Exploring Context and Content Links in Social Media: A Latent Space Method

Guo-Jun Qi, Charu Aggarwal, *Fellow*, *IEEE*, Qi Tian, *Senior Member*, *IEEE*, Heng Ji, *Member*, *IEEE*, and Thomas S. Huang, *Life Fellow*, *IEEE*

Abstract—Social media networks contain both content and context-specific information. Most existing methods work with either of the two for the purpose of multimedia mining and retrieval. In reality, both content and context information are rich sources of information for mining, and the full power of mining and processing algorithms can be realized only with the use of a combination of the two. This paper proposes a new algorithm which mines both context and content links in social media networks to discover the underlying latent semantic space. This mapping of the multimedia objects into latent feature vectors enables the use of any off-the-shelf multimedia retrieval algorithms. Compared to the state-of-the-art latent methods in multimedia analysis, this algorithm effectively solves the problem of sparse context links by mining the geometric structure underlying the content links between multimedia objects. Specifically for multimedia annotation, we show that an effective algorithm can be developed to directly construct annotation models by simultaneously leveraging both context and content information based on latent structure between correlated semantic concepts. We conduct experiments on the Flickr data set, which contains user tags linked with images. We illustrate the advantages of our approach over the state-of-the-art multimedia retrieval techniques.

Index Terms—Context and content links, latent semantic space, low-rank method, social Media, multimedia information networks.

1 Introduction

The development and popularity of Web 2.0 applications has made it much easier for millions of users to create and share their personal multimedia objects (**MO**s) than ever before. Many image and video sharing web sites have become extremely popular, as is evidenced by their burgeoning membership. Many such sites are built upon information and social network infrastructures such as Flickr, Youtube, and Facebook that connect millions of users with one another. Users are able to share their MOs with each other, and also provide the ability to tag each other's objects. Such sites represent a kind of rich multimedia information networks (**MIN**) [4] for social media [30], [21] in which the objects are linked to one another in the site with content links. By "content links," we refer to the visual and/or acoustic similarities between objects in a content feature space (see Fig. 1a). At the same time, the sharing process of such sites naturally creates Context Objects (**CO**s) because of the rich information provided by the different users directly or indirectly. Some examples of such COs are tags (e.g., user tag and geo-tags), related attributes (colors, textures, and even categories from weakly labeled data) [5], and users who share MOs as well as their queries connected to MOs by click-through records (see Fig. 1b). This helps to create an even richer MIN with context links which connect the MOs with their related COs. For example, the MOs clicked by users in the same query session probably contain the same semantic meaning. It is also the same for the MOs which share the same user tags[1] in MINs. It is often very useful for multimedia retrieval by mining the semantics in these context links. In this paper, we define a MIN as an information network with two kinds of semantic objects—*MOs* and *COs*. See Fig. 2 for an example. The MOs are connected in a relational graph structure, with both content and context relationships. While content relationships are directly useful for retrieval, the context relationships also contain rich semantic information which should be leveraged for effective retrieval.

In this paper, we show that a compact latent space can be discovered to summarize the semantic structure in MINs, which can be seamlessly applied in the state-of-the-art multimedia information retrieval systems (see Fig. 2 for an example). Specifically, this algorithm maps each MO into a latent feature vector that encodes the information in both context and content information. Based on these latent feature vectors, MOs can be effectively classified, indexed, and retrieved in a vector space by many mature off-the-shelf vector-based multimedia retrieval methods, like clustering, Re-Ranking [26], and Support Vector Machine (SVM) [20] for multimedia retrieval. Thus, our approach is a

- *G.-J. Qi and T.S. Huang are with the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801. E-mail: {qi4, huang}@ifp.uiuc.edu.*
- *C. Aggarwal is with the IBM T.J. Watson Research Lab, Yorktown Heights, NY 10598. E-mail: charu@us.ibm.com.*
- *Q. Tian is with the Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249. E-mail: qitian@cs.utsa.edu.*
- *H. Ji is with the Department of Computer Science, The City University of New York, New York, NY 10031. E-mail: hengji@cs.qc.cuny.edu.*

Manuscript received 28 June 2010; revised 6 July 2011; accepted 19 Aug. 2011; published online 1 Oct. 2011.
Recommended for acceptance by A. Torralba.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2010-06-0486.
Digital Object Identifier no. 10.1109/TPAMI.2011.191.

1. In this paper, we mainly concentrate on the context links associated with user tags. While the results in this paper are general enough to be applied to any kind of context links, we mainly focus on tag links because of the richness of their semantic information as compared to other kinds of context links.

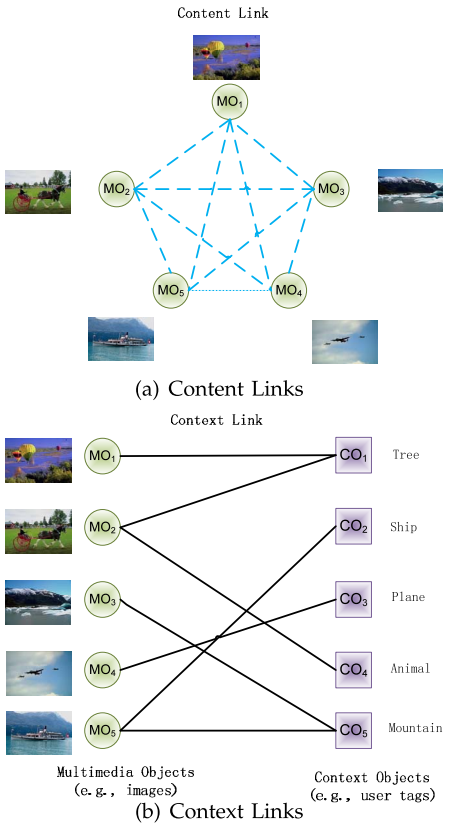


Fig. 1. Context and content links in MINs.

"general purpose technique" which can be leveraged to improve the effectiveness of a wide variety of techniques.

The general approach of learning latent semantic space has been extensively studied in the field of information retrieval. Popular techniques include Latent Semantic Indexing (LSI) [15], Probabilistic Latent Semantic Indexing (PLSI) [14], and Latent Dirichlet Allocation (LDA) [6]. These algorithms have also been applied to multimedia domain for problems such as indexing and retrieval [7], [17], [16], [29]. For example, [7], [17] learn latent feature vectors by LSI for natural scene images, and the learned features can be used effectively with general purpose SVM classifiers. Some preliminary results have shown the effectiveness of these algorithms; however, all these methods suffer from the problem with sparse context links, which we solve with the use of content links.

1. **Sparse context links.** These are the virtual links which are created as a result of user feedback (e.g., tags), and may be represented as the linkages between the MOs and the contextual objects such as tags. In the real-world contextual links, the number of user tags attached to an MO is usually quite small. In some extreme cases, only a few or even no tag may be attached to an object, which often leads to sparse contextual links. In such cases, it is hard to derive meaningful latent features for MOs because the determination of the correlation structure in the latent space requires a sufficient number of such contextual objects to occur together.

   A reasonable solution to this problem is to exploit the content links between MOs. In this paper, we will show how the content links can effectively complement the sparse contextual links by incorporating acoustical and/or visual information to discover the underlying latent semantic space.
2. **Omitting content information in LSI modeling of context links.** In this paper, content links represent the content similarities between MOs, i.e., those visually and/or acoustically similar objects are assumed to have strong content links between them. Content links contain important knowledge complementary to that embedded in context links. However, to the best of our knowledge, the existing latent space methods, LSI, PLSI, and LDA, cannot seamlessly incorporate the content and context links in a unified framework. Some attempts have been made to jointly model content and context information to learn the latent space [17], [29]. They quantify the MOs into visual words, which are treated in a way similar to some COs by linking them to MOs. However, such approaches greatly increase the number of parameters in the latent space model, and make it more prone to quantization-induced noise and overfitting due to the sparse context links.

   In contrast, we will show that content and context links can be seamlessly modeled to learn the underlying latent space. The content information does not have to be quantified into some discrete elements

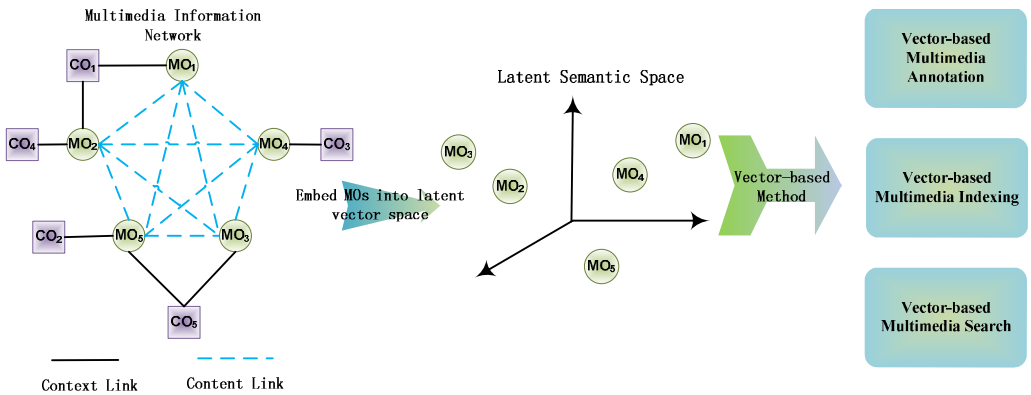


Fig. 2. Learning latent semantic space from context as well as content links simultaneously.

such as visual words described in [17]. Instead, the content link structure will be directly leveraged to discover latent features together with context links.

Therefore, we propose an elegant mapping of MINs to the latent space which can support an emerging paradigm of multimedia retrieval which unifies the information in context and content links. In other words, the goal of this approach is to annotate the images with some manually defined concepts, using visual and contextual features for learning a latent space. Specifically, by feeding the latent vectors into existing classification models, it can be applied to multimedia annotation, which is one of the most important problems in multimedia retrieval. Furthermore, we show a more sophisticated algorithm which can directly incorporate the discriminant information in training example for multimedia annotation without using mapping as a prestep. It jointly explores the context and content information based on a latent structure in the semantic concept space. Moreover, even given a new MO with no context links, this extended algorithm can still annotate it. This solves the out-of-sample problem and greatly extends the applicability of the algorithm in multimedia retrieval applications.

1.1 Related Work

Analysis and inference with multimodal data [2], [13], [19] is one of most important research topics in computer vision and patter recognition areas. Existing methods usually assume that in each data piece, there are a number of complementary cues associated with each other. For example, in a video clip, we observe a sequence of video frames as its visual cue, as well as the incident audio track. In the multimodal problem, the data in different modalities are always associated with each other. In other words, one data modality is always associated with its counterparts in another modalities. Many representative works concentrate on such a problem. SimpleMKL [19] addresses the multimodal problem by learning a linear combination of multiple kernels with a weighted 2-norm formulation. Bekkerman and Jeon [2] explore the multimodal nature of multimedia collections within the unsupervised learning framework. Guillaumin et al. [13] propose using semi-supervised learning to explore both labeled and unlabeled images in photo sharing websites while exploring the associated keywords in the text modality. Competitive results show these multimodal algorithms can gain much better performance as compared with single modal algorithms.

However, in social media applications, COs are not always associated with COs. For example, the new images in a test set usually do not have any accompanying user tags. In this case, multimodal methods cannot be applied due to the missing COs. We will discover the missing links between context and content objects, which is one of main problems we will address in this paper. In social media, structured MINs are the most natural data structure to represent the interaction between content and COs. This paper proposes a principled method to fuse the content and COs in such a social media network structure. We especially attempt to capture the links in MIN by embedding the content objects into a latent space. Similar linear embedding techniques like metric learning [28] have been proposed to reveal the underlying space structure. However, it is nontrivial to extend these embedding techniques to MIN. Perhaps the most relevant work is proposed by Blei et al. [6], who use a latent method for associating the annotated tags with the local regions in images. Its limitation is that this method can only assign existing user tags to images, but cannot handle the concepts beyond these tags.

The remainder of this paper is organized as follows: Section 2 reviews a set of state-of-the-art multimedia retrieval paradigms and motivates unifying both context and content links in social media. In Section 3, we briefly review the basic ideas of latent methods which are closely related to the proposed method. The proposed latent method is then detailed in Section 4. In Section 5, we develop an advanced algorithm for multimedia annotation by exploring the context and content information with the latent structure between the correlated semantic concepts for annotation. Experimental results are presented in Section 6 on a real-world multimedia data set crawled from Flickr. Finally, conclusions are made in Section 7.

2 Multimedia Retrieval Paradigms

In the following, we briefly review some existing multimedia retrieval paradigms and discuss the advantages of unifying analyses of both context and content links in social media. Based on whether context and/or content links are used, multimedia retrieval has evolved from the Content-based Multimedia Retrieval (**CMR**) [22] in the first paradigm, to the context-based multimedia retrieval (**CxMR**) in the second paradigm, and to the Context-and-Content-based Multimedia Retrieval (**C2MR**) as the latest paradigm.

2.1 Content-Based Multimedia Retrieval

The CMR approach attempts to model high-level concepts from low-level concepts extracted from the MOs. In a typical multimedia retrieval system like QBIC [12] and Virage [1], the query is formulated by some example MOs and/or text-based keywords. Then, the relevant MOs are retrieved based on their content features. The advantage of CMR is that it is an automatic retrieval approach. Once the concepts are modeled, no human labels are required to maintain it. However, due to the technical limit of artificial intelligence and multimedia analysis, its accuracy is often too low to output satisfactory retrieval results due to the semantic gap between low-level content features and high-level semantics.

2.2 Context-Based Multimedia Retrieval

With the development of Web 2.0 infrastructures, rich context links are often connected to MOs on the media-rich web sites such as Flickr, Youtube, and Facebook. In contrast to pure content information, these links provide extra semantic information to retrieve and index MOs in the Web environment. For a simple example, the images of "sea" and "sky" have similar color features which are difficult to distinguish by similarity in content feature space. However, by leveraging the user tags in their context links and mapping them into a new latent space by LSI, PLSI, and LDA, they can be distinguished with the semantics in their COs. Context-based Multimedia Retrieval (CxMR) approaches have been widely used in many practical multimedia search engines such as Google Images, which

utilize the context links such as surrounding text and user tags. Although the information in the context links is useful in many cases, they are often sparse and noisy. In some cases, it can lead to questionable performance, when the context contains much more irrelevant information to the mining process. This is often evident from the Google Image results when the images do not match the corresponding search at all.

2.3 Context-and-Content Multimedia Retrieval

Unifying the information in both context and content links is an appealing approach to solving the limits inherent in the two paradigms discussed above. Context links provide high-level semantic information which can be effective for resolving the ambiguity in the content feature space due to the semantic gap inherent in a pure content-based approach. Similarly, content links between MOs can serve as regularization, which can avoid the overfitting problem due to the sparse and noisy context links. The combination of two techniques provides the solution to effective multimedia retrieval in the rich Web 2.0 environment, which is so-called *Multimedia Retrieval 2.0*. This approach formulates multimedia retrieval by unifying the content and context-based approaches. As compared with the above existing multimedia retrieval systems, the advantages of our algorithm include:

1. We propose a general-purpose scheme which is broadly applicable. Many advanced vector-based retrieval systems can be seamlessly used with the proposed approach.
2. Context and content links are explored in a unifying framework. Hence, the learned latent space ought to be more optimal than the other methods which separately mine these two kinds of links in MINs.
3. Specifically, for the multimedia annotation problem, a more sophisticated algorithm is developed by leveraging the assumption that the semantic concepts for annotation are correlated and thus a latent structure exists in such a semantic concept space. Also, the context-and-content links are simultaneously explored to optimize the annotation performance.

3 Latent Semantic Indexing

In this section, we briefly review LSI, which is closely related to the algorithms proposed in the later section of this paper. In conventional methods for LSI, we map MOs to latent feature vectors. Suppose we have n MOs $\{d_1, d_2, \ldots, d_n\}$ and m COs $\{c_1, c_2, \ldots, c_m\}$ such as user tags. The context links between these n MOs and the m COs are denoted by a $n \times m$ matrix A. The elements $A_{i,j} \in \mathbb{R}^{n \times m}$ of this matrix represent the weights of context links, e.g., $A_{i,j} = 1$ if the jth CO is assigned to the ith MO, or $A_{i,j} = 0$ otherwise. The goal of LSI is to construct a set of feature vectors $\{X_1, X_2, \ldots, X_n\}$ in a latent semantic space $\mathbb{R}^k$ to represent these MOs. LSI performs a Singular Vector Decomposition (SVD) on the matrix A as follows:

$$A = U\Sigma V^T. \tag{1}$$

Here, U and V are orthogonal matrices such that $U^TU = V^TV = I$, and the diagonal matrix Σ has the singular values as its diagonal elements. By retaining the largest k singular values in Σ and approximating others to be zero, LSI creates an approximated diagonal matrix $\widetilde{\Sigma}$ with fewer singular values. This diagonal matrix is used to approximate A as $\hat{A} = U\widetilde{\Sigma}V^T$. Then the matrix $X = U\widetilde{\Sigma} \in \mathbb{R}^{n \times k}$ yields a new feature representation, each row of which is a k-dimensional feature vector of one MO, i.e., $X = [X_1 \;\; X_2 \;\; \cdots \;\; X_n]^T$. The computational complexity of SVD on the matrix A grows quadratically with the number of COs. If the content features extracted from MOs are quantified into description words (e.g., visual words) as COs, the computational cost will increase rapidly. On the other hand, as stated in Section 1, the link matrix A is usually quite sparse, with few context links. This may result in overfitting of the latent feature vectors since the small number of context links may not reflect the underlying correlation structure in a robust way.

PLSI is another algorithm which models the latent space by context links. Each MO is associated with a set of latent topic variables $\{h_1, h_2, \ldots, h_k\}$ with conditional probabilities $P(h_j|MO)$, $1 \le j \le k$. Similarly, for the latent topic h_l, the conditional probability of the context object CO_j is denoted by $P(CO_j|h_l)$. The conditional probability of CO_j given MO_i can be expressed as a product of these values:

$$P(CO_j|MO_i) = \sum_{l=1}^{k} P(CO_j|h_l)P(h_l|MO_i). \tag{2}$$

The probabilities $P(h_l|MO_i), P(CO_j|h_l)$, $1 \le l \le k$, can be estimated by using Maximum Likelihood (ML) and standard EM algorithms. We can use these to construct the latent feature vector $X(MO)$ of the multimedia object MO as follows:

$$X(MO) = [P(h_1|MO), P(h_2|MO), \ldots, P(h_k|MO)]^T. \tag{3}$$

PLSI has similar drawbacks as LSI because it does not consider the content links. Furthermore, the number of parameters in PLSI grows linearly with the number n of MOs. This suggests that the model is prone to overfitting [6] due to the sparse context links. Some alternative PLSI algorithms have been proposed for using context information during latent space discovery. They quantize the content features into COs (e.g., visual words) and use some extra conditional probabilities to model their relations with latent topics [29]. Although content information is used in such a model, it has many more parameters which need to be estimated. This results in overfitting.

LDA is another technique from this family of latent space methods. It assumes that the probability distributions of MOs over latent topics are generated from the same Dirichlet distribution [6]. This simplified assumption is key to avoiding the (large parameter) overfitting issue of PLSI. However, the simplifying assumption has the pitfall that the assumed Dirichlet distribution over MOs may not reflect their true distribution in the multimedia corpus.

While most of these algorithms focus on learning the latent space solely with context links, some efforts have been made to incorporate content information [31]. In order to incorporate content information into context analysis, it uses two separate matrices to factorize the content and context links (in addition to the latent matrix for MOs). However, it does not consider the geometric structure of the

distribution of MOs in the corpus. From a practical perspective, the extra latent matrix for either content or context links is unnecessary in multimedia retrieval. Instead, in this paper, we will learn a shared latent space from content and context links simultaneously so that it can mine the link structure in an integrated manner without introducing any additional model parameters. Moreover, the proposed formulation has a better optimization topology, i.e., it is a global convex optimization problem so that better numerical stability can be achieved.

We propose to model the geometric structure of MOs by their content links to capture their distribution in the underlying latent space. In other words, our intuitive assumption is that *the MOs with stronger content links ought to be closer to each other in the latent space*. By this assumption, the content links can be encoded into latent space together with context links.

4 Latent Space Modeling in Social Media

In this section, we propose methods for combining the content links with context links in order to discover the latent semantic space for MOs.

First, we show that the LSI problem is closely related to low-rank matrix approximation [8], [9]. Due to the noises in the context links, a noise term ε exist on the matrix A such that

$$A = H + \varepsilon. \tag{4}$$

Here, the matrix H denotes the noise-free context links, after the noise ε has been removed.

To derive H, some extra prior ought to be assumed on H. Inspired by LSI with a low-rank approximation of A, we impose a low-rank prior to recover H by minimizing the noisy term simultaneously as

$$\begin{aligned} &\min \|\varepsilon\|_F^2 + \gamma \mathrm{rank}(H) \\ &s.t., A = H + \varepsilon, \end{aligned} \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm (i.e., the squared summation of all elements in a matrix), γ is the balancing parameter, and $\mathrm{rank}(\cdot)$ is the rank function.

There is an intuitive interpretation for the low-rank prior. Let H_i, $1 \le i \le n$, denote the row vectors of H, which is the associated noise-free tag vector for the ith MO. Each tag vector represents the occurrence of the corresponding tag in the multimedia corpus. As illustrated in Fig. 3, the tag vectors of synonyms should be the same (or within a positive multiplier of one another), such as the tag vector H_{Person} and H_{Human} for the synonym terms "person" and "human." Moreover, many tags do not independently occur in the corpus since they are semantically correlated. For example, the tag "animal" often correlates with its subclasses such as "cat" and "tiger." This indicates, from the viewpoint of linear algebra, that the tag vector of "animal" could be located in a latent subspace spanned by those of its subclasses. Since the rank of matrix H is the maximum number of independent row vectors, it follows from the above dependency among tags that H ought to have a low rank structure. As revealed by the latent methods in the last section, user tags can be generated by mixing few latent topics. The topic vectors that represent occurrences of the associated topics in the multimedia corpus span a latent semantic space which contains most of the tag vectors. Therefore, the rank of H should be no more than the maximum number of independent topic vectors in the latent space. Hence, we can impose a low-rank prior to estimate the noise-free H from the observed noisy A.

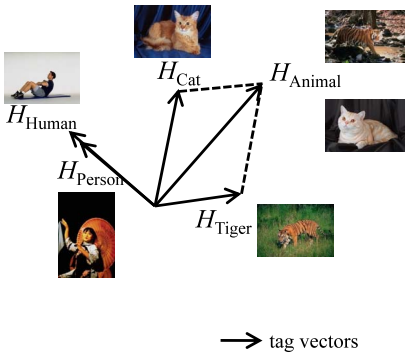


Fig. 3. Illustration of latent low-rank structure among the tag vectors.

It is NP-hard to directly solve the optimization problem of determining the lowest rank approximation [8]. Recently, nuclear norm was proposed as a convex surrogate for matrix rank [27], [8]. Its convexity is an advantage in being able to perform an effective optimization process. The norm is computed as the sum of all the singular values of the matrix. Let $\|A\|_*$ denote the nuclear norm of A, then $\|A\|_* = \sum_i \sigma_i(A)$ where $\sigma_i(A)$ are singular values of A. Then (5) can be rewritten as

$$\min_H \|A - H\|_F^2 + \gamma \|H\|_*. \tag{6}$$

The relationship between the above formulation and LSI can be presented more formally in the following result [8].

Theorem 1. *$\min_H \|A - H\|_F^2 + \gamma \|H\|_*$ has a unique analytical solution as $H_\gamma = U\mathrm{diag}((\sigma - \frac{\gamma}{2})_+)V^T$, where U, V, and $\mathrm{diag}(\sigma)$ form SVD for A as $A = U\mathrm{diag}(\sigma)V^T$. Here $\mathrm{diag}(\sigma)$ is a diagonal matrix with the singular values in vector σ as its diagonal elements. $(\sigma - \frac{\gamma}{2})_+$ is a component-wise operation that $(x)_+ = \max(0, x)$.*

The difference is that LSI directly selects the largest k singular values of A, but (6) subtracts $\frac{\gamma}{2}$ from each singular value and thresholds them by 0.

Suppose the resulting H is of rank k, then the SVD of H has form as $H = U\Sigma_k V^T$, where Σ_k is a $k \times k$ diagonal matrix. Similarly as LSI, the row vectors of $X = U\Sigma_k$ can be used as the latent vector representations of MOs in latent space. It is also worth noting that minimizing the rank of H gives a smaller k so that the obtained latent vector space can have lower dimensionality, and then the storage and computation in this space could be more efficient in practice.

However, (6) does not encode the content links, and the sparse context links may not result in a reliable latent space to represent MOs. Suppose we are given a matrix Q of content links, where $Q_{i,j}$ can represent the similarity measurement between the ith MO and the jth MO. For example, we can extract some low-level feature vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ from the visual and/or acoustic content of MOs; then $Q_{i,j}$ could be represented as follows:

$$Q_{i,j} = \exp\left\{-\frac{\|\mathbf{f}_i - \mathbf{f}_j\|^2}{\sigma^2}\right\}. \quad (7)$$

The relationship above uses Gaussian kernel with radius σ.

By linking all the MOs with Q, they can be embedded into a low-dimensional manifold structure [11], [3]. More specifically, we assume that *the MOs with stronger links ought to be closer to each other in the latent semantic space*. This assumption is analogous to the Laplace-Beltrami operator on manifolds [11], and makes a smooth regularization on the underlying geometric structure between MOs in the latent space. It can avoid the overfitting problem induced by sparse context links, and it can also incorporate the content links into modeling the latent space geometry. Based on this assumption, we introduce the quantity Ω to measure the smoothness of MOs in the underlying latent space:

$$\begin{aligned}\Omega(X) &= \frac{1}{2}\sum_{i,j=1}^{n} Q_{i,j}\|X_i - X_j\|_2^2 \\ &= \frac{1}{2}\sum_{i,j=1}^{n} Q_{i,j}\big(X_i - X_j\big)\big(X_i - X_j\big)^T.\end{aligned} \quad (8)$$

Here, $\|\cdot\|_2$ is l_2 norm, and X_i and X_j are the ith and jth row of X. It is easy to see that by minimizing the above regularization term, a pair of MOs with larger $Q_{i,j}$ will have closer feature vectors X_i and X_j in the latent space. With some matrix operations, $\Omega(X)$ can be further simplified as follows:

$$\begin{aligned}\Omega(X) &= \frac{1}{2}\sum_{i,j=1}^{n} Q_{i,j}\big(X_iX_i^T - X_iX_j^T - X_jX_i^T + X_jX_j^T\big) \\ &= \sum_{i,j=1}^{n} Q_{i,j}X_iX_i^T - \sum_{i,j=1}^{n} Q_{i,j}X_iX_j^T \\ &= \mathrm{trace}\big(XX^TD\big) - \mathrm{trace}\big(XX^TQ\big) \\ &= \mathrm{trace}\big(XX^T(D-Q)\big) \\ &= \mathrm{trace}\big(X^T(D-Q)X\big) = \mathrm{trace}\big(X^TLX\big).\end{aligned} \quad (9)$$

Here, D is a diagonal matrix with its elements as the sum of each row of Q, and $L = D - Q$ is the positive semidefinite Laplacian matrix. By using the factorization $H = XV^T$ and $V^TV = I$, we can simplify as follows:

$$\begin{aligned}\mathrm{trace}\big(H^TLH\big) &= \mathrm{trace}\big(VX^TLXV^T\big) \\ &= \mathrm{trace}\big(X^TLXV^TV\big) = \mathrm{trace}\big(X^TLX\big).\end{aligned} \quad (10)$$

Now we can formulate the new model to discover the latent semantic space by adding (10) into (6), which minimizes the following problem:

$$\min_H \mathcal{F}(H) = \|A - H\|_F^2 + \lambda\mathrm{trace}\big(H^TLH\big) + \gamma\|H\|_*. \quad (11)$$

Here λ is a tradeoff parameter. We note that the nuclear norm is convex, and L is a positive semidefinite matrix. Therefore, the above optimization problem has the desirable property that it is convex with a global optimum. Note that when there are images without any associated COs (e.g., testing images with no user tags), the term of the least-square error in the above equation is computed on the images with COs. It is the matrix completion problem in [8]. In this case, the second term plays the role of sharing and connecting the context knowledge between tagged and untagged images by their visual similarities.

It is worth noting that no links are established between COs in the above formulation. The reason we do not consider these links is that in order to link the COs (e.g., user tags), external knowledge is required to measure the similarity between them, such as WordNet and Google distance for linking textual user tags. Although these links can provide extra information, misleading knowledge may be introduced from the external resources, which do not comply with the visual evidence. For example, there is domain gap between text and visual similarities, and two textual tags that are strongly correlated in text documents may not co-occur in images. Thus, in the context of multimedia retrieval, we shall not incorporate context links in the formulation.

In contrast to (6), (11) does not have a closed-form solution. Fortunately, this problem can be solved by the Proximal Gradient method [25], which uses a sequence of quadratic approximations of the objective function (11) in order to derive the optimal solution. We define $K(H) = \|A - H\|_F^2 + \lambda\mathrm{trace}(H^TLH)$, and observe that $\mathcal{F}(H) = K(H) + \gamma\|H\|_*^2$ is summation of the differentiable function K and the nuclear norm. This helps in defining the update step as well. Given $H_{\tau-1}$ in the last step $\tau - 1$, it can be updated by solving the following optimization problem which quadratically approximates $\mathcal{F}(H)$ by Taylor expansion of $K(H)$ at $H_{\tau-1}$ [25]:

$$\begin{aligned}H_\tau = &\arg\min_H K(H_{\tau-1}) + \langle\nabla K(H_{\tau-1}), H - H_{\tau-1}\rangle \\ &+ \frac{\alpha}{2}\|H - H_{\tau-1}\|_F^2 + \gamma\|H\|_* \\ = &\arg\min_H \frac{\alpha}{2}\|H - G_\tau\|_F^2 + \gamma\|H\|_* \\ &+ K(H_{\tau-1}) - \frac{1}{2\alpha}\|\nabla K(H_{\tau-1})\|_F^2.\end{aligned} \quad (12)$$

Note that the last two terms in the rightmost side of the above equation do not depend on H_τ so they can be ignored when minimizing w.r.t. H_τ. The values of G_τ and α in the above expression are defined as follows:

$$\begin{aligned}G_\tau &= H_{\tau-1} - \frac{1}{\alpha}\nabla K(H_{\tau-1}) \\ &= H_{\tau-1} - \frac{2}{\alpha}\big(H_{\tau-1} - A + \lambda L^TH_{\tau-1}\big),\end{aligned} \quad (13)$$

$$\alpha = 2\sigma_{\max}\big(I + \lambda L^T\big), \quad (14)$$

where the coefficient α satisfies the Lipschitz condition such that $\|\nabla_R K(R) - \nabla_T K(T)\|_F \le \alpha\|R - T\|_F$ for any R, T, and $\sigma_{\max}(\cdot)$ denotes the largest singular value.

In each step, (12) provides an analytical solution to H_τ, as illustrated in Theorem 1. Algorithm 1 summarizes the optimization procedure.

Algorithm 1. Proximal Gradient for minimizing (11)

input A for the context links, Q for the content links, balance parameters λ and γ.

1 Initialize $H_0 \leftarrow 0$ and $\tau \leftarrow 1$.

2 Set $\alpha \leftarrow 2\sigma_{\max}(I + \lambda L^T)$.

repeat

2 Compute G_τ in (13).

3 Set $H_\tau \leftarrow U\mathrm{diag}(\sigma - \frac{\gamma}{\alpha})_+V^T$ which optimizes (12) by Theorem 1. Here $U\mathrm{diag}(\sigma)V^T$ gives the SVD of G_τ.

4 $\tau \leftarrow \tau + 1$.

until Convergence or maximum iteration number achieves.

5 Multimedia Annotation from Context and Content Links

Multimedia annotation plays the critical role in multimedia retrieval, and it aims at annotating semantic concepts to MOs. As mentioned before, once the latent feature vectors are learned, they can be fed into some existing vector-based classifiers to detect semantic concepts for annotation. Instead of learning a latent space for MOs as a prestep, we develop an alternative algorithm in this section that directly learns the annotation model from training examples. Our method explores both the context and content information based on the latent structure between the correlated semantic concepts for annotation. Since it is a supervised algorithm, we will refer to it as Supervised Context-and-Content Multimedia Retrieval (S-C2MR) in this paper (in contrast to the U-C2MR algorithm in the last section). It is worth noting that even given a new MO without any associated context links, S-C2MR can still annotate it. In other words, S-C2MR can readily handle *out-of-sample problems* in the case of new MOs. This greatly extends the applicability of content and context-based multimedia annotation in many practical applications.

For a set of l semantic concepts, the goal of multimedia annotation is to predict the labels of these concepts on the MOs. A set of n MOs is used as the training data set to learn the annotation model on which the labels of l concepts are given. Let $y_{i,u}$ denote the training label of the uth concept for the ith MO, where $y_{i,u} = +1$ denotes the positive label and $y_{i,u} = -1$ denotes the negative label. Meanwhile, a set of d-dimensional raw feature vectors $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ (e.g., the visual features for images and audio-visual features for videos) are extracted from the training set. To predict the labels, l linear classifiers are to be learned, where $W_u \in \mathbb{R}^d$, $u = 1, 2, \ldots, l$, are the coefficient vectors for these linear classifiers. Then, $\tilde{y}_{i,u} = W_u^T \mathbf{f}_i$ is the prediction score for the uth concept on the ith MOs. Stacking W_u into a $d \times l$ matrix $W = [W_1, W_2, \ldots, W_l]$, $Y_i = W^T \mathbf{f}_i$ is the l-dimensional label vectors for all the l concepts on the ith MO.

In the learning phase, we learn the model parameter W. The aim is to ensure that the prediction scores given by W should match with the ground truth labels on the training set as much as possible. Let $m_{i,u} = y_{i,u}\tilde{y}_{i,u} = y_{i,u} W_u^T \mathbf{f}_i$; then it should be as large as possible by the maximum margin principle. We use the logistic loss function $h_\theta(x) = \frac{1}{\theta}\log(1 + \exp(-\theta x))$ to measure the margin with θ controlling its shape, and the margin can be maximized by minimizing the total logistic loss over all the training examples:

$$\mathcal{L}(W) = \sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(m_{i,u}\big) = \sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(y_{i,u} W_u^T \mathbf{f}_i\big). \quad (15)$$

To incorporate the information from the context links, when learning W we define an $n \times n$ symmetric matrix S, where each entry $S_{i,j}$ counts the number of COs that the ith and the jth MOs share. Actually, S can be computed as $S = AA^T$, and it summarizes the information in the context links. Similarly to the smoothness assumption made in the last section on the content links, it is also reasonable to assume that if two MOs share more COs, they ought to be semantically similar and the predicted label vectors on them should be as close as possible. Formally, this smoothness condition can be obtained by minimizing the following:

$$\begin{aligned}\Gamma(W) &= \frac{1}{2}\sum_{i,j=1}^{n} S_{i,j}\|Y_i - Y_j\|_2^2 \\ &= \frac{1}{2}\sum_{i,j=1}^{n} S_{i,j}\|W^T\mathbf{f}_i - W^T\mathbf{f}_j\|_2^2 \\ &= W^T F(J - S)F^T W \\ &= W^T F K F^T W.\end{aligned} \quad (16)$$

Here, $F = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n]$ is the $d \times n$ data matrix with the raw feature vectors as its columns, J is a diagonal matrix whose element is the sum of each corresponding row vector of S, and $K = J - S$ is the Laplacian matrix for the context links in contrast to the Laplacian matrix L for the content links in (9). The third equality in the above equation can be derived in a similar manner to (9).

Similarly to the tag vectors illustrated in Fig. 3, the target semantic concepts for annotation will not appear independently. The correlation between these concepts implies that a linear dependency structure exists among the predictions of these concepts on the MOs. In other words, these concepts form a low-dimensional latent space in which these concepts are (linearly) dependent on each other. Since each column vector of W corresponds to the prediction coefficients for the associated concept, the linear dependent structure among concept predictions implies that W ought to be of low rank. Combining (15) and (16) together with the above latent assumption of concept space, we can solve W by minimizing

$$\sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(y_{i,u} W_u^T \mathbf{f}_i\big) + \eta\,\mathrm{trace}\big(W^T F K F^T W\big) + \mu\|W\|_*, \quad (17)$$

where η and μ are the balancing parameters. Again, this optimization problem can be solved by the Proximal Gradient algorithm in a similar way as in the last section. In detail, let us denote

$$B(W) = \sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(y_{i,u} W_u^T \mathbf{f}_i\big) + \eta\,\mathrm{trace}\big(W^T F K F^T W\big), \quad (18)$$

then, given the fixed $W^{(\tau-1)}$ at iteration $\tau - 1$, (17) can be quadratically approximated by Taylor expanding $B(W)$ at $W^{(\tau-1)}$:

$$\begin{aligned}&P_\tau\big(W, W^{(\tau-1)}\big) \\ &= B\big(W^{(\tau-1)}\big) + \big\langle \nabla B\big(W^{(\tau-1)}\big), W - W^{(\tau-1)}\big\rangle \\ &\quad + \frac{\alpha}{2}\big\|W - W^{(\tau-1)}\big\|_F^2 + \mu\|W\|_* \\ &= \frac{\alpha}{2}\big\|W - G^{(\tau)}\big\|_F^2 + \mu\|W\|_* \\ &\quad + B\big(W^{(\tau-1)}\big) - \frac{1}{2\alpha}\big\|\nabla B\big(W^{(\tau-1)}\big)\big\|_F^2,\end{aligned} \quad (19)$$

where

$$G^{(\tau)} = W^{(\tau-1)} - \frac{1}{\alpha}\nabla B\big(W^{(\tau-1)}\big). \tag{20}$$

Here $\nabla B(W^{(\tau-1)})$ is an $l \times n$ matrix which is the gradient of $B(W)$ at $W^{(\tau-1)}$.

$B(W)$ consists of two terms, and we compute their gradients respectively. Note that the first term of logistic loss is always differentiable, so we have

$$\begin{aligned}&\frac{\partial}{\partial W_u}\left(\sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(y_{i,u}W_u^T\mathbf{f}_i\big)\right)\\ &= \sum_{i=1}^{n} y_{i,u}h'_\theta\big(y_{i,u}W_u^T\mathbf{f}_i\big)\mathbf{f}_i,\end{aligned} \tag{21}$$

where $h'_\theta(z) = \frac{-1}{1+e^{\theta z}}$ is the derivative of logistic loss function h at z. Denoting M as an $n \times l$ matrix with each entry $M_{i,u} = y_{i,u}h'_\theta(y_{i,u}W_u^T f_i)$, we have the gradient w.r.t. W:

$$\nabla\left(\sum_{i=1}^{n}\sum_{u=1}^{l} h_\theta\big(y_{i,u}W_u^T f_i\big)\right) = F \cdot M. \tag{22}$$

Therefore, the gradient of $B(W)$ is

$$\nabla B(W) = F \cdot M + 2\eta FKF^T W. \tag{23}$$

Then the new $W^{(\tau)}$ at iteration τ can be solved by

$$\begin{aligned}W^{(\tau)} &= \arg\min_{W} P_\tau\big(W, W^{(\tau-1)}\big)\\ &= \arg\min_{W} \frac{\alpha}{2}\big\|W - G^{(\tau)}\big\|_F^2 + \mu\|W\|_*,\end{aligned} \tag{24}$$

which has analytical solution according to Theorem 1. Note that as pointed out in [25], the convergence of the proximal gradient algorithm can be accelerated by making an initial estimate of α (here, we initialize α by $\sigma_{\max}(\nabla B(W^{(\tau-1)}))$ in each iteration) and multiplying it by a constant factor ρ ($= 0.7$ in our case) until $B(W^{(\tau)}) + \mu\|W^{(\tau)}\|_* \le P_\tau(W^{(\tau)}, W^{(\tau-1)})$.

Algorithm 2. Supervised Content-and-Context-based Multimedia Annotation

input Matrix S, balance parameters η and μ.

1 Initialize $W^{(0)} \leftarrow 0$ and $\tau \leftarrow 1$.

repeat

2 Compute the gradient of $B(W)$ at $W^{(\tau-1)}$ as (22).

3 Set $G^{(\tau)} = W^{(\tau-1)} - \frac{1}{\alpha}\nabla B(W^{(\tau-1)})$.

4 Set $W^{(\tau)} \leftarrow U\mathrm{diag}(\sigma - \frac{\mu}{\alpha})_+ V^T$, where $U\mathrm{diag}(\sigma)V^T$ is the SVD of $G^{(\tau)}$.

8 $\tau \leftarrow \tau + 1$.

until Convergence or maximum iteration number achieves.

In the inference phase, given the raw feature vector $\mathbf{f}$ of a new MO, its labels on l concepts can be predicted by $\tilde{y}(\mathbf{f}) = sign(W^T\mathbf{f})$.

Finally, we distinguish the proposed supervised content-and-context multimedia annotation algorithm from other latent models, including the one proposed in the last section. Previous latent methods, such as Latent Semantic Analysis [15], Probabilistic Latent Semantic Analysis [14], and Latent Dirichlet Allocation [6], are restricted to latent factor discovery. On the contrary, in this section, the goal of our approach is to directly model the semantic concepts from the content and context links while exploring their latent semantic correlations.

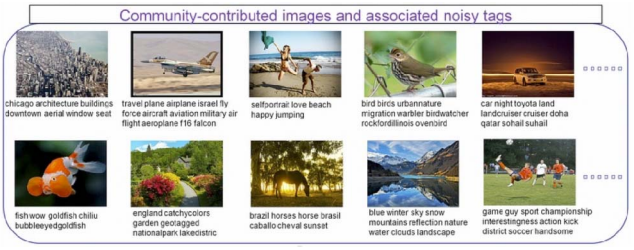


Fig. 4. Examples of Flickr images and associated community-contributed tags.

6 Experiments

To evaluate the proposed latent space method and its application in C2MR, we conduct experiments on a public multimedia data set with a large number of images as MOs and noisy user tags as COs. It is compared with the other paradigms of multimedia retrieval algorithms, such as CMR and CxMR. We evaluate these algorithms in the multimedia annotation problem, and their performances can be compared in quantity with the available labeling ground truth in the data set.

6.1 Data Set

Experiments are conducted on a publicly available Flickr data set.[2] It contains 55,615 images which are crawled from the photo sharing web site Flickr.com. The crawled images are linked to 1,000 user tags, which are annotated by users registered in Flickr. The context links between images and tags are quite sparse. In this data set, most of images only have fewer than 10 tags, and the average number of tags per image is 7.3. Fig. 4 illustrates some example of images and their associated user tags.

Beyond these images and user tags, 81 concepts are defined in the data set for image annotation. Note that these 81 concepts are different from the user tags, and their ground truth labels are manually collected by the data set developer. In contrast, tags are annotated by amateur users in Flickr, which contains much irrelevant noise information. The whole data set is partitioned into training set and test set for this annotation problem. The training set contains 27,807 images and the remaining 27,808 images are in the test set. In the training set, the training labels are given for all 81 concepts to learn prediction model. The annotation performances are then evaluated on test set.

Visual features extracted from the image corpus include the 64D color histogram and 73D edge direction histogram. These two kinds of features are concatenated together to form a 137D vector feature [10]. Features are normalized by subtracting each dimension of feature by its mean, and then dividing the resulting feature by three times of the standard variation of this dimension. After that, the feature vectors of all samples are normalized so that the square sum of all the elements in each feature vector is one [10].

2. http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm.

6.2 Performance Evaluation

The goal of multimedia retrieval is to retrieve a list which is relevant to the target concept. All the retrieved images are ranked according to their prediction scores in a descent order. The relevant images are expected to be ranked higher in the retrieved list. Therefore, to evaluate the ranking performance, we adopt Average Precision (AP) to measure the retrieval performance for each concept. Let R be the number of true positive images in the test set and R_j be the number of the relevant images in the top j images in the rank list. Let $I_j = 1$ if the jth image is relevant and 0 otherwise. Then AP is defined as

$$\frac{1}{R}\sum_{j}\frac{R_j}{j}I_j. \tag{25}$$

The AP corresponds to the area under a noninterpolated recall/precision curve and it favors highly ranked relevant images. In the experiments, AP is computed for each concept on the test set to measure the algorithm performance.

6.3 Comparison between Three Paradigms

First, we compare the proposed algorithm with the other three paradigms of multimedia retrieval algorithms. For the sake of fair comparison, the SVM model is trained based on the learned latent space and/or visual features.

1. CMR—Only visual features are used to model the 81 concepts. No user tags are used in this algorithm. In other words, we train SVM for each concept on visual features and the resulting SVM is used to predict the classification scores for retrieval. The Gaussian kernel is used in SVM for comparison.
2. CxMR—First, a latent space is learned solely from the context links between user tags and images based on PLSI. Then the SVM model is trained for each concept based on the obtained latent feature vectors to predict the scores. In the next section, we will compare with an advanced LSI variant—CLMF (i.e., combining Content and Link using Matrix Factorization [31]). We do not assume that user tags are available in the test set; thus, in this paradigm of latent methods, the user tags are predicated by their nearest neighbors in the training set.
3. C2MR—C2MR contains two different types—unsupervised C2MR and supervised C2MR.
    a. U-C2MR—Unsupervised C2MR. The algorithm in Section 3 is applied to model the latent space, which maps the MOs into a latent space from both content and context links. The parameters λ and γ in (10) are chosen from {0.2, 0.5, 1.0, 2.0} via a 5-folder cross-validation on training set in terms of the resulting AP. Then, SVM is used to train classification models from the learned latent space.
    b. S-C2MR—Supervised C2MR. The algorithm in Section 4 is developed for multimedia annotation. Different from U-C2MR, it directly learns classifier for the semantic concepts. The parameters η and μ in (17) are chosen from {0.2, 0.5, 1.0, 2.0} via a 5-folder cross-validation on training set, and the shape parameter θ for the logistic loss is empirically set to be 1.0.

Figs. 6 and 7 illustrate the performances on all the compared algorithms. From the results, we have the following observations.

Among CMR, CxMR, and C2MR, the proposed C2MR, both supervised and unsupervised versions, gain the best performances in terms of mean average precision (MAP) over all the 81 concepts. As for U-C2MR, it improves CMR by 246.8 percent and CxMR by 37.6 percent. Furthermore, S-C2MR improves CMR by 264.2 percent and CxMR by 44.5 percent. Meanwhile, of all 81 concepts, the proposed content and context multimedia retrieval methods (U-C2MR and S-C2MR) perform best on 58 concepts. On the remaining concepts, their performances only slightly deteriorate compared to the other algorithms.

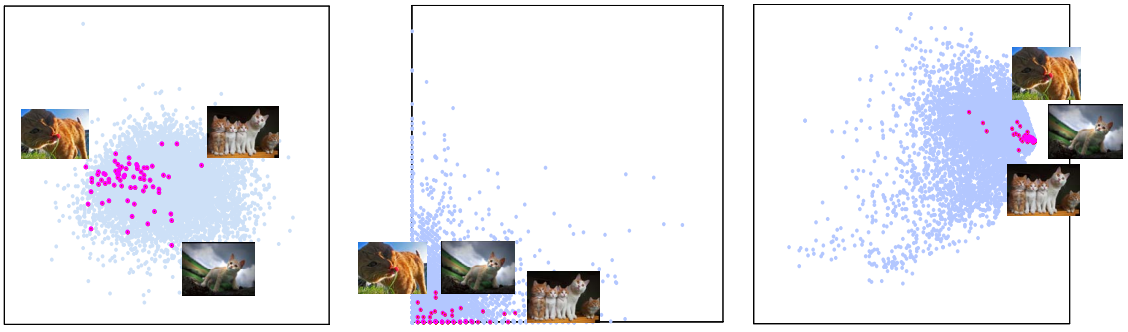
Comparing these three paradigms of multimedia retrieval methods, CMR performs worst since no semantic information in user tag is used. CxMR performs much better than CMR, although the tag link is sparse and noise. By regularizing the tag links by content links, C2MR significantly improves CxMR here. This is because, by mining the similarity information in content links between MOs, visually similar Flickr images can implicitly "share" the tag links between each other, which relieves the problem with sparse tag links. On the other hand, the noise in tags can also be somewhat reduced in a latent semantic space by embedding context links and visual geometric structure in content links simultaneously.

Finally, we illustrate how different algorithms map MOs into a 2D latent space in Fig. 5. It shows that the proposed method maps the MOs with the same class (i.e., "cat" in this example) close to each other so that they have consistent feature representation in the underlying latent space. It gives an intuitive interpretation of better performance of the proposed algorithm since it often becomes much easier to identify the region corresponding to a certain semantic class in the latent space where the objects of this class are mapped together.

6.4 Comparison with Related Algorithms

We also compare the proposed algorithm with the other closely related algorithms.

1. Fusion—We combine the 137D visual content features and the obtained context features in CxMR. The combined features are used to train SVM model for each concept. There are the following two different fusion strategy—early-fusion and late-fusion [23].
    a. Early-Fusion: The two kinds of features are concatenated and directly fed into SVM to train model for each concept.
    b. Late-Fusion: Two SVM models are learned from visual and PLSI features, respectively, to predict scores for each concept, and the final prediction scores are given by linearly combining them in a late fusion step.
2. SGSSL_dn—Sparse graph-based semi-supervised learning approach together with handling tag noises [24]. In this algorithm, a concept space is explicitly constructed from the context links. Moreover, a sparse graph is constructed by datum-wise



(a) CMR (b) CxMR (c) U-C2MR

Fig. 5. Illustration of different algorithms of mapping of MOs into a 2D latent space. The gray points correspond to the MOs in the corpus, and the red ones correspond to those of "cat" images. (a) CMR: Mapping MOs into the 2D space by applying principal component analysis to visual features of images. (b) CxMR: Mapping MOs by PLSI into the 2D space. (c) U-C2MR: Mapping MOs into the 2D space by the proposed latent method in Section 3.

one-vs-kNN reconstructions of all samples in which a training label refinement strategy is proposed to handle the noise in the user tags.

3. ML-DML—Multi-Label Distance Metric Learning [18]. This algorithm learns a semantic distance metric between visual features from user tags. Based on the learned distance, SVM is used to model each concept with a Gaussian kernel by exponentiating the obtained negative multilabel distance. Since it leverages user tags, it is compared with C2MR in the following.
4. CLMF—Combining Content and Link using Matrix Factorization [31]. This algorithm combines the content and link analysis using matrix factorization. It attempts to symmetrically factorize context matrix and asymmetrically factorize content matrix. In this model, some extra latent variables are used to model context topics.

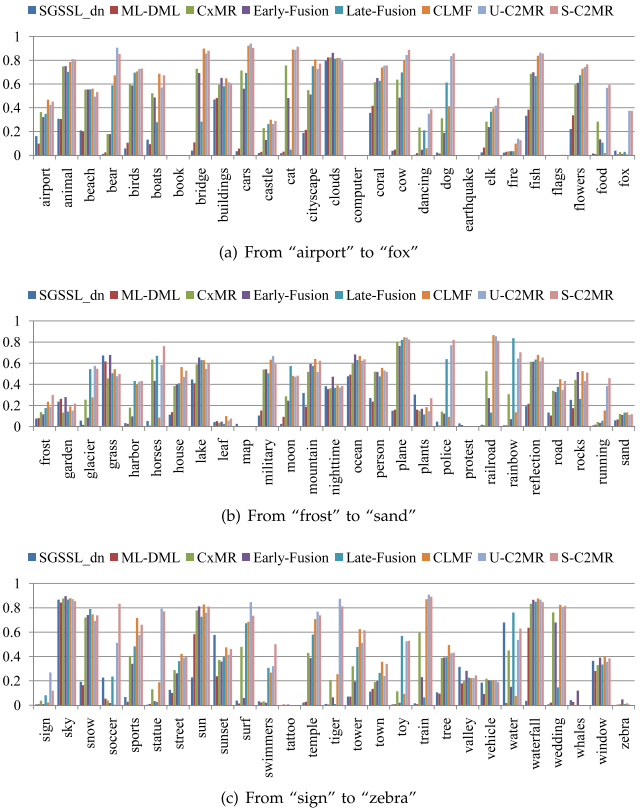


Fig. 6. Comparison of different algorithms over 81 concepts on the Flickr data set in terms of AP. The figure can be enlarged in the electronic version.

By comparison, in Fig. 7 C2MR shows it can more effectively model the two links than the other fusion methods in terms of MAP. U-C2MR improves Early-Fusion by 52.7 percent, Late-Fusion by 35.3 percent, SGSSL_dn by 225.8 percent, and ML-DML by 247.0 percent and CLMF by 15.6 percent. S-C2MR improves Early-Fusion by 60.3 percent, Late-Fusion by 42.1 percent, SGSSL_dn by 242.1 percent and ML-DML by 264.2 percent, and CLMF by 21.4 percent.

In Fusion methods, Late-Fusion outperforms Early-Fusion. It indicates that simply concatenating context and content feature vectors together into a higher dimensional vector cannot effectively utilize the context and content links. On the contrary, it is proven in the experiments that C2MR models a more informative latent space from the content and context links.

Finally, the comparison between ML-DML, SGSSL_dn, and C2MR also shows C2MR can better utilize the information in the links of MINs. Although SGSSL_dn attempts to handle the noisy tags in context links, it does not solve the problem with sparse context links. Moreover, the concept space in this approach constructed from user tags is usually far from perfect due to the semantic gap. This makes it difficult to further improve the performance of multimedia retrieval built on this concept space. Although ML-DML also utilizes user tags to learn a discriminant metric structure in visual feature space, it does not explore the geometric structure in either content links as U-C2MR or the context links as S-C2MR. Moreover, it does not look into the intrinsic latent space of either the tag vectors, as U-C2MR or the label vectors of semantic concepts, as S-C2MR.

Although CLMF attempts to incorporate content information into context analysis, it uses two matrices to separately factorize the context and content links. On the contrary, the proposed model learns a shared latent matrix H from content and context links simultaneously. Indeed, from the practical perspective, one extra matrix for either content or context links is unnecessary in multimedia retrieval, and it needs extra training samples to learn a satisfactory model. With more compact latent structure, the proposed algorithm is more compact than CLMF with shared latent matrix and thus has better

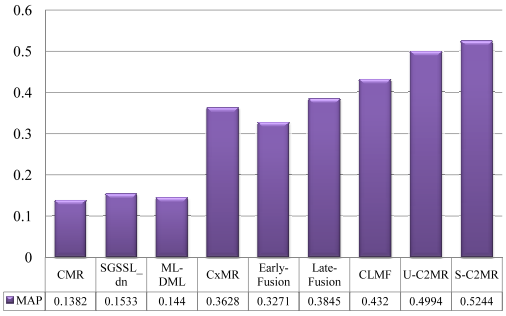


Fig. 7. Comparison of different algorithms over 81 concepts on Flickr data set in terms of MAP.

performance, as shown in experiment. Moreover, the proposed model can reduce the noise-induced uncertainty by low-rank prior, and the sparse context links are complemented by embedding MOs into their content linkage structure.

6.5 Comparison between U-C2MR and S-C2MR

Finally, we compare U-C2MR and S-C2MR. As shown in Fig. 7, S-C2MR performs slightly better than U-C2MR by 5 percent improvement. The reason is that S-C2MR aims at directly learning the semantic concepts for annotation in a unified framework and it utilizes extra discriminant information to learn the corresponding model for the target concepts.

6.6 Computing Time

Experiments are conducted on a platform with Intel Xeon CPU 2.80 GHz and 8 G physical memory. Table 1 illustrates the computing time of different algorithms compared above. Since CMR is conducted directly on low-level feature space without modeling the latent space, its computing time is not listed. By comparison, both U-C2MR and S-C2MR are more computationally efficient than CxMR and SGSSL_dn, and have a similar computation load as CLMF. On the other hand, although U-C2MR and S-C2MR perform more slowly than ML-DML, they improve the performance of ML-DML significantly as shown in the above.

7 Conclusion

In this paper, we propose an algorithm which discovers the latent semantic space from both context and content links in MINs. The algorithms solve the problem with sparse context links by enriching the MINs with content links, and MOs are embedded into a geometric structure underlying their content information. We extend the traditional LSI algorithm by low-rank approximation in which the information from the content links is seamlessly incorporated. The learned latent semantic space can be applied for many applications, such as multimedia annotation and retrieval. Specifically, we develop a context-and-content-based multimedia annotation algorithm which can learn the concept models from the context links and content links simultaneously based on the intrinsic low-rank structure in the latent concept space. For evaluation, we compare the proposed algorithm with other multimedia retrieval paradigms with either content or context links on a real-world Flickr data set. Other related algorithms in MINs are compared as well. The results show that the proposed algorithm is quite effective to integrate the content and context links for semantic retrieval over all 81 concepts from Flickr data set.

TABLE 1
Comparison of Computing Time (in Seconds) by Latent Methods and the Other Related Methods

	Algorithms	Computing Time
Latent Methods	CMR	N/A
	CxMR	8152.50 secs
	CLMF	3045.31 secs
	U-C2MR	2347.78 secs
	S-C2MR	3749.48 secs
Other Methods	SGSSL_dn	22680.0 secs
	ML-DML	349.57 secs

Acknowledgments

Research was sponsored by the US Army Research Laboratory Cooperative Agreement Number W911NF-09-2-0053 and the US National Science Foundation under Grant IIS-1144111. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The work was also supported in part to Guo-Jun Qi by an IBM PhD fellowship award, as well as in part to Dr. Qi Tian by US National Science Foundation grant IIS 1052851, Faculty Research Awards by Google, FXPAL, and NEC Laboratories of America, respectively. This work was originally submitted to International ACM Conference on Multimedia 2010.

References

[1] J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R.C. Jain, and C.-F. Shu, "Virage Image Search Engine," *Proc. SPIE*, vol. 2670, no. 76, 1996.
[2] R. Bekkerman and J. Jeon, "Multi-Modal Clustering for Multimedia Collections," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.
[3] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems*, vol. 1, no. 14, pp. 585-591, 2001.
[4] A.B. Benitez, J.R. Smith, and S.-F. Chang, "Medianet: A Multimedia Information Network for Knowledge Representation," *Proc. SPIE Conf. Series*, 2000.
[5] T.L. Berg, A.C. Berg, and J. Shih, "Automatic Attribute Discovery and Characterization from Noisy Web Images," *Proc. 11th European Conf. Computer Vision*, 2010.
[6] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, pp. 993-1022, Jan. 2003.
[7] A. Bosch, A. Zisserman, and X. Munoz, "Scene Classification via pLSA," *Proc. European Conf. Computer Vision*, 2006.
[8] E. Candés and Y. Plan, "Matrix Completion with Noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925-936, June 2010.
[9] E.J. Candés and P. Randall, "Highly Robust Error Correction by Convex Programming," *IEEE Trans. Information Theory*, vol. 54, no. 7, pp. 2829-2840, July 2006.
[10] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-Wide: A Real-World Web Image Database from National University of Singapore," *Proc. ACM Int'l Conf. Image and Video Retrieval*, 2009.

[11] F.R.K. Chung, "Spectral Graph Theory," *Proc. Regional Conf. Series in Math.,* 1997.
[12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The Qbic System," *Intelligent Multimedia Information Retrieval,* MIT Press, pp. 7-22, 1997.
[13] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal Semi-Supervised Learning for Image Classification," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.
[14] T. Hofmann, "Probabilistic Latent Semantic Analysis," *Proc. Uncertainty in Artificial Intelligence,* 1999.
[15] T.K. Landauer, P.W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, pp. 259-284, 1998.
[16] P.P. Martin Labský and M. Vacura, "Web Image Classification for Information Extraction," *Proc. Int'l Workshop Representation and Analysis of Web Space,* 2005.
[17] F. Monay and D. Gatica-Perez, "pLSA-Based Image Auto-Annotation: Constraining the Latent Space," *Proc. 12th ACM Ann. Int'l Conf. Multimedia,* pp. 348-351, 2004.
[18] G.-J. Qi, X.-S. Hua, and H.-J. Zhang, "Learning Semantic Distance from Community-Tagged Media Collection," *Proc. 17th ACM Int'l Conf. Multimedia,* 2009.
[19] A. Rakotomamonjy, F.R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Machine Learning Research,* vol. 9, pp. 2491-2521, Nov. 2008.
[20] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Recognition.* Cambridge Univ. Press, 2004.
[21] S. Sizov, "Geofolk: Latent Spatial Semantics in Web 2.0 Social Media," *Proc. Third ACM Int'l Conf. Web Search and Data Mining,* 2010.
[22] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1349-1380, Dec. 2000.
[23] C.G.M. Snoek, M. Worring, J.C. van Gemert, J.M. Geusebroek, and A.W.M. Smeulders, "The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia," *Proc. 14th Ann. ACM Int'l Conf. Multimedia,* 2006.
[24] J. Tang, S. Yan, R. Hong, G.-J. Qi, and T.-S. Chua, "Inferring Semantic Concepts from Community-Contributed Images and Noisy Tags," *Proc. 17th ACM Int'l Conf. Multimedia,* 2009.
[25] K.C. Toh and S. Yun, "An Accelerated Proximal Gradient Algorithm for Nuclear Norm Regularized Least Squares Problems," preprint on Optimization Online, Apr. 2009.
[26] S. Wang, Q. Huang, S. Jiang, L. Qin, and Q. Tian, "Visual Contextrank for Web Image Re-Ranking," *Proc. First ACM Workshop Large-Scale Multimedia Retrieval and Mining,* pp. 121-128, Oct. 2009.
[27] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization," *Proc. Neural Information Processing Systems,* Dec. 2009.
[28] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell, "Distance Metric Learning with Application to Clustering with Side-Information," *Proc. Advanced Neutral Information Processing System,* 2003.
[29] Q. Yang, Y. Chen, G.R. Xue, W. Dai, and Y. Yu, "Heterogeneous Transfer Learning for Image Clustering via the Social Web," *Proc. Joint Conf. 47th Ann. Meeting of the ACL and the Fourth IJCNLP of the AFNLP,* pp. 1-9, Aug. 2009.
[30] J. Yu, X. Jin, J. Han, and J. Luo, "Social Group Suggestion from User Image Collections," *Proc. 19th Int'l Conf. World Wide Web,* 2010.
[31] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining Content and Link for Classification Using Matrix Factorization," *Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* 2007.

Guo-Jun Qi received the BS degree from the University of Science and Technology of China in Automation, Hefei, Anhui, China, in 2005. His research interests include pattern recognition, machine learning, computer vision, and multimedia. In 2011, he was a recipient of the IBM PhD fellowship award. He was the winner of the best paper award at the 15th ACM International Conference on Multimedia, Augsburg, Germany, 2007. He has been with the Beckman Institute and the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign since 2009. He also has served as a program committee member and reviewer for many academic conferences and journals in the fields of computer vision, pattern recognition, machine learning, and multimedia.

Charu Aggarwal received the BS degree from IIT Kanpur in 1993 and the PhD degree from the Massachusetts Institute of Technology in 1996. Since then he has worked in the field of performance analysis, databases, and data mining. He is a research scientist at the IBM T.J. Watson Research Center in Yorktown Heights, New York. He has published more than 155 papers in refereed conferences and journals and has been granted more than 50 patents. Because of the commercial value of the above-mentioned patents, he has received several invention achievement awards and has been designated a Master Inventor at IBM three times. He is a recipient of an IBM Corporate Award (2003) for his work on bioterrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, and a recipient of an IBM Research Division Award (2008) for his scientific contributions to data stream research. He has served on the program committees of most major database/data mining conferences, and served as a program vice-chair for the SIAM Conference on Data Mining, 2007, the IEEE ICDM Conference, 2007, the WWW Conference 2009, and the IEEE ICDM Conference, 2009. He served as an associate editor of the *IEEE Transactions on Knowledge and Data Engineering* from 2004 to 2008. He is an associate editor of the *ACM Transactions on Knowledge from Data Discovery*, an action editor of *Data Mining and Knowledge Discovery*, an associate editor of the *ACM SIGKDD Explorations*, and an associate editor of the *Knowledge and Information Systems Journal*. He is a fellow of the IEEE for "contributions to knowledge discovery and data mining techniques" and a life-member of the ACM.

Qi Tian received the BE degree in electronic engineering from Tsinghua University, China, in 1992, the MS degree in electrical and computer engineering from Drexel University in 1996, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2002. He is currently an associate professor in the Department of Computer Science at the University of Texas at San Antonio (UTSA). His research interests include multimedia information retrieval and computer vision. He has published more than 120 refereed journal and conference papers. His research projects were funded by the NSF, ARO, DHS, SALSI, CIAS, and UTSA and he received faculty research awards from Google, NEC Laboratories of America, FXPAL, Akiira Media Systems, and HP Labs. He took a one-year faculty leave at Microsoft Research Asia (MSRA) during 2008-2009. He was the coauthor of a Best Student Paper in ICASSP 2006, and coauthor of a Best Paper candidate in PCM 2007. He received the 2010 ACM Service Award. He has been serving as program chair, organization committee member, and TPC for numerous IEEE and ACM conferences including ACM Multimedia, SIGIR, ICCV, ICME, etc. He has been a guest editor for the *IEEE Transactions on Multimedia*, *Journal of Computer Vision and Image Understanding*, *Pattern Recognition Letter*, *EURASIP Journal on Advances in Signal Processing, Journal of Visual Communication and Image Representation*, and is on the editorial board of the *IEEE Transactions on Circuit and Systems for Video Technology* (*TCSVT*), *Journal of Multimedia* (JMM), and *Journal of Machine Vision and Applications* (MVA). He is a senior member of the IEEE.

Heng Ji received the PhD degree in computer science from New York University (NYU) in 2007. She is an assistant professor and doctoral faculty member in computer science at Queens College and the Graduate Center of City University of New York (CUNY), and the director of the BLENDER Lab. Her research interests focus on information extraction and knowledge discovery. She has published several book chapters and many conference and journal papers. In 2006, she was awarded the Sandra Bleistein Prize from the Courant Institute of Mathematical Sciences of NYU for the most notable achievement by a woman in math and computer science. In 2009, she was the recipient of Google Research Award. In 2010, she received a five-year Faculty Early Career Development (CAREER) Award from the US National Science Foundation (NSF). In 2011, she received the CUNY Chancellor's Salute to Scholar award. Since 2008 she has also received several research awards from the US Army Research Lab, NSF, and Defense Advanced Research Projects Agency. She has been coorganizing the NIST TAC Knowledge Base Population task in 2010 and 2011. She is a member of the IEEE.

Thomas S. Huang received the ScD degree from MIT in 1963. He is the William L. Everitt Distinguished Professor in the University of Illinois at Urbana-Champaign Department of Electrical and Computer Engineering and the Coordinated Science Lab (CSL), and a full-time faculty member in the Beckman Institute Image Formation and Processing and Artificial Intelligence groups. His research interests include computer vision, image compression and enhancement, pattern recognition, and multimodal signal processing. He is a life fellow of the IEEE.