SCALABLE MAMMOGRAM RETRIEVAL USING ANCHOR GRAPH HASHING

Jingjing Liu[⋆] *Shaoting Zhang*[†] *Wei Liu*[‡] *Xiaofan Zhang*[†] *Dimitris N. Metaxas*[⋆]

[⋆] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[†] Department of Computer Science, University of North Carolina at Charlotte, NC, USA
[‡] IBM T. J. Watson Research Center, NY, USA

ABSTRACT

Mammogram analysis is known to provide early-stage diagnosis of breast cancer in reducing its morbidity and mortality. In this paper, we propose a scalable content-based image retrieval (CBIR) framework for digital mammograms. CBIR is of great significance for breast cancer diagnosis as it can provide doctors image-guided avenues to access relevant cases. Clinical decisions based on such cases offer a reliable and consistent supplement for doctors. In our framework, we employ an unsupervised algorithm, Anchor Graph Hashing (AGH), to compress the mammogram features into compact binary codes, and then perform searching in the Hamming space. In addition, we also propose to fuse different features in AGH to improve its search accuracy. Experiments on the Digital Database for Screening Mammography (DDSM) demonstrate that our system is capable of providing content-based accesses to proven diagnosis, and aiding doctors to make reliable clinical decisions. What's more, our system is applicable to large-scale mammogram database, such that high number analogical cases would be retrieved as clinical references.

Index Terms— Digital mammogram, scalable image retrieval, hashing, Hamming space

1. INTRODUCTION

Breast cancer is the second-most common and deadly cancer among women [1]. Since the cause of breast cancer is undiscovered, for the time being, there are no effective ways to prevent it. Fortunately, due to the adoption of mammography screening, early-stage diagnosis of breast cancer significantly reduces its morbidity and mortality. However, breast cancer diagnosis in mammogram screening involves in error prone decision-making. In a pioneering work, it is reported that up to 30% of lesions are possible to be misinterpreted during routine screening [2].

Computer-aided diagnosis (CAD) can play as a clinical auxiliary in detecting the abnormalities in mammograms. A recent study shows the use of CAD in the interpretation of screening mammogram can increase the detection rate of early-stage malignancies [3]. In the past decades, many CAD techniques related to mammography have been proposed and attracted the attention of both computer scientists and radiologists. Most of these work focused on mass detection/classification [4] [5], and microcalcifications (MCs) detection/pattern classification [6] [7]. Regardless of improved detection rate, CAD systems commonly result in excessive false positives of malignancy, which would have adverse effect on clinical decision-making [8].

In recent years, researchers become incrementally interested in content-based image retrieval (CBIR) for medical images [9] [10] [11]. Specifically for mammogram analysis, CBIR can provide doctors with content-based manner to get accesses to clinically analogical cases. These cases of visual similarities can further facilitate decision-making on breast cancer. Different from CAD which computes the likelihood of malignancy, in practice, CBIR aims at providing radiologists with proven diagnosis and other suitable information, by recalling mammograms of past cases visually relevant to a query [12] [13] [14]. With the popularity of mammography, mammogram are available in ever increasing quantities. Consequentially, leveraging clinical information from large rather than small mammogram database becomes more pivotal. Retrieval on a large number of mammographic cases could provide comprehensive reference to radiologists. However, to the best of our knowledge, few effort has been devoted to scalable mammogram retrieval.

In this paper, we investigate a scalable mammogram retrieval system on more than 5222 mammographic ROIs obtained from the Digital Database of Screening Mammography (DDSM). Encouraged by the recent success of hashing methods on scalable web-image retrieval [15] [16], we employ the *Anchor Graph Hashing* (AGH) approach [17]. AGH derives compact binary codes from mammograms that preserve neighborhood structure inherent in image feature space with high probability, thus resulting in less memory space and computation complexity. In addition, we propose to seamlessly fuse both holistic and local features in AGH on the distance level. We conduct experiments on the aforementioned mammogram repository, to evaluate both retrieval precision and classification accuracy.

2. METHODOLOGY

Given a mammographic ROI, the CBIR seeks out relevant cases in targeted database, based on visual similarities. The

978-1-4673-1961-4/14/$31.00 ©2014 IEEE

framework of our retrieval system is illustrated in Fig.1. It consists of two main phases: offline learning and online query. During the offline phase, we extract image features from mammogram database and compress them into binary codes, by using *Anchor Graph* and spectral embedding. Such binary codes preserve the similarities in original image feature space with high probability. In the online phase, the image features of a query ROI are also converted into binary codes, with generalized hashing functions. Then we perform efficient searching in Hamming space to retrieve the most similar ROIs with smallest distance. The proven diagnosis of these retrieved ROIs can facilitate clinical decision-making on the query mammogram.

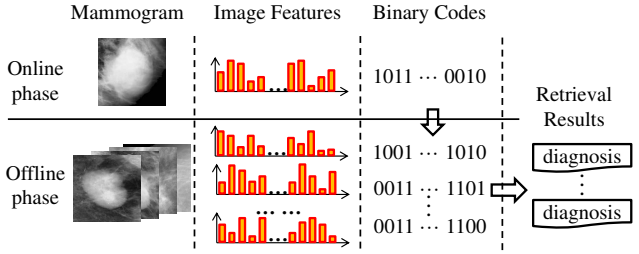

Fig. 1. Framework of mammogram retrieval using hashing.

Spectral Embedding: The r-bit Hamming embedding for n data points is obtained by minimizing mapping errors (for conciseness, a data point here refer to the visual features of one mammographic ROI). Using spectral relaxation, the optimization formulation becomes:

$$\min_{Y} \quad \frac{1}{2}\sum_{i,j=1}^{n}\|Y_i - Y_j\|^2 A_{ij} = \text{tr}\left(Y^T L Y\right)$$
$$\text{s.t.} \quad Y \in \mathbb{R}^{n\times r},\ \mathbf{1}^T Y = 0,\ Y^T Y = nI_{r\times r} \tag{1}$$

where A is the $n \times n$ similarity matrix. The graph Laplacian is then defined as $L = D - A$ with $D = \text{diag}(A\mathbf{1})$ ($\mathbf{1} = [1, ..., 1]^T \in \mathbb{R}^n$). The solution Y is given by r eigenvectors corresponding to the r smallest eigenvalues (abandoning eigenvalue 0) of L. Final desired binary codes are given by $\text{sgn}(Y) : \mathbb{R}^d \mapsto \{1, -1\}^r$.

Anchor Graph: The main drawback of the above formulation is the intractable cost ($\mathcal{O}(dn^2)$) of building the underlying graph for large n. One solution is to use an efficient Anchor Graph that employs a small set of m *anchors* $\mathcal{U} = \{\boldsymbol{u}_j \in \mathbb{R}^d\}_{j=1}^m$ ($m \ll n$). Then, similarities between n points and m anchors can be measured as:

$$Z_{ij} = \begin{cases} \dfrac{\exp\left(-\mathcal{D}^2\left(\boldsymbol{x}_i, \boldsymbol{u}_j\right)/t\right)}{\sum_{j'\in\langle i\rangle}\exp\left(-\mathcal{D}^2(\boldsymbol{x}_i, \boldsymbol{u}_{j'})/t\right)}, & \forall j \in \langle i\rangle \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\langle i\rangle \subset [1 : m]$ denotes the indices of s ($s \ll m$) nearest anchors of $\boldsymbol{x}_i$ in $\mathcal{U}$ according to the distance function $\mathcal{D}(\cdot)$; t is the bandwidth parameter.

This provides an approximation to adjacency matrix A as $\hat{A} = Z\Lambda^{-1}Z^T$. It is low-rank and doubly stochastic, where $\Lambda = \text{diag}(Z^T\mathbf{1})$. The consequent graph Laplacian of Anchor Graph is $L = I - \hat{A}$. The spectral embedding matrix Y is obtained by solving the eigenvalue system of a small matrix $M = \Lambda^{-1/2}Z^T Z\Lambda^{-1/2}$ instead of L. Let $\{\sigma_1, ..., \sigma_r\}$ ($1 > \sigma_1 \geq ... \geq \sigma_r > 0$) denote the r eigenvalues of M, $V = [\boldsymbol{v}_1, ..., \boldsymbol{v}_r]$ the corresponding eigenvectors, and $\Sigma = \text{diag}(\sigma_1, ..., \sigma_r) \in \mathbb{R}^{r\times r}$:

$$Y = \sqrt{n}Z\Lambda^{-1/2}V\Sigma^{-1/2} = ZW \tag{3}$$

where $W = [\boldsymbol{w}_1, ..., \boldsymbol{w}_r] \in \mathbb{R}^{m\times r}$, $\boldsymbol{w}_k = \sqrt{n/\sigma_k}\Lambda^{-1/2}\boldsymbol{v}_k$

Hashing Functions: Eq.(3) generates binary codes only for offline mammographic ROIs. To handle any "out-of-sample" online query, graph Laplacian eigenvectors are extended to general hash function $h : \mathbb{R}^d \mapsto \{1, -1\}$, using Nyström method. Given m anchor points $\mathcal{U} = \{\boldsymbol{u}_j\}_{j=1}^m$ and an arbitrary point $\boldsymbol{x}$, the hash functions used in the AGH are:

$$h_k(\boldsymbol{x}) = \text{sgn}\left(\boldsymbol{w}_k^T\boldsymbol{z}(\boldsymbol{x})\right),\ k = 1, ..., r. \tag{4}$$

$$\boldsymbol{z}(\boldsymbol{x}) = \frac{\left[\delta_1\exp\left(-\frac{\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_1)}{t}\right), ..., \delta_m\exp\left(-\frac{\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_m)}{t}\right)\right]^T}{\sum_{j=1}^m \delta_j\exp\left(-\frac{\mathcal{D}^2(\boldsymbol{x},\boldsymbol{u}_j)}{t}\right)} \tag{5}$$

where $\delta_j \in \{1, 0\}$ and $\delta_j = 1$ if and only if anchor $\boldsymbol{u}_j$ is one of s nearest anchors of query $\boldsymbol{x}$ in $\mathcal{U}$, according to the distance function $\mathcal{D}(\cdot)$.

The algorithm of AGH is shown in Alg.1. Owing to the bit-manipulation on binary codes, AGH needs little physical memory and could achieve linear time in training and constant time in searching a new query, without significant loss of precision. Therefore, our system is competitive for scalable mammogram retrieval.

Algorithm 1 Anchor Graph Hashing Algorithm.

Input:

n data points $\{\boldsymbol{x}_i\}_{i=1}^n$ and one arbitrary query sample $\boldsymbol{x}$.

Hamming Embedding:

1. Obtain m anchor points $\mathcal{U} = \{\boldsymbol{u}_j \in \mathbb{R}^d\}_{j=1}^m$;
2. Compute graph Laplacian $L = I - Z\Lambda^{-1}Z^T$ of Anchor Graph, using eq.(2);
3. By minimizing eq.(1), we obtain $Y = ZW$, where $W = \sqrt{n}\Lambda^{-1/2}V\Sigma^{-1/2}$; the corresponding binary codes for the n points are sgn(Y);

Query Hashing:

Using eq.(5), $\boldsymbol{x}$'s kth-bit is binarized as $h_k(\boldsymbol{x}) = \text{sgn}\left(\boldsymbol{w}_k^T\boldsymbol{z}(\boldsymbol{x})\right)$, where $\boldsymbol{w}_k = \sqrt{n/\sigma_k}\Lambda^{-1/2}\boldsymbol{v}_k$.

Output:

$\boldsymbol{x}$'s top-K nearest points in $\{\boldsymbol{x}_i\}_{i=1}^n$.

Feature Fusion in AGH: Traditional AGH is only able to compress one type of image feature. However, as demonstrated in many image analysis problems, one feature may not be able to capture comprehensive information. Particularly, holistic and local features represent different yet complementary information [18] [19]. Therefore, we propose to improve traditional AGH by seamlessly fusing multiple features. By random sampling ROIs from training database as anchors when constructing Anchor Graph, we aggregate the SIFT features [20] and GIST features [21] on distance level (before the hashing stage). Joint Equal Contribution (JEC) method [22] is employed to compute the accumulated distances. We assign equal contributions to individual distances based on SIFT and GIST when calculating image similarities. In addition, based on the fact that AGH is independent of distance metric, we choose correlation and Euclidean distance for Bag-Of-Word (BoW) and GIST features respectively, to maximize the information gain. Using such feature aggregation scheme, we successfully extend AGH to incorporate both holistic and local information.

3. EXPERIMENTS

Experimental Setting: We validate our system on the DDSM database [23]. It contains mammograms acquired from 2470 persons, in each there are four images captured from different views (i.e., RIGHT_CC, RIGHT_MLO, LEFT_CC, and LEFT_MLO). All the images from different scanners are normalized corresponding to the same optical density. Since the clinical analysis of mammograms is based on the visual characteristics of suspicious regions of masses, we follow the conventional manner of extracting mammographic ROIs [12] [24], and then perform the retrieval task. Rectangular ROIs centered on the known location of each annotated mass (benign or malignant) are cropped. We also extract false positive mass (nonmass) ROIs from mammograms without mass issues, using a mass detection CAD system. We separate the database into offline and online categories with different patients. The online query database contains 34 benign, 68 malignant, and 115 false positive (FP) mass ROIs; the offline database contains 928 benign, 1246 malignant, and 2831 FP mass ROIs. In brief, we query 217 mammographic ROIs in the offline library of 5005 ROIs. 1000 dimensional BoW descriptor from SIFT features with *tf-idf* scheme, and 1024 dimensional GIST features are generated for each ROI. The following experiments are conducted on a Dell Workstation with a 3.4GHz processor of eight cores and 16G RAM.

Retrieval Precision: In our experiments, FP mass ROIs are labelled as nonmass class, and benign/malignant ROIs as mass. A retrieved ROI is regarded as relevant if it belongs to the same class of the query ROI. Precision is defined as the fraction of retrieved images that are relevant to a query image, and the retrieval performance is based on the average precision across all queries. We validate AGH with varying numbers of hash bits and plot the precision-scope curves, as shown in Fig.2(a). AGH with $r = 48$ obtains the best retrieval performance. Even with increasing scopes, the precisions still remain at least 80%. Based on 48-bit Hashing, we then illustrate precision-scope curves in Fig.2(b)-(d), considering overall, mass, and nonmass queries, respectively. Fusion of SIFT and GIST features improves the retrieval precision of mass ROIs, since mass ROIs with spiculated bright mass in the center should share holistic and local characteristics. Such improvement is weaken on nonmass ROI retrieval, which corresponds to the fact that global and local visual characteristics of nonmass issues are sometimes contradictory. Two mass ROIS queries and their retrieved cases are shown in Fig.3.

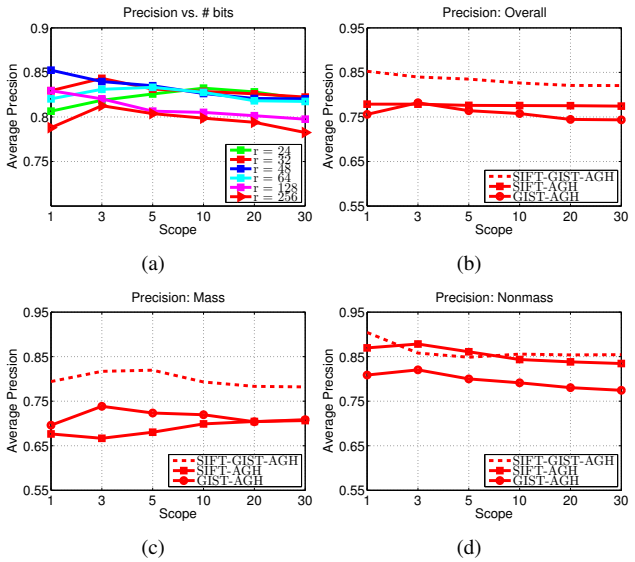

Fig. 2. Retrieval Precision-Scope curve: (a) Overall precision with different number of hash bits; (b)-(d) Precision of overall, mass, and nonmass queries with $r = 48$.

Classification Accuracy: Based on the types of retrieved ROIs, we further evaluate ROI classification performance, i.e., identifying a query ROI as mass or nonmass. We compare our system with two baseline methods: kNN and SVM. SVM slightly outperforms the other two models, which is unsurprising given that SVM is a supervised method. The overall classification accuracies of both SIFT-SVM and GIST-SVM are around 90%. However, SVM cannot provide relevant cases to doctors as clinical reference. Although with single features type the classification accuracy of AGH is slightly worse than kNN, after feature fusion, our system obtains better results than kNN. Compared with kNN ($N = 5$), the overall classification accuracy is enhanced by 2%, and 5% for mass class in particular. SIFT-GIST-AGH ($r = 48$, scope = 5) achieves 89.4% overall accuracy: 85.3% for mass ROIs and 93.04% for nonmass, which shows its capability of aiding clinical decisions. Moreover, our system is applicable to large scale databases.

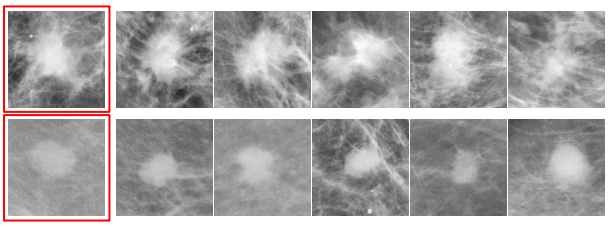

Fig. 3. Each row corresponds to one query ROIs (in red bounding boxes) and top 5 retrieved ROIs.

4. CONCLUSION

In this paper, we investigated the scalable mammogram retrieval using an unsupervised hashing method. By converting image features into binary codes, AGH achieves quick image search, without significant loss of precision. In addition, we also proposed to seamlessly fuse both holistic and local features in AGH, using different distance metrics during the construction of Anchor Graph. Our system obtains promising retrieval precision and classification in a database of mammographic ROIs. It can provide doctors with image-guided access to proven diagnosis, and aid further clinical decisions.

5. REFERENCES

[1] J. Tang, R.M. Rangayyan, J. Xu, I. El-Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Trans on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 236–251, 2009.

[2] R. Strickland and H. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *TMI*, vol. 15, no. 2, pp. 218–229, 1996.

[3] T.W. Freer and M.J. Ulissey, "Screening mammography with computer-aided detection: prospective study of 12860 patients in a community breast center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001.

[4] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," *PR*, vol. 39, no. 4, pp. 646–668, 2006.

[5] A.Oliver, J. Freixenet, J. Martí, E. Pérez, J. Pont, E. R.E. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[6] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcification in digitized mammogram films.," *TMI*, vol. 19, no. 2, pp. 115–126, 2000.

[7] H.D. Cheng, X. Cai, X. Chen, L. Hu, and X. Lou, "Computer-aided detection and classification of microcalcifications in mammograms: a survey," *PR*, vol. 36, no. 12, pp. 2967–2991, 2003.

[8] N. Houssami, R. Given-Wilson, and S. Ciatto, "Early detection of breast cancer: Overview of the evidence on computer-aided detection in mammography screening," *Journal of Medical Imaging and Radiation Oncology*, vol. 53, no. 2, pp. 171–176, 2009.

[9] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1–23, 2003.

[10] F. Valente, C. Costa, and A. Silva, "Dicoogle, a pacs featuring profiled content based image retrieval," *PLoS ONE*, vol. 8, no. 5, 2013.

[11] X. Yu, S. Zhang, B. Liu, L. Zhong, and D. Metaxas, "Large scale medical image search via unsupervised pca hashing," in *CVPRW*, 2013, pp. 393–398.

[12] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, and C. E. Floyd, "Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information," *Medical Physics*, vol. 30, no. 8, pp. 2123–2130, 2003.

[13] I. El-Naqa, Y. Yang, N.P. Galatsanos, R.M. Nishikawa, and M.N. Wernick, "A similarity learning approach to content-based image retrieval: application to digital mammography," *TMI*, vol. 23, no. 10, pp. 1233–1244, 2004.

[14] G. D. Tourassi, R. Ike, S. Singh, and B. Harrawood, "Evaluating the effect of image preprocessing on an information-theoretic CAD system in mammography," *Academic Radiology*, vol. 15, no. 5, pp. 626–634, 2008.

[15] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *VLDB*, 1999, pp. 518–529.

[16] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *NIPS*, 2008.

[17] W. Liu, J. Wang, and S.F. Chang, "Hashing with graphs," in *ICML*, 2011.

[18] S. Zhang, M. Yang, T. Cour, K. Yu, and D. Metaxas, "Query specific fusion for image retrieval," in *ECCV*, pp. 660–673. 2012.

[19] S. Zhang, J. Huang, H. Li, and D. Metaxas, "Automatic image annotation and retrieval using group sparsity," *TSMC, Part B*, vol. 42, no. 3, pp. 838–849, 2012.

[20] D.G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, 1999, vol. 2, pp. 1150–1157.

[21] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[22] A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *ECCV*, 2008, pp. 316–329.

[23] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W.P. Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the International Workshop on Digital Mammography*, 2001, pp. 212–218.

[24] B. Zheng, A. Lu, A. Hardesty, J.H. Sumkin, C.M. Hakim, M.A. Ganott, and D. Gur, "A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment," *Medical Physics*, vol. 33, no. 1, pp. 111–117, 2006.