A Probabilistic Framework for Multimodal Retrieval using Integrative Indian Buffet Process

Authored by:

Larry S. Davis
Bahadir Ozdemir

Abstract

We propose a multimodal retrieval procedure based on latent feature models. The procedure consists of a nonparametric Bayesian framework for learning underlying semantically meaningful abstract features in a multimodal dataset, a probabilistic retrieval model that allows cross-modal queries and an extension model for relevance feedback. Experiments on two multimodal datasets, PASCAL-Sentence and SUN-Attribute, demonstrate the effectiveness of the proposed retrieval procedure in comparison to the state-of-the-art algorithms for learning binary codes.

1   Paper Body

As the number of digital images which are available online is constantly increasing due to rapid advances in digital camera technology, image processing tools and photo sharing platforms, similaritypreserving binary codes have received significant attention for image search and retrieval in largescale image collections [1, 2]. Encoding high-dimensional descriptors into compact binary strings has become a very popular representation for images because of their high efficiency in query processing and storage capacity [3, 4, 5, 6]. The most widely adapted strategy for similarity-preserving binary codes is to find a projection of data points from the original feature space to Hamming space. A broad range of hashing techniques can be categorized as data independent and dependent schemes. Locality sensitive hashing [3] is one of the most widely known data-independent hashing techniques. This technique has been extended to various hashing functions with kernels [4, 5]. Notable data-dependent hashing techniques include spectral hashing [1], iterative quantization [6] and spherical hashing [7]. Despite the increasing amount of multimodal data, especially in multimedia domains e.g. images with tags, most existing hashing techniques, unfortunately, focus on unimodal data. Hence, they inevitably suffer from the semantic gap, which is defined in [8] as the lack of coincidence between low level visual features and high level semantic interpretation of an image. On the other

hand, joint analysis of multimodal data offers improved search and cross-view retrieval capabilities e.g. text-to-image queries by bridging the semantic gap. However, it also poses challenges associated with handling cross-view similarity. Most recent studies have concentrated on multimodal hashing. Bronstein et al. proposed crossmodality similarity learning via a boosting procedure [9]. Kumar and Udupa presented a cross-view similarity search [10] by generalizing spectral hashing [1] for multi-view data objects. Zhen and Yeung described two recent methods: Co-regularized hashing [11] based on a boosted co-regularization framework and a probabilistic generative approach called multimodal latent binary embedding [12] based on binary latent factors. Nitish and Salakhutdinov proposed a deep Boltzmann machine for multimodal data [13]. Recently, Rastegari et al. proposed a predictable dual-view hashing [14] that aims to minimize the Hamming distance between binary codes obtained from two different views by utilizing multiple SVMs. Most of the multimodal hashing techniques are computationally ex1

pensive, especially when dealing with large-scale data. High computational and storage complexity restricts their scalability. Although many hashing approaches rely on supervised information like semantic class labels, class memberships are not available for many image datasets. In addition, some supervised approaches cannot be generalized to unseen classes that are not used during training [15] even though new classes emerge in the process of adding new images to online image databases. Besides, every user?s need is different and time varying [16]. Therefore, user judgments indicating the relevance of an image retrieved for a query are utilized to achieve better retrieval performance in the revised ranking of images [17]. Development of an efficient retrieval system that embeds information from multiple domains into short binary codes and takes relevance feedback into account is quite challenging. In this paper, we propose a multimodal retrieval method based on latent features. A probabilistic approach is employed for learning binary codes, and also for modeling relevance and user preferences in image retrieval. Our model is built on the assumption that each image can be explained by a set of semantically meaningful abstract features which have both visual and textual components. For example, if an image in the dataset contains a side view of a car, the words ?car?, ?automobile? or ?vehicle? will probably appear in the description; also an object detector trained for vehicles will detect the car in the image. Therefore, each image can be represented as a binary vector, with entries indicating the presence or absence of each abstract feature. Our contributions can be summarized in three aspects: 1. We propose a Bayesian nonparametric framework based on the Indian Buffet Process (IBP) [18] for integrating multimodal data in a latent space. Since the IBP is a nonparametric prior in an infinite latent feature model, the proposed method offers a flexible way to determine the number of underlying abstract features in a dataset. 2. We develop a retrieval system that can respond to cross-modal queries by introducing new random variables indicating relevance to a query. We present a Markov chain Monte Carlo (MCMC) algorithm for inference of the relevance from data. 3. We formulate relevance feedback as pseudo-images to alter the distribution of images in the latent space so that

the ranking of images for a query is influenced by user preferences. The rest of the paper is organized as follows: Section 2 describes the proposed integrative procedure for learning binary codes, retrieval model and processing relevance feedback in detail. Performance evaluation and comparison to state-of-the-art methods are presented in Section 3, and Section 4 provides conclusions.

2

Our Approach

In our data model, each image has both textual and visual components. To facilitate the discussion, we assume that the dataset is composed of two full matrices; our approach can easily handle images with only one component and it can be generalized to more than two modalities as well. We denote the data in the textual and visual space by X? and Xv , respectively. X? is an N ? D? matrix whose rows corresponds to images in either space where ? is a placeholder used for either v or ? . The values in each column of X? are centered by subtracting the sample mean of that column. The dimensionality of the textual space D? and the dimensionality of the visual space Dv can be different. We use X to represent the set {X? , Xv }. 2.1

Integrative Latent Feature Model

We focus on how textual and visual values of an image are generated by a linear-Gaussian model and its extension for retrieval systems. Given a multi-modal image dataset, the textual and visual data matrices, X? and Xv , can be approximated by ZA? and ZAv , respectively. Z is an N ? K binary matrix where Znk equals to one if abstract feature k is present in image n and zero otherwise. A? is a K ? D? matrix where the textual and visual values for abstract feature k are stored in row k of A? and Av , respectively (See Figure 1 for an illustration). The set {A? , Av } is denoted by A. 2

Our initial goal is to learn abstract features present in the dataset. Given X , we wish to compute the posterior distribution of Z and A using Bayes? rule

$$p(Z, A | X) \propto p(X? | Z, A?)p(A?)p(Xv | Z, Av)p(Av)p(Z)$$

(1)

where Z, A? and Av are assumed to be a priori independent. In our model, the vectors for textual and visual properties of an image are generated from Gaussian distributions with covariance matrix $(?x?)2 I$ and expectation $E[X?]$ equal to ZA? . Similarly, a prior on A? is defined to be Gaussian with zero mean vector and covariance matrix $(?a?)2 I$. Since we do not know the exact number of abstract features present in the dataset, we employ the Indian Buffet Process (IBP) to generate Z, which provides a flexible prior that allows K to be determined at inference time (See [18] for details). The graphical model of our integrative approach is shown in Figure 2. Unobserved

Observed Visual features for image

visual textual Abstract features for image Textual features for image

Figure 1: The latent abstract feature model proposes that visual data Xv is a product of Z and Av with some noise; and similarly the textual data X? is a product of Z and A? with some noise.

Figure 2: Graphical model for the integrative IBP approach where circles indicate random variables, shaded circles denote observed values, and the blue

square boxes are hyperparameters. The exchangeability property of the IBP leads directly to a Gibbs sampler which takes image n as the last customer to have entered the buffet. Then, we can sample Znk for all initialized features k via p(Znk = 1—Z?nk , X) ? p(Znk = 1—Z?nk)p(X —Z). (2) where Z?nk denotes entries of Z other than Znk . In the finite latent feature model (where K is fixed), the conditional distribution for any Znk is given by p(Znk = 1—Z?nk) =

? m?n,k + K ? N+K

(3)

where m?n,k is the number of images possessing abstract feature k apart from image n. In the m infinite case like the IBP, we obtain p(Znk = 1—Z?nk) = ?n,k for any k such N that m?n,k ¿ 0. ? We also need to draw new features associated with image n from Poisson N , and the likelihood term is now conditioned on Z with new additional columns set to one for image n. 3

For the linear-Gaussian model, the collapsed likelihood function p(X —Z) = p(X? —Z)p(Xv —Z) can be computed using

Z exp ? 2(?1?)2 tr X? T (I ? ZMZT)X? ? ? ? ? ? ? x p(X —Z) = p(X —Z, A)p(A) dA = (4) ?D ? N D? (2?) 2 (?x?)(N ?K)D? (?a?)KD? —M— 2 (? ?)2 ?1 and tr(?) is the trace of a matrix [18]. To reduce the computational where M = ZT Z + (?x?)2 I a complexity, Doshi-Velez and Ghahramani proposed an accelerated sampling in [19] by maintaining the posterior distribution of A? conditioned on partial X? and Z. We use this approach to learn binary codes, i.e. the feature-assignment matrix Z, for multimodal data. Unlike the hashing methods that learn optimal hyperplanes from training data [6, 7, 14], we only sample Z without specifying the length of binary codes in this process. Therefore, the binary codes can be updated efficiently if new images are added in a long run of the retrieval system. 2.2

Retrieval Model

We extend the integrative IBP model for image retrieval. Given a query, we need to sort the images in the dataset with respect to their relevance to the query. A query can be comprised of textual and visual data, or either component can be absent. Let q? be a D? -dimensional vector for the textual values and qv be a Dv -dimensional vector for the visual values of the query. We can write Q = {q? , qv }. As for the images in X , we consider a query to be generated by the same model described in the previous section with the exception of the prior on abstract features. In the retrieval part, we consider Z as a known quantity and we fix the number abstract features to K. Therefore, the feature-assignments for the dataset are not affected by queries. In addition, queries are explained by known abstract features only. We extend the Indian restaurant metaphor to construct the retrieval model. A query corresponds to the (N + 1)th customer to enter the buffet. The previous customers are divided into two classes as friends and non-friends based on their relevance to the new customer. The new customer now samples from at most K dishes in proportion to their popularity among friends and also their unpopularity among non-friends. Consequently, the dishes sampled by the new customer are expected to be similar to those of friends and dissimilar to those of non-friends. Let r be an N -dimensional vector

4

where rn equals to one if customer n is a friend of the new customer and zero otherwise. For this finitely long buffet, the sampling probability of dish k by the new customer can be written PN m0k +?/K as N +1+?/K where m0k = n=1 (Znk)rn (1 ? Znk)1?rn , that is the total number of friends who tried dish k and non-friends who did not sample dish k. Let z0 be a K-dimensional vector where zk0 records if the new customer (query) sampled dish k. We place a prior over rn as Bernoulli(?). Then, we can sample zk0 from p(zk0 = 1—z0?k , Q, Z, X) ? p(zk0 = 1—Z)p(Q—z0 , Z, X). The probability as below:

p(zk0

(5)

= 1—Z) can be computed efficiently for k = 1, . . . , K by marginalizing over r

p(zk0 = 1—Z) =

X

p(zk0 = 1—r, Z)p(r) =

r?{0,1}N

?mk + (1 ? ?)(N ? mk) + ? N +1+ K

? K

.

(6)

The collapsed likelihood of the query, p(Q—z0 , Z, X), is given by the product of textual and visual likelihood values, p(q? —z0 , Z, X?)p(qv —z0 , Z, Xv). If either textual or visual component is missing, we can simply integrate out the missing one by omitting the corresponding term from the equation. The likelihood of each part can be calculated as follows: Z p(q? —z0 , Z, X?) = p(q? —z0 , A?)p(A? —Z, X?) dA? = N (q? ; ??q , ??q). (7) where the mean and covariance matrix of the normal distribution are given by ??q = z0 MZT X? and ??q = (?x?)2 (z0 Mz0T + I), akin to the update equation in [19] (Refer to (4) for M). Finally, we use the conditional expectation of r to rank images in the dataset with respect to their relevance to the given query. Calculating the expectation E[r—Q, Z, X] is computationally expensive; 4

however, it can be empirically estimated using the Monte Carlo method as follows: 0(i) I I K X ? X Y p zk —rn = 1, Z ? n —Q, Z, X] = 1 E[r p(rn = 1—z0(i) , Z) = 0(i) I i=1 I i=1 p zk —Z k=1

(8)

where z0(i) represents i.i.d. samples from (5) for i = 1, . . . , I. The last equation required for computing (8) is p(zk0 = 1—rn = 1, Z) =

Znk + ?m?n,k + (1 ? ?)(N ? 1 ? m?n,k) + ? N +1+ K

? K

.

(9)

The retrieval system returns a set of top ranked images to the user. Note that we compute the expectation of relevance vector instead of sampling directly since binary values indicating the relevance are less stable and they hinder the ranking of images. 2.3

Relevance Feedback Model

In our data model, user preferences can be described over abstract features. For instance, if abstract feature k is present in the most of positive samples i.e. images judged as relevant by the user and it is absent in the irrelevant ones, then we can say that the user is more interested in the semantic subspace represented by abstract feature k. In the revised query, the images having abstract feature k are expected to be ranked in higher positions in comparison to the initial query. We can achieve this desirable property from query-specific alterations to the sampling probability in (5) for the corresponding abstract features. Our approach is to add pseudo-images to the feature-assignment matrix Z before the computations of the revised query. For the Indian restaurant analogy, pseudoimages correspond to some additional friends of the new customer (query), who do not really exist in the restaurant. The distribution of dishes sampled by those imaginary customers reflects user relevance feedback. Thus, the updated expectation of the relevance vector has a bias towards user preferences. Let Zu be an Nu ?K feature-assignment matrix for pseudo-images only; then the number of pseudoimages, Nu , determines the influence of relevance feedback. Therefore, we set an upper limit on Nu as the number of real images, N , by placing a prior distribution as Nu ? Binomial(?, N) where ? is a parameter that controls the weight of feedback. Let mu,k be the number of pseudo-images containing abstract feature k; then this number has an upper bound Nu by definition. For abstract feature k, a prior distribution conditioned on Nu can be defined as mu,k —Nu ? Binomial(?k , Nu) where ?k is a parameter that can be tuned by relevance judgments. Let z00 be a K-dimensional feature-assignment vector for the revised query; then we can sample each zk00 via p(zk00 = 1—z00?k , Q, Z, X) ? p(zk00 = 1—Z)p(Q—z00 , Z, X)

(10)

where the computation of the collapsed likelihood is already shown in (7). Note that we do not actually generate all entries of Zu but only the sum of its columns mu and number of rows Nu for computing the sampling probability. We can write the first term as: p(zk00 = 1—Z) =

N X Nu =0

p(Nu)

Nu X

p(mu,k —Nu)

mu,k =0

X

p(zk00 = 1—r, Zu , Z)p(r)

r?{0,1}N

N X ?mk + (1 ? ?)(N ? mk) + N j = ? (1 ? ?)N ?j ? N +1+ K +j j j=0

(11) ? K

+ ?k j

Unfortunately, this expression has no compact analytic form; however, it can be efficiently computed numerically by contemporary scientific computing software even for large values of N . In this equation, one can alternatively fix rn to 1 if the user marks observation n as relevant or 0 if it is indicated to be

irrelevant. Finally, the expectation of r is updated using (8) with new i.i.d. samples z00(i) from (10) and the system constructs the revised set of images. 5

3

Experiments

The experiments were performed in two phases. We first compared the performance of our method in category retrieval with several state-of-the-art hashing techniques. Next, we evaluated the improvement in the performance of our method with relevance feedback. We used the same multimodal datasets as [14], namely PASCAL-Sentence 2008 dataset [20] and the SUN-Attribute dataset [21]. In the quantitative analysis, we used the mean of the interpolated precision at standard recall levels for comparing the retrieval performance. In the qualitative analysis, we present the images retrieved by our proposed method for a set of text-to-image and image-to-image queries. All experiments were performed in the Matlab environment1 . 3.1

Datasets

The PASCAL-Sentence 2008 dataset is formed from the PASCAL 2008 images by randomly selecting 50 images belonging to each of the 20 categories. In experiments, we used the precomputed visual and textual features provided by Farhadi et al. [20]. Amazon Mechanical Turk workers annotate five sentences for each of the 1000 images. Each image is labelled by a triplet of ¡object, action, scene¿ representing the semantics of the image from these sentences. For each image, the semantic similarity between each word in its triplet and all words in a dictionary constructed from the entire dataset is computed by the Lin similarity measure [22] using the WordNet hierarchy. The textual features of an image are the sum of all similarity vectors for the words in its triplet. Visual features are built from various object detectors, image classifiers and scene classifiers. These features contain the coordinates and confidence values that object detectors fire and the responses of image and scene classifiers trained on low-level image descriptors. The SUN-Attribute dataset [21], a large-scale dataset of attribute-labeled scenes, is built on top of the existing SUN categorical dataset [23]. The dataset contains 102 attribute labels annotated by 3 Amazon Mechanical Turk workers for each of the 14,340 images from 717 categories. Each category has 20 annotated images. The precomputed visual features [21, 23] include gist, 2?2 histogram of oriented gradient, self-similarity measure, and geometric context color histograms. The attribute features is computed by averaging the binary labels from multiple annotators where each image is annotated with attributes from five types: materials, surface properties, functions or affordances, spatial envelope attributes and object presence. 3.2

Experimental Setup

Firstly, all features were centered to zero and normalized to unit length; also duplicate features were removed from the data. We reduced the dimensionality of visual features in the SUN dataset from 19,080 to 1,000 by random feature selection, which is preferable to PCA for preserving the variance among visual features. The Gibbs sampler was initialized with a randomly sampled feature assignment matrix Z from a IBP prior. We set ? = 1 in all experiments to keep binary codes short. The other hyperparameters ?a? and ?x? were deter-

7

mined by adding Metropolis steps to the MCMC algorithm in order to prevent one modality from dominating the inference process. In the retrieval part, the relevance probability ? was set to 0.5 so that all abstract features have equal prior probability from (6). Feature assignments of a query were initialized with all zero bits. For relevance feedback analysis, we set ? = 1 (equal significance for the data and feedback) and we decide each ?k as follows: PI 0(i) Let z?k0 = I1 i=1 zk where each z0(i) is drawn from (5) for a given query; and z?k0 = P T 1 rt 1?rt where t represents the index of each image judged by the user and t=1 (Ztk) (1 ? Ztk) T T is the size of relevance feedback. The difference between these two quantities, ?k = z?k0 ? z?k0 , controls ?k which is defined by a logistic function as 1 ?k = (12) ?(c? k +?0,k) 1+e p(z 0 =1—Z)

where c is a constant and ?0,k = ln p(zk0 =0—Z) (refer to (6) for p(zk0 —Z)). We set c = 5 in our k experiments. Note that ?k = p(zk0 = 1—Z) when z?k0 is equal to z?k0 . 1

Our code is available at http://www.cs.umd.edu/?ozdemir/iibp

6

3.3

Experimental Results

We compared our method, called integrative IBP (iIBP), with several hashing methods including locality sensitive hashing (LSH) [3], spectral hashing (SH) [1], spherical hashing (SpH) [7], iterative quantization (ITQ) [6], multimodal deep Boltzmann machine (mDBM) [13] and predictable dualview hashing (PDH) [14]. We divided each dataset into two equal sized train and test segments. The train segment was first used for learning the feature assignment matrix Z by iIBP. Then, the other binary code methods were trained with the same code length K. We used supervised ITQ coupled with CCA [24] and took the dual-view approach [14] to construct basis vectors in a common subspace. However, LSH, SH and SpH were applied on single-view data since they do not support cross-view queries. All images in the test segment were used as both image and text queries. Given a query, images in the train set were ranked by iIBP with respect to (8). For all other methods, we use Hamming distance between binary codes in the nearest-neighbor search. Mean precision curves are presented in Figure 3 for both datasets. Unlike the experiments in [14] performed in a supervised manner, the performance on the SUN-Attribute dataset is very low due to the small number of positive samples compared to the number of categories (Figure 3b). There are only 10 relevant images among 7,170 training images. Therefore, we also used Euclidean neighbor ground truth labels computed from visual data as in [6] (Figure 3c). As seen in the figure, our method (iIBP) outperforms all other methods. Although unimodal hashing methods perform well on text queries, they suffer badly on image queries because the semantic similarity to the query does not necessarily require visual similarity (Figures 3-4 in the supplementary material). By the joint analysis of visual and textual spaces, our approach improves the performance for image queries by bridging the semantic gap [8]. iIBP

mDBM

0.7 0.6

8

ITQ
SpH
SH 0.5
0.08
0.45
Mean Precision
Mean Precision
Mean Precision
0.4
0.06
0.4
0.35
0.05 0.04
0.3
0.02
0.1 0
0.2
0.4
0.6
Recall
0.8
(a) PASCAL-Sentence Dataset (K = 23)
1
0.3
0.25
0.03
0.2
LSH
0.09
0.07
0.5
0
PDH
0.2
0.15
0.01
0.1
0
0.05
0
0.2
0.4
0.6
Recall
0.8
(b) SUN Dataset ? Class label ground truth (K = 45)

1
0
0.2
0.4
0.6
Recall
0.8
1
(c) SUN Dataset ? Euclidean ground truth (K = 45)

Figure 3: The result of category retrieval for all query types (image-to-image and text-to-image queries). Our method (iIBP) is compared with the-state-of-the-art methods. For qualitative analysis, Figure 4a shows the top-5 retrieved images from the PASCAL-Sentence 2008 dataset for image queries. Thanks to the integrative approach, the retrieved images share remarkable semantic similarity with the query images. Similarly, most of the retrieved images for the text-to-image queries in Figure 4b comprise the semantic structure in the query sentences. In the second phase of analyses, we utilized the rankings in the first phase to decide relevance feedback parameters independently for each query. We picked the top two relevant images as positive samples and top two irrelevant images as negative samples. We set each ?k by (12) and reordered the images using the relevance feedback model excluding the ones used as user relevance judgements. Those images were omitted from precision-recall calculations as well. Figure 5 illustrates that relevance feedback slightly boosts the retrieval performance, especially for the PASCAL-Sentence dataset. The computational complexity of an iteration is $O(K 2 + KD?)$ for a query and $O(N (K 2 + KD? + KDv))$ for training [19]. The feature assignment vector $z0$ of a query usually converges in a few 7

Query
Retrieval Set
A bird perching on a tree
A boat sailing along a river
A furniture located in a room
A child sitting in a room
A flower pot placed in a house
(b) Text-to-image queries
(a) Image-to-image queries

Figure 4: Sample images retrieved from the PASCAL-Sentence dataset by our method (iIBP) iterations. A typical query took less than 1 second in our experiments for I = 50 with our optimized Matlab code. Text Query w/ feedback Text Query w/o feedback

0.7
0.012
0.6
0.01
Mean Precision
0.5

0.5 0.45
0.006
0
0.2
0.4
0.6
Recall
0.8

(a) PASCAL-Sentence Dataset (K = 23)

1
0 0
0.2
0.15
0.002
0.2
0.3
0.25
0.004
0.3
0.4
0.35
0.008
0.4
0.1
0.55
Mean Precision
0.014
Mean Precision
0.8
Image Query w/ feedback Image Query w/o feedback
0.1 0.2
0.4
0.6
Recall
0.8

(b) SUN Dataset ? Class label ground truth (K = 45)

1
0.05
0
0.2
0.4
0.6
Recall
0.8
1

(c) SUN Dataset ? Euclidean ground truth (K = 45)

Figure 5: The result of category retrieval by our approach (iIBP) with relevance feedback for text and image queries. Revised retrieval with relevance feedback is compared with initial retrieval.

4

Conclusion

We proposed a novel retrieval scheme based on binary latent features for multimodal data. We also describe how to utilize relevance feedback for better retrieval performance. The experimental results on real world data demonstrate that our method outperforms state-of-the-art hashing techniques. In our future work, we would like to develop a user inference to get relevance feedback and a deterministic variational method for inference the integrative IBP based on a truncated stick-breaking approximation. Acknowledgments This work was supported by the NSF Grant 12621215 EAGER: Video Analytics in Large Heterogeneous Repositories. 8

2 References

[1] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In Advances in Neural Information Processing Systems 21, pages 1753?1760, 2009. [2] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pages 1?8, June 2008. [3] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In Proceedings of the 25th International Conference on Very Large Data Bases, VLDB ?99, pages 518?529, 1999. [4] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In IEEE 12th International Conference on Computer Vision, 2009, pages 2130?2137, Sept 2009. [5] M. Raginsky and S. Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In Advances in Neural Information Processing Systems 22, pages 1509?1517, 2009. [6] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pages 817?824, June 2011. [7] J.-P. Heo, Y. Lee, J. He, S.-F. Chang, and S.-E. Yoon. Spherical hashing. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pages 2957?2964, June 2012. [8] 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 Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349? 1380, Dec 2000. [9] M. M. Bronstein, E. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pages 3594?3601, June 2010. [10] S. Kumar and R. Udupa. Learning hash functions for cross-view similarity search. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Two, IJCAI?11, pages 1360?1365, 2011. [11] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In Advances in Neural Information

Processing Systems 25, pages 1376?1384, 2012. [12] Y. Zhen and D.-Y. Yeung. A probabilistic model for multimodal hash function learning. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ?12, pages 940?948, 2012. [13] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. In Advances in Neural Information Processing Systems 25, pages 2222?2230, 2012. [14] M. Rastegari, J. Choi, S. Fakhraei, H. Daume III, and L. S. Davis. Predictable Dual-View Hashing. Proceedings of The 30th International Conference on Machine Learning, pages 1328?1336, 2013. [15] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pages 2160?2167, June 2012. [16] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems, 8(6):536?544, 2003. [17] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan. A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4):723?742, April 2012. [18] Z. Ghahramani and T. L. Griffiths. Infinite latent feature models and the indian buffet process. In Advances in Neural Information Processing Systems 18, pages 475?482, 2005. [19] F. Doshi-Velez and Z. Ghahramani. Accelerated sampling for the indian buffet process. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ?09, pages 273?280, 2009. [20] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV?10, pages 15?29, Berlin, Heidelberg, 2010. [21] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pages 2751? 2758, June 2012. [22] D. Lin. An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML ?98, pages 296?304, 1998. [23] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pages 3485?3492, June 2010. [24] H. Hotelling. Relations Between Two Sets of Variates. Biometrika, 28(3/4):321?377, December 1936.

9