA compared to that of tf-idf.

For evaluation purposes, a test set containing 200 images was randomly chosen and removed from the training set along with their text information. As is demonstrated in Fig. 3, the PLSA outperforms the term frequency-inverse document frequency (TF-IDF) method [11]. TF-IDF is a classical global weighting scheme for vector space model, where terms appearing in documents are weighted proportionally to term frequency and inversely proportional to the document frequency.

In Fig. 4, recall-precision curves are plotted for the PLSA, having been initialized by the RC and the DRC for 5 and 10 iterations. The results indicate the superiority of DRC over RC for the same number of iterations. For evaluation purposes, the average recall precision curves over 1000 repetitions of the PLSA training for each initialization are shown.

Furthermore, it was noted that better results are obtained by increasing the number of topics, as can be seen in Fig. 5. This may be attributed to the fact that geo-cluster derived documents may contain multiple topics. Indeed, these documents consist of the tags of many photos taken by several people and each one may possess multiple semantic topics. The recall-precision curves were obtained by averaging recall-precision pairs in 100 repetitions.

Fig. 6 discloses the classification rates obtained for the 13 visual topic classes. It is seen that the proposed classification method performs extremely well for scenes of flying birds or cloudy sky. Good results are also obtained for all other classes. Clearly, the GIST features reaffirm the reputation of being the most effective features for scene matching tasks. However, their performance is not at the same level as in object recognition tasks.

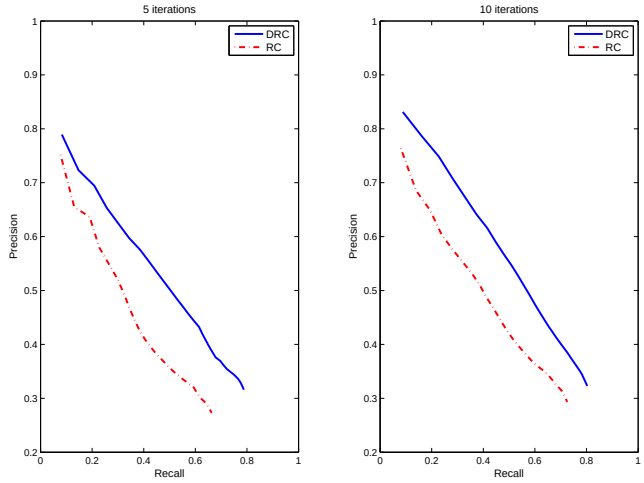

Fig. 4. RC and DRC recall-precision curves for 5 and 10 iterations.

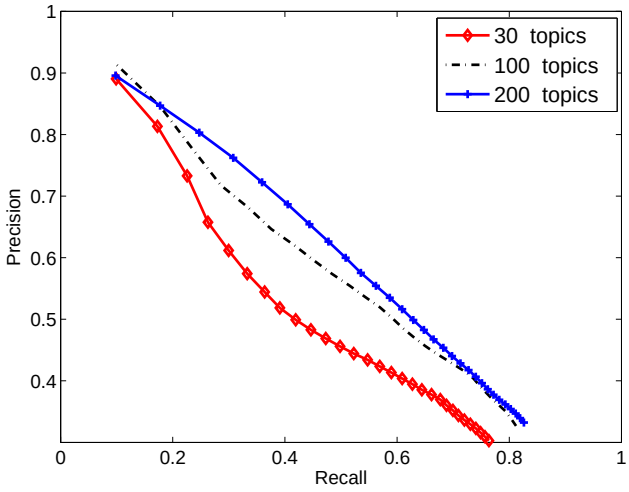

Fig. 5. PLSA recall-precision curves for 30, 100, 200 topics.

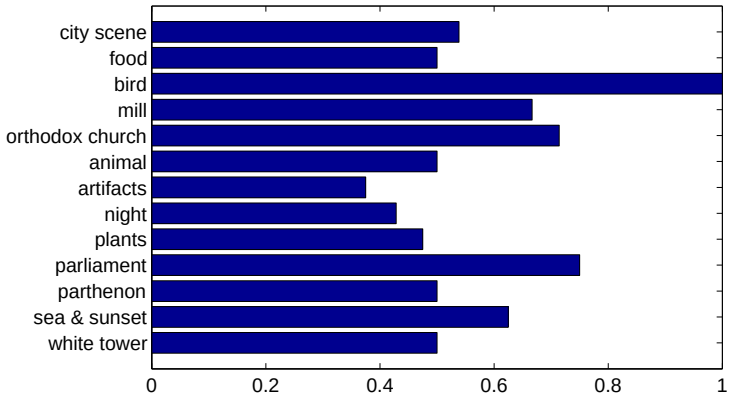

Fig. 6. Accuracy results of the visual image classification.

Fig. 7 demonstrates the accuracy of tourism recommendation for the 5 top ranking positions. The best results are obtained for the 1st ranking position. The precision does not degrade, falling below 60%, when additional ranking positions are taken into account, indicating the effectiveness of the proposed method.

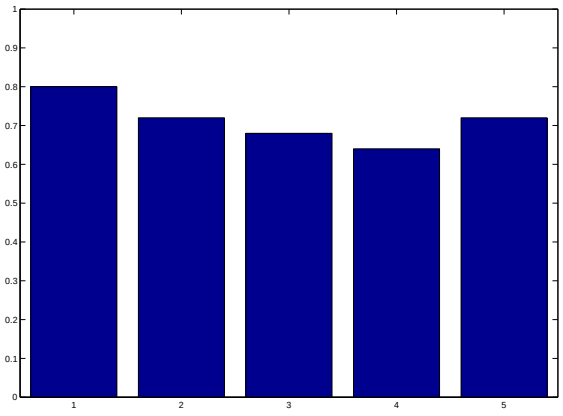

Fig. 7. Recommendation precision for the top 5 ranking positions.

8 Conclusions and Future Work

Here, a novel and efficient annotation and recommendation system was proposed. A method to organize large collections of images was developed based on clustering and classification. PLSA was enhanced by an effective initialization method and used in order to extract semantic information from image metadata. The annotation procedure was also supplemented, exploiting image visual attributes. Thanks to hypergraph learning, tourism recommendation has been implemented. The dataset used in the experiments can be expanded to cover the entire Greek territory. Several online updating methods can be applied to PLSA, improving system performance. Finally, the proposed system could be favored by the exploitation of the social media information, being available on the web in order to provide personalized recommendation.

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

References

1. Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine, in Proc. IEEE CVPR, pp. 1085–1092, 2009.
2. S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, Cluster-based landmark and event detection on tagged photo collections, Multimedia, pp. 52–63, 2011
3. L. Cao, J. Luo, A. Gallagher, A worldwide tourist recommendation system based on geotagged web photos, in Proc. IEEE ICASSP, pp. 2274–2277, Dallas, Texas, USA, 2010.
4. C. Wang, D. Blei, and L. Fei-Fei, Simultaneous image classification and annotation, in Proc. IEEE CVPR, pp. 1903–1910, Florida, USA, June 2009.
5. T. Hofmann, Probabilistic latent semantic analysis, in Proc. 15th Conf. Uncertainty in Artificial Intelligence, pp. 289–296, 1999.
6. N. Bassiou and C. Kotropoulos, On-line PLSA: Batch updating techniques including out of vocabulary words, IEEE Trans. Neural Networks and Learning Systems, 2014, to appear.
7. N. Bassiou and C. Kotropoulos, RPLSA: A novel updating scheme for probabilistic latent semantic analysis, Computer, Speech, and Language, vol. 25, no. 4, pp. 741–760, 2011.
8. 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.
9. S. Agarwal, K. Branson, and S. Belongie, Higher order learning with graphs, in Proc. 23rd ICML, pp. 17–24, 2006.
10. J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, Z. Lijun, and X. He, Music recommendation by unified hypergraph: Combining social media information and music content, in Proc. ACM Conf. Multimedia, pp. 391–400, 2010.
11. G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM, pp. 613–620, 1975.