Large Scale Medical Image Search via Unsupervised PCA Hashing

Xiang Yu, Shaoting Zhang, Bo Liu, Lin Zhong, Dimitris N. Metaxas
Department of Computer Science, Rutgers University
Piscataway, NJ, USA
{xiangyu, shaoting, lb507, linzhong, dnm}@cs.rutgers.edu

Abstract

Medical image search is a significant way to provide similar clinical cases for doctors. Text based and content based image retrieval techniques have been widely investigated in the last decades. However, handling text-missing images and large scale medical database is still challenging. Traditional methods may encounter unsolvable efficiency problem or storage problem when tackling millions of images with general computers. In this paper, we employ an efficient PCA hashing based method for mapping raw features into locality preserving binary code. We focus on investigating the efficiency of PCA hashing while maintaining its competitive performance in medical image search. Ranking aggregation is used to achieve fusion of different features or fusion of retrieval results, which significantly improves single feature retrieval rate and thus compensates the overall accuracy. Without significantly sacrificing the retrieval accuracy, the benefit is a huge gain in physical memory and runtime efficiency. Experimental results show that hashing methods achieve far lower memory and far less time consuming handling large scale database.

1. Introduction

Last half century has witnessed fast development of digitalized medical image acquisition techniques, such as, Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and ultrasound. These various modalities of medical images provide significant help for diagnosis. However, as the amount of images is growing explosively, how to retrieve these images efficiently and effectively is becoming an urgent issue.

Searching by image content has been extensively investigated in the past several decades [20, 22, 15]. An amount of applications appear in medical research and diagnostics [14, 3, 18]. Ranking images by content, i.e., image feature similarity, we can find candidates from the image repository, which are most relevant to the query image. A comprehensive overview about the development of medical image retrieval can be found in [14]. Dina *et al.* [4] proposed a supervised method to combine text and image information for annotation and retrieval. Devrim *et al.* [24] focused on extracting efficient features for medical image search. Kim *et al.* [10] explored the information of relevant regions in medical Content Based Image Retrieval (CBIR). Most of the previous investigations on medical image retrieval are based on small datasets, in which case, the storage and computational complexity are not major concerns. However, as database volume and feature dimensionality are increasing, those methods with promising performance in small datasets become impractical, if both of the time complexity and physical memory are taken into consideration. This motivates us to design a medical image search framework which can potentially tackle scalable image database.

Feature similarity measurement lies in the heart of CBIR and it largely determines the efficiency of image retrieval systems. To achieve fast similarity search in large scale dataset, vocabulary tree based methods [2, 15] and hashing based methods [19, 25] are explored to seek a tradeoff between search precision and efficiency. Many previous systems in medical image retrieval are based on the vocabulary tree, while hashing methods haven't been widely applied in this area. Hashing methods are effective because of their lower memory requirement and higher efficiency, especially when data dimension is high. It has been theoretically and experimentally proved that binary code derived from hashing can map similar images to the same entry with high probability [17]. A large efficiency gain in data storage and computation can be achieved by the compact binary code rather than the original feature. One of the pioneering works [23, 21, 6, 30] of this method is Locality Sensitive Hashing (LSH) [1, 11], which randomly projects the data and generates binary code with a random threshold. Recently, by making use of the data distribution, many data dependent code learning methods have been proposed [7, 27, 26, 13, 12]. These methods can better preserve feature similarity compared with data independent hashing methods, such as LSH, of which compact binary codes are measured by hamming distance.

Encouraged by the success of the above mentioned hashing methods in large scale image search, in this paper we aim to leverage *data-dependent* hashing-based techniques to assist scalable medical image search. In our framework, each image is initially denoted by a high dimensional feature descriptor. As many other binary coding methods do, we employ Principal Component Analysis (PCA) first to reduce data dimensionality. Furthermore, to get the compact binary code, we employ a recently proposed *optimal rotation* based hashing method [7], which minimizes the error between the principal component features and the derived binary codes. We finally conduct search experiments on public scalable medical image database ImageCLEF [9] to validate the effectiveness of the hashing method with standard retrieval measurements. Results demonstrate that retrieval time and memory are far less than general methods while precision is little sacrificed.

The rest of the paper is organized as following. Section 2 introduces our medical image retrieval framework dealing with large image repository. Experiments are conducted in section 3, comparing hashing results with original feature results and the rank aggregated result with single feature results. Section 4 concludes our work according to the experimental results.

2. Methodology

2.1. Framework

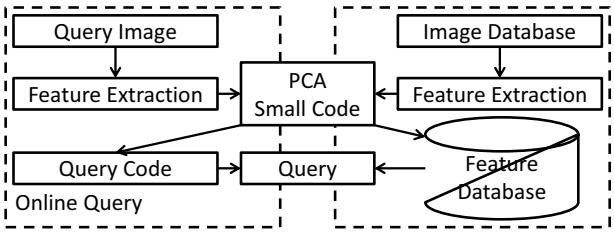


Figure 1. The workflow of Medical Image Retrieval Framework consisting the offline storage and online query modules.

Our image query framework is illustrated in Figure 1. It mainly includes two parts: (1) Off-line database construction. By extracting raw features, such as GIST [16], HSV and RGB, we obtain visual image features including texture, color and illumination. Then features are compressed using our PCA hashing method. The binary codes representing medical images are stored at feature database. (2) Online query. Given a pathological query image, the same feature extraction step is taken and the feature vector of query image is also translated into binary codes. As feature fusion techniques can further improve the performance [5, 8, 29, 28], after finding Nearest Neighbors (NNs) from different feature databases, our framework would generate the query list by aggregating these rank results [5]. This is an efficient way to improve accuracy. Nearest neighbors reflect local topological structure essentially. As long as the binary codes preserve the local topology, NN is a canonical and efficient way to depict the similarity.

The online query part is undoubtedly the key factor to decide retrieval performance. While running the query system, the feature database is pre-loaded into PC memory for efficiency consideration. Since the image number could be millions, our PCA hashing technology crucially determines whether the PC memory will overflow. In running time aspect, feature extraction, feature encoding and similarity calculation are three time consuming steps. Therefore, we design and choose most expressive features, develop efficient algorithm of binary code and implement the rank aggregation. Figure 2 shows some visual retrieval rankings of medical images.

2.2. PCA and Optimal Rotation based Hashing

Generally, optimization of time complexity and feature storage comes from the dimensionality reduction of feature vectors. Karhunen-Loeve Transform (KLT) linearly projects data into uncorrelated dimensions. Further, the principal components are picked to remove noise and insignificant components while preserving most information of original features. So we search a way linearly projecting raw features into orthogonal principal components and binarizing the reduced features with minimal loss.

PCA based Hashing: Raw features (e.g. GIST, HSV) are usually hundreds or thousands of float numbers long, and may be redundant and noisy. Our goal is to map the primal features into concise but expressive codes. Such mapping could keep the most information of the original data. Larger variance reveals more information according to Shannon's information theory. In the mean while, from statistical aspect, redundancy mainly refers to the correlation among feature vectors. For example, two feature vectors have no correlation if they are orthogonal to each other. Thus, we expect to find a transformation T such that the new feature vectors maintain the largest variance while they are orthogonal to each other. The objective function is:

$$\arg\max_T V(T) = \arg\max_T \sum_k \mathbf{E}(\|xt_k\|_2^2) \qquad (1)$$

Ensembling each feature vector x into a feature matrix X:

$$V(T) = \frac{1}{n}\sum_k t_k X^T X t_k = \frac{1}{n} tr(T^T X^T X T), \quad T^T T = I \qquad (2)$$

From above, actually we take Principal Component Analysis (PCA) on the feature matrix X. We try to find the projection matrix T on $X^T X$ so that the variance, which

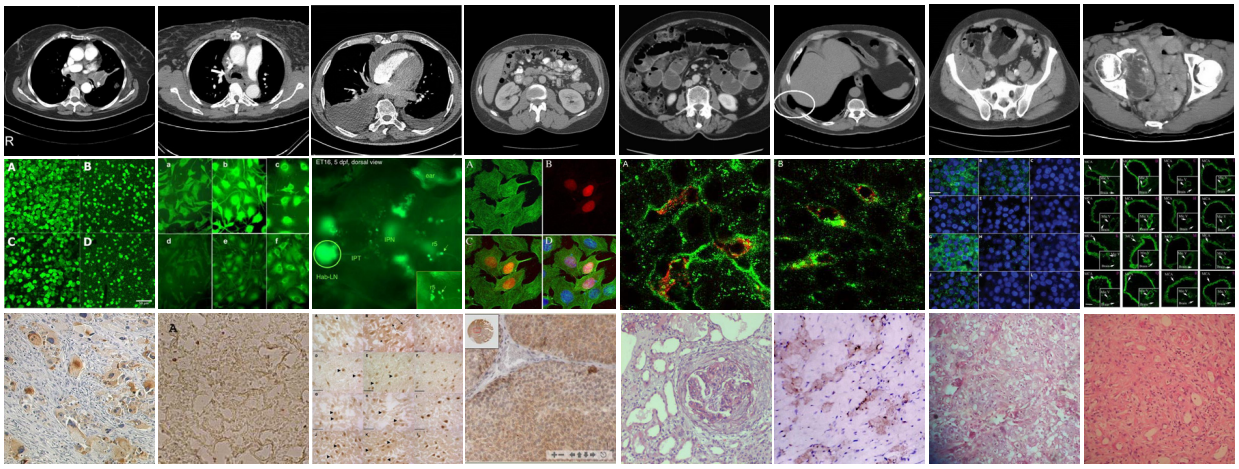


Figure 2. Visual modality retrieval rankings of Medical images, computed tomography (CT), fluorescence microscopy (FL) and histopathology (HX), the first column are query images, the other columns are retrieval images by ranking

are the eigenvalues of $X^T X$, is maximized. If the limited length of shortened feature is a, we would pick the first a eigenvalues and their corresponding eigenvectors to form the projection matrix T. A straightforward way to hash the reduced feature is binarizing the reduced feature into 0-1 code. However, brute force binarization incurs large quantization error, which cannot preserve the locality of original features.

Optimal Rotation based PCA Hashing: To solve the above mentioned problem, we seek to further shrink the features while keeping the feature space's neighborhood structure. Hashing based method would maintain the neighborhood property by randomized hashing functions. In the same way, we want to search some mapping function to keep the neighborhood structure. Getting aid from image encoding area, quantization is a significant strategy to compress the image size. Here if the feature vector is quantized into binary code, the storage of feature database would decrease greatly. Thus, we seek to minimize the binarization error between the binarized feature matrix Q and the feature matrix L generated by PCA. However, since $Q = sgn(L)$ (sign function outputs 1 for positive input, otherwise 0) is determined with fixed L, it seems the error is a fixed quantization error.

Fortunately, the projection matrix W is not unique. Different W may cause different quantization errors. Actually, all those W are linearly dependent with each other. The translation is from one coordinate system to another. Equivalently, we could formulate the translation by a rotation matrix R. If we add such orthogonal rotation matrix R onto the projection matrix, it still holds the optimality for new projection matrix TR at PCA step. Thus, our aim is to find R, such that the quantization error is minimized.

$$\phi(L,R) = \arg\min_{R} \|Q - LR\|_F^2, Q = sgn(LR) \quad (3)$$

The quadratic program is hard to solve by optimization problem solvers because sign function can not be differentiated. We take Frobenius norm measurement during the optimization process. Intuitively, given two variables, in order to pursue the minimum, we would fix one variable and minimize the objective function over the other and vice versa. The basic idea is to alternatively fix R and Q and push the objective function towards local minima.

(1) If R is fixed, the problem becomes:

$$\phi(L,R) = \|Q\|_F^2 + \|LR\|_F^2 - 2tr(Q^T LR) \quad (4)$$

Since Q is $n \times t$, $\|Q\|_F^2 \leq na$. The R step is:

$$\begin{aligned}\tilde{\phi}(L,R) &= \arg\max tr(Q^T LR) \\ &= \arg\max\Big(\sum_{i=1}^{n}\sum_{j=1}^{t} Q_{ij}P_{ij}\Big), P = LR \end{aligned} \quad (5)$$

As $Q_{ij} = \{1, -1\}$, P and R are fixed, Q_{ij} must be positive 1 if P_{ij} is positive and negative 1 if P_{ij} is negative to achieve maximum. Thus $Q_{ij} = P_{ij}$.

(2) If Q is fixed from the last step, we want to search new R to minimize the quantization error. It is actually the orthogonal Procrustes problem

$$\tilde{\phi}(L,R) = \arg\max tr(RQ^T L) \quad (6)$$

$$\begin{aligned} &= \arg\max tr(RU\Lambda S^T) \\ &= \arg\max tr(\Lambda S^T RU), Q^T L = U\Lambda S^T \end{aligned} \quad (7)$$

Equation 6 and Equation 7 hold because diagonal elements remain the same no matter the order of two matrices multiplication. Since R and $Q^T L$ are both $a \times a$ matrices, we consider the SVD decomposition over $Q^T L$. Since S, R and U are all orthogonal matrices, $S^T RU$ is also an orthogonal matrix, whose entries are no greater than 1. In order to

Algorithm 1 Alternative R&Q optimization.

Input: Initialized random rotation matrix $R \in \mathbf{R}^{a\times a}$, data matrix L after PCA.
Output: Optimized rotation matrix R.
repeat
R **step:** fix R, $Q = sgn(LR)$
Q **step:** fix Q, do SVD on $Q^T L$,$Q^T L = U\Lambda S^T$, $R = SU^T$
until halting criterion is true

meet the maximum, those diagonal entries of $S^T RU$ must be 1, and all other entries equal 0. Otherwise the output is just a portion of the diagonal elements of Λ:

$$I = S^T RU \Leftrightarrow R = SU^T \tag{8}$$

Though the approach is not guaranteed to reach global minima, the near-optimal objective value could bring us good enough retrieval precision but takes milliseconds, which can be verified from the experiments. The procedure is summarized in Alg.1.

The stop criterion of Alg.1. can be set as limited iteration number or two consecutive quantization errors' gap is within certain fixed margin. Practically both of the two methods are effective to obtain R and Q.

3. Experiments

3.1. Experimental settings

Our experiments are conducted on the public ImageCLEF [9] medical database, containing 231k images with 18 different modalities, such as CT, MRI, ultrasound, etc. The modality information of 137k images from ImageCLEF database is manually labeled as the ground truth for evaluation. We take the 5 largest volume modalities as our retrieval tasks, which are graphs(GX), histopathology(HX), computed tomography(CT), fluorescence(FL) and x-ray(XR). Our goal is to demonstrate the effectiveness of the optimal rotation based PCA hashing technique.

In feature extraction aspect, we use GIST feature, HSV and RGB color feature. GIST feature depicts texture information of images, which is also known as Gabor multi-scale multi-directional wavelets. In Gabor model with Euclidean distance measurement, noise is assumed Gaussian and thus feature difference is *l2*-norm based. HSV and RGB are typical color models for feature extraction step.

For different features, we run 545 query images which are not in the 137k image database. Our main concern is: when features are reduced by PCA, and they are continuously binarized by PCA hashing, how would the performance decrease from without such dimensionality reduction strategies. Thus we design four main tasks: (1) original GIST with *l2* measurement (GIST), (2) applying PCA on GIST using *l2* measurement(GIST-PCA), (3) Binarizing the PCA result using hamming measurement(GIST-PCA-BIN), (4) The hashing method using hamming measurement(GIST-PCA-SC). Note that those four tasks are raised on HSV and RGB features in parallel. In the last step, rank aggregation to combine GIST, HSV and RGB results is adopted to boost the performance.

3.2. Experimental Results

Figure 3 presents precision-recall curve comparison on those four step-in methods, GIST, GIST-PCA, GIST-PCA-BIN and our method GIST-PCA-SC. GIST performs the best of the four methods, since other methods lose information through dimensionality reduction. Using only 32-256 bits, the hashing method shown in red curve, is just slightly worse than the GIST and GIST-PCA methods, but far above the PCA hashing GIST-PCA-BIN result. The precision gap between the hashing one and GIST is within 0.1 when code length is 256 bits. As the code length decreases from 256 to 32 bits, it can be noticed that the precision gap between GIST-PCA-SC and GIST does not increase significantly. The gap remains within 0.1 no matter the recall rate is, which shows that the hashing method is not sensitive to specific query image. All the evidence reveals that the performance of the hashing method does not drop significantly when code length varies from 256 to 32 bits.

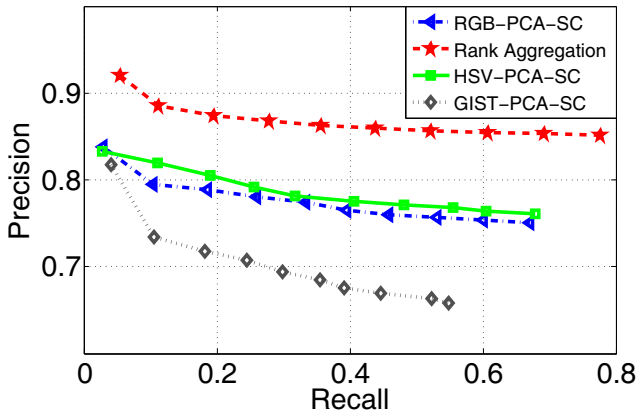
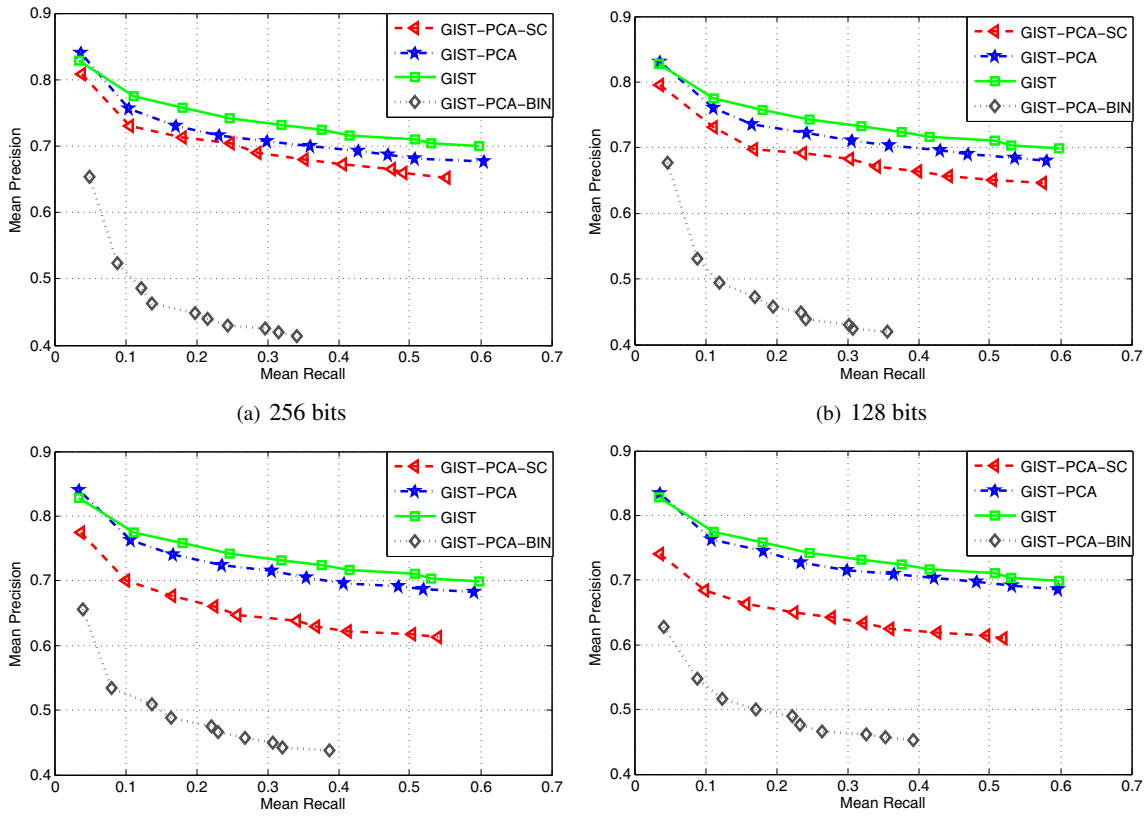


Figure 4. Rank aggregation on GIST/HSV/RGB-PCA-SC retrieval results.

Table 1 compares the other three methods, the hashing method with 256, 128, 64 and 32 bits and the final Rank aggregation result from GIST, HSV and RGB features. They are measured based on precision at top 5, 10 retrievals, mean Average Precision (mAP), query time and memory cost. Quantitatively, the largest gap between GIST and GIST-PCA-SC is no larger than 0.1 and again convinced that the drop of performance due to code length compression is not striking. Moreover, the final rank aggregated result performs even better than original GIST method from



(a) 256 bits (b) 128 bits

(c) 64 bits (d) 32 bits

Figure 3. Precision recall curve for the hashing method (GIST-PCA-SC) with 256, 128, 64 and 32 bits binary code, compared with original GIST feature, applying PCA on GIST (GIST-PCA), binarizing on PCA result (GIST-PCA-BIN).

Table 1. Comparison of GIST, GIST-PCA, GIST-PCA-BIN and GIST-PCA-SC (which is denoted as SC in the table) methods over precision at 5, 10, mean Average Precision (mAP), query time for one search and memory cost of database 137k images.

	GIST	GIST PCA	GIST PCA,BIN	SC (256 bits)	SC (128 bits)	SC (64 bits)	SC (32 bits)	Rank Aggregate
P@5	0.776	0.765	0.532	0.737	0.732	0.713	0.689	**0.782**
P@10	0.760	0.734	0.492	0.714	0.700	0.679	0.667	**0.753**
mAP	0.771	0.751	0.527	0.738	0.725	0.700	0.685	**0.767**
time(ms)	0.998	0.998	0.0172	0.0172	0.0116	0.0041	0.0038	**0.0172**
memory(bits)	6.2G	1.6G	26M	26M	13M	6.5M	3.2M	**26M**

Table 1. Though hashing decreases accuracy by losing information from original features, rank aggregation attempts to combine features from different aspects and improve the accuracy. The striking improvement is memory. We inferred the GIST memory by its storage on hard disk which is equivalent in memory as the training data could not be fully loaded into memory. The statistics reveals that hashing binary code brings 238 times less memory than original GIST method based on 256 bits binary code, which makes it possible to load the whole database into memory. For running time evaluation, we provide the average time for one search in the database. Given our 137k dataset, we need 137 seconds to finish the whole query of the dataset by GIST. But only 2.35 seconds are needed by 256 bits GIST-PCA-SC to accomplish a query. The case of 32 bits approaches 0.52 seconds to complete a query of 137k dataset, which makes the query of whole database real time. Figure 4 shows the result of Rank Aggregation on GIST-PCA-SC, HSV-PCA-SC and RGB-PCA-SC retrieval results, from which we obtained an obvious enhancement in the final accuracy performance, which is a significant improvement of single feature methods.

Experimental results demonstrate that the hashing method achieves better performance than original single feature methods. Further the feature size shrinks from thousands of float numbers to 256 bits, which leads to 238 times less memory consuming than original GIST method. While the query time decreases from 137 seconds to 0.52 seconds, which is 263 times faster and guarantees instant query for entire database.

4. Conclusions

In this paper, we employed a PCA hashing based technique to achieve scalable search in medical image database. It overcomes fallacies of traditional retrieval methods in volume and speed aspects. Verified from public database and standard evaluation, it is able to tackle million-volume image repository in real time almost without pulling down performance in accuracy. Low memory cost and real-time query speed reveal its wide applicability in medical image field. However, in this paper we only tested on the modality recognition problem, which is a relatively easy task in medical image retrieval. Future work includes validation on challenging clinical use cases.

References

[1] A.Andoni and P.Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Communications of the ACM*, 2008.

[2] C. Bohm, S. Berchtold, and D. A. Keim. Searching in high-dimensional spaces-index structures for improving the performance of multimedia databases. *ACM Computing Survey*, 33:322–373, 2001.

[3] J. Caicedo, F. Gonzalez, and E. Romero. Content-based medical image retrieval using low-level visual features and modality identification. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152, pages 615–622. 2008.

[4] D. Demner-Fushman, S. Antani, M. Simpson, and G. R.Thoma. Annotation and retrieval of clinically relevant images. *International Journal of Medical Informatics*, 78:59–67, 2009.

[5] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In *ACM SIGMOD*, 2003.

[6] R. Fergus, A. Torralba, and Y. Weiss. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.

[7] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011.

[8] H. Jegou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *IEEE Trans. Pattern Anal. Machine Intell.*, 32:2–11, 2010.

[9] J. Kalpathy-Cramer, H. Muller, S. Bedrick, I. Eggel, A. G. S. de Herrera, and T. Tsikrika. Overview of the clef 2011 medical image classification and retrieval tasks, 2011.

[10] E. Kim, S. Antani, X. Huang, L. Long, and D. Demner-Fushman. Using relevant regions in image search and query refinement for medical cbir. In *SPIE Medical Imaging*, 2011.

[11] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. In *ICCV*, 2009.

[12] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised hashing with kernels. In *CVPR*, pages 2074–2081. IEEE, 2012.

[13] W. Liu, J. Wang, S. Kumar, and S.-F. Chang. Hashing with graphs. In *ICML*, pages 1–8, 2011.

[14] H. Muller, N. Michoux, D. Bandon, and A. Geissbuhler. A review of content-based image retrieval systems in medical applicationsłclinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23, 2004.

[15] D. Nister and H. Stewenius. Scalable recognition with a vocabular tree. In *CVPR*, pages 2161–2168, 2006.

[16] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[17] M. Raginsky and S. Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In *The Neural Information Processing Systems*, 2009.

[18] M. Rahman, P. Bhattacharya, and B. Desai. A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *Trans on Information Technology in Biomedicine*, 11(1):58 –69, 2007.

[19] R.Salakhutdinov and G.Hinton. Semantic hashing. In *SIGIR*, 2007.

[20] Y. Rui, T. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[21] R. Salakhutdinov and G. Hinton. Semantic hashing. In *SIGIR*, 2007.

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[23] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, 2008.

[24] D. Unay, A. Ekin, and R. Jasinschi. Medical image search and retrieval using local binary patterns and klt feature points. In *ICIP*, pages 997–1000, 2008.

[25] J. Wang, S. Kumar, and S.-F. Chang. Semi-supervised hashing for scalable image retrieval. In *CVPR*, pages 3424–3431, 2010.

[26] J. Wang, S. Kumar, and S.-F. Chang. Sequential projection learning for hashing with compact codes. In *ICML*, 2010.

[27] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008.

[28] S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. Metaxas. Automatic image annotation using group sparsity. In *CVPR*, 2010.

[29] S. Zhang, M. Yang, T. Cour, K. Yu, and D. Metaxas. Query specific fusion for image retrieval. In *ECCV*, 2012.

[30] X. Zhang, Z. Li, L. Zhang, W. Ma, and H.-Y. Shum. Efficient indexing for large scale visual search. In *ICCV*, 2009.