COMPUTER-AIDED DIAGNOSIS OF MAMMOGRAPHIC MASSES USING VOCABULARY TREE-BASED IMAGE RETRIEVAL

Menglin Jiang[1]*, Shaoting Zhang*[2]*, Jingjing Liu*[1]*, Tian Shen*[3]*, Dimitris N. Metaxas*[1]

[1]Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[2]Department of Computer Science, UNC Charlotte, Charlotte, NC, USA
[3]Hwatech Medical Info-Tech Co., Xi'An, China

ABSTRACT

Computer-aided diagnosis of masses in mammograms is important to the prevention of breast cancer. Many approaches tackle this problem through content-based image retrieval (CBIR) techniques. However, most of them fall short of scalability in the retrieval stage, and their diagnostic accuracy is therefore restricted. To overcome this drawback, we propose a scalable method for retrieval and diagnosis of mammographic masses. Specifically, for a query mammographic region of interest (ROI), SIFT descriptors are extracted and searched in a vocabulary tree, which stores all the quantized descriptors of previously diagnosed mammographic ROIs. In addition, to fully exert the discriminative power of SIFT descriptors, contextual information in the vocabulary tree is employed to refine the weights of tree nodes. The retrieved ROIs are then used to determine whether the query ROI contains a mass. This method has excellent scalability due to the low spatial-temporal cost of vocabulary tree. Retrieval precision and diagnostic accuracy are evaluated on 5005 ROIs extracted from the digital database for screening mammography (DDSM), which demonstrate the efficacy of our approach.

Index Terms— Mammographic masses, computer-aided diagnosis (CAD), content-based image retrieval (CBIR)

1. INTRODUCTION

For years, breast cancer remains the leading cause of cancer-related death among women. Nevertheless, early diagnosis could improve the chances of recovery dramatically. Currently, among all the imaging techniques for breast examination, mammography is the most effective and the only widely accepted method. Many computer-aided diagnosis (CAD) methods have been proposed to facilitate the detection of masses in mammograms, which is an important indicator of breast cancer. Most of these approaches consist of two steps, namely detection of suspicious regions and classification of these regions as mass or normal tissue [3, 4, 8, 11].

As an alternative solution, some CAD methods utilize content-based image retrieval (CBIR) techniques. Specifically, they compare the current case with previously diagnosed cases stored in a reference database, and return the most relevant cases along with the likelihood of a mass in the current case. Compared with classification-based approaches, these methods could provide more clinical evidence to assist the diagnosis, and therefore attract more and more attention. For example, template matching based on mutual information was utilized to retrieve mammographic regions of interest (ROIs), and similarity scores between the query ROI and its best matches were used to determine whether it contained a mass [13]. This approach was further studied using more similarity measures (such as normalized mutual information) [12]. Features related to shape, edge sharpness and texture were adopted to search for mammographic ROIs with similar masses [1]. For the same purpose, 14 image features and a k-nearest neighbor (k-NN) algorithm were applied in [18]. This method was improved by removing poorly effective ROIs from the reference database [9]. These methods have shown great value of CBIR techniques in retrieval and analysis of mammographic masses. However, they did not consider scalability and were tested on at most 3200 mammographic ROIs. This drawback limited the diagnostic accuracy, since the larger a reference database is, the more likely that relevant cases are found and a correct decision is made [9].

In this paper, we propose to solve the above problem through a scalable image retrieval framework. Specifically, SIFT descriptors extracted from database ROIs are quantized and indexed in a vocabulary tree. To enhance the discriminative power of SIFT descriptors, statistical information about neighbor nodes in the tree is utilized to refine the weights of tree nodes following [15]. Given a query ROI, i.e. a mass region asserted by another CAD system, SIFT descriptors are extracted and searched in the tree to find similar database ROIs. These ROIs along with the similarities to the query ROI are used to determine whether the query contains a mass or not. Such method could retrieve from millions of images efficiently due to its low cost of memory and computational time.

2. PROPOSED APPROACH

In this section, we first introduce the mammographic ROI retrieval framework based on vocabulary tree, then present our

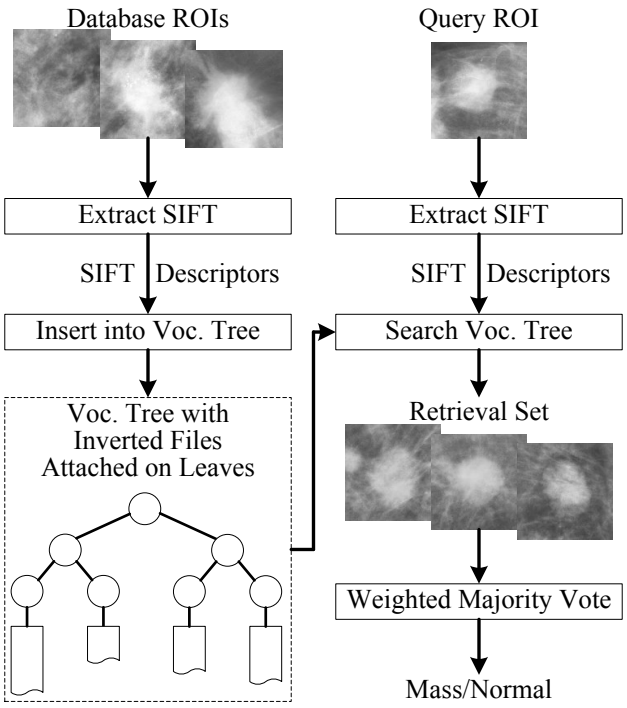

Fig. 1. Overview of the proposed approach.

refinement on the weights of tree nodes, and describe how to make a diagnostic decision using the retrieval set. The overview of our approach is shown in Fig. 1.

Mammogram Retrieval with a Vocabulary Tree: Our approach builds upon a popular CBIR framework that indexes local image features using vocabulary tree and inverted files [10, 7]. The local feature we choose here is scale-invariant feature transform (SIFT) [6]. It has been successfully applied to medical image retrieval and analysis [2], owing to its excellent robustness and discriminability.

In this framework, a large set of SIFT descriptors extracted from a separate database are used to train a vocabulary tree through hierarchical k-means clustering. Specifically, k-means algorithm is first run on the entire training data, defining k clusters and their centers. This process is then recursively applied to all the clusters, splitting each cluster into k sub-clusters. After L recursions, a vocabulary tree of depth L and branch factor k is built. Then, all SIFT descriptors extracted from database mammographic ROIs are quantized and indexed using this vocabulary tree and inverted files. Each SIFT descriptor is propagated down the tree by choosing the closest cluster center at each level. The ID of associated database ROI is then added to the inverted file attached to the leaf node. Given a query mammographic ROI q, SIFT features are extracted and quantized. The similarity score between q and a database ROI d is calculated based on how similar their paths are. The tree nodes are weighted using term frequency-inverse document frequency (TF-IDF) scheme or its variations, where TF means the weight of a node is proportional to its frequency in a query ROI, and IDF indicates that the weight is offset by its frequency in all database ROIs.

Assume q is represented by a set of paths (descriptors) $q = \{P_i^q\}_{i=1}^m$, where m is the number of descriptors. Each path consists of L nodes $P_i^q = \left\{v_{i,\ell}^q\right\}_{\ell=1}^L$, where $v_{i,\ell}^q$ denotes the node on the ℓ-th level. Similarly, d is represented by $d = \left\{P_j^d\right\}_{j=1}^n$, where n is the number of descriptors, and $P_j^d = \left\{v_{j,\ell}^d\right\}_{\ell=1}^L$. The similarity score $s\left(q,d\right)$ between q and d is calculated as the average similarity between all pairs of paths:

$$s\left(q,d\right) = \frac{1}{m \cdot n} \sum\nolimits_{i,j} s_P\left(P_i^q, P_j^d\right) \tag{1}$$

where the normalization factor $1/(m \cdot n)$ is used to achieve fairness between database ROIs with few and many descriptors. The similarity between two paths is defined as the weighted count of their common nodes:

$$s_P\left(P_i^q, P_j^d\right) = \sum\nolimits_{\ell} w\left(v_{i,\ell}^q\right) \cdot \delta\left(v_{i,\ell}^q, v_{j,\ell}^d\right) \tag{2}$$

where w is a weighting function, and δ is the Kronecker delta function, i.e. $\delta\left(a,b\right) = 1$ if $a = b$ and $\delta\left(a,b\right) = 0$ otherwise. In [7], w is defined following the IDF principle using entropy:

$$w\left(v\right) = idf\left(v\right) = \log \frac{N}{N_v} \tag{3}$$

where N is the total number of database ROIs, and N_v is the number of ROIs with at least one path through node v. Note that multiple descriptors in q quantized to the same node v contribute $w\left(v\right)$ multiple times to $s\left(q,d\right)$, which is equivalent to TF.

The above framework allows the use of a very large vocabulary since its computational cost is logarithmic in the number of leaf nodes. As the vocabulary size increases, leaf nodes become smaller and more discriminative. Therefore, the retrieval accuracy is improved. In addition, smaller nodes mean that less descriptors from the database need to be considered during the similarity calculation. Thus, the retrieval speed is accelerated.

Reweighting of Vocabulary Tree Nodes: The IDF scheme calculates a node's weight based on the whole database, ignoring how frequently it occurs in a specific mammogram. However, generally speaking, descriptors with high frequencies in a mammogram are less informative than those with low frequencies. As shown in Fig. 2, a majority of descriptors are extracted from normal tissue around a mass. Although their IDFs are generally smaller than those of the descriptors extracted from the edge of the mass, they still dominate the similarity score due to large TFs. To avoid such over-counting, inspired by descriptor contextual weighting [15], we incorporate the mammogram-specific node frequencies into IDF scheme to down-weight these descriptors.

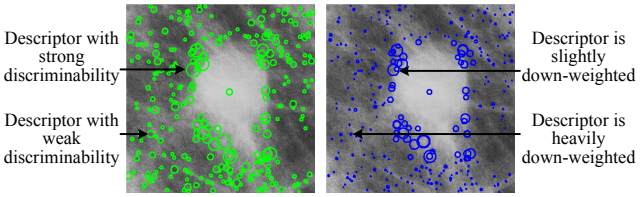

Fig. 2. Effect of reweighting. The left image shows the original IDF weights of each descriptor (only 300 are drawn), and the right image shows the refined new weights. The radius of a circle associated with a descriptor is proportional to its weight.

Suppose the node paths P_i^q of query ROI q and P_j^d of database ROI d have the same node $v \in P_i^q \cap P_j^d = \left\{v_{i,\ell}^q\right\}_{\ell=1}^L \cap \left\{v_{j,\ell}^d\right\}_{\ell=1}^L$, the node's weight $w(v)$ in (3) is modified to:

$$w_{i,j}^{q,d}(v) = w_P\left(P_i^q\right) \cdot w_P\left(P_j^d\right) \cdot idf(v) \tag{4}$$

where the reweighting factors $w_P\left(P_i^q\right)$ and $w_P\left(P_j^d\right)$ are calculated based on the frequencies of nodes along paths P_i^q and P_j^d respectively. Specifically, let $n\left(q, v_{i,\ell}^q\right)$ be the number of paths of q that pass through node $v_{i,\ell}^q$, $w_P\left(P_i^q\right)$ is defined as:

$$w_P\left(P_i^q\right) = \sqrt{\frac{\sum_\ell w\left(v_{i,\ell}^q\right)}{\sum_\ell w\left(v_{i,\ell}^q\right) \cdot n\left(q, v_{i,\ell}^q\right)}} \tag{5}$$

where $w\left(v_{i,\ell}^q\right)$ is a weighting coefficient, usually set to $idf\left(v_{i,\ell}^q\right)$ empirically. The square root is due to the weighting of both $w_P\left(P_i^q\right)$ and $w_P\left(P_j^d\right)$.

Note that $w_P\left(P_i^q\right)$ is shared for all nodes $v_{i,\ell}^q$ along path P_i^q. In order to determine the importance of a descriptor P_i^q, $w_P\left(P_i^q\right)$ takes into account the descriptors in q quantized to neighbor tree leaves since they also contribute to $n\left(q, v_{i,\ell}^q\right)$. Consequently, nodes in a subtree with more descriptors are heavily down-weighted. The effect of reweighting is illustrated in Fig. 2.

Diagnosis of Mammographic Masses: After the retrieval stage, a query mammographic ROI is classified according to its best matched database ROIs using majority logic. Formally speaking, let $\{d_i\}_{i=1}^K$ denote the top K similar database ROIs for q, each d_i has a class tag $c(d_i) \in \{\oplus, \ominus\}$, with the label $\oplus$ for mass and $\ominus$ for normal tissue. q is classified by a weighted majority vote of $\{d_i\}_{i=1}^K$, where the contribution of d_i is weighted by its similarity to q:

$$c(q) = \arg\max_c \sum\nolimits_i s(q, d_i) \cdot \delta(c, c(d_i)) \tag{6}$$

Table 1. Retrieval precision of vocabulary tree (Voc) and our method (Voc + Reweighting) at different K.

K	Method	Mass	Normal	Total
1	Voc	76.0%	82.0%	79.0%
	Voc+Reweighting	**83.0%**	**86.0%**	**84.5%**
5	Voc	77.6%	80.6%	79.1%
	Voc+Reweighting	83.4%	84.8%	84.1%
10	Voc	74.5%	78.7%	76.6%
	Voc+Reweighting	79.9%	82.1%	81.0%

3. EXPERIMENTS

Our experiment dataset was constructed from the digital database for screening mammography (DDSM) [5] following the conventions in [9, 12, 18]. First, all the mammograms acquired on different scanners were normalized according to DDSM's instructions. Second, 2274 mammographic ROIs centered on a mass (1279 malignant and 995 benign) annotated by experienced radiologists were extracted. Third, 2931 false positives generated by a CAD system were used as normal regions. This CAD system is based on a cascade of boosted Haar classifiers [14] and trained on a separate mammogram dataset. Finally, of the above ROIs, 2174 mass ROIs and 2831 normal ROIs, 5005 ROIs in total, were used to construct a reference database. The remaining 200 ROIs, with a half for masses and the other for normal regions, were used as queries. Note that compared with experiments which randomly select normal regions [13], our experiment setting is more similar to realistic situations and more challenging. It is also worth pointing out that our approach can retrieve in real-time from a database of 1 million images, which has been substantiated on general CBIR datasets [7, 15].

We first evaluate the retrieval performance of the proposed approach. A system employing a vocabulary tree without reweighting is used as the baseline approach. The evaluation measure used here is *retrieval precision*, which is defined as the percentage of retrieved database ROIs that are relevant to query ROI. Overall the precision changes slightly as the number of retrievals increases from K=1 to K=10. The precisions achieved at top K=1, 5, and 10 retrievals are summarized in Table 1. Our method achieves higher precision than that of the baseline system. Detailed results show that many incorrect retrievals are due to the visual similarity between malignant masses and ROIs with a bright core and spiculated edge. For example, as shown in Fig. 3, a ROI depicting normal tissue is incorrectly retrieved for a mass ROI. A possible solution is to conduct a more reliable (but less efficient) re-matching between query ROI and its retrieved database ROIs, e.g. using spatial contextual information of local features, and remove irrelevant database ROIs. Another possible improvement is to re-rank the retrieval set according to associated diagnostic

Table 2. Classification accuracy of vocabulary tree (Voc) and our method (Voc + Reweighting) at different K.

K	Method	Mass	Normal	Total
1	Voc	76.0%	82.0%	79.0%
	Voc+Reweighting	83.0%	86.0%	84.5%
5	Voc	81.0%	85.0%	83.0%
	Voc+Reweighting	**86.0%**	**88.0%**	**87.0%**
10	Voc	78.0%	81.0%	79.5%
	Voc+Reweighting	85.0%	86.0%	85.5%

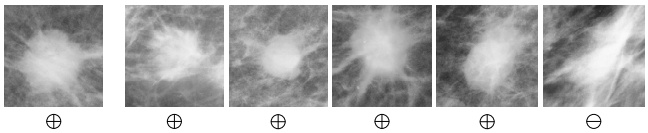

Fig. 3. An example of a query mass ROI (left) and its top 5 best-matched database ROIs. For each ROI, its class is shown. The query ROI is correctly classified as mass according to a weighted majority vote of the 5 database ROIs.

information, such as the patient's age and breast tissue density rating. These textual features can be combined with SIFT features using feature selection and fusion methods [16, 17].

The diagnostic performance is measured using *classification accuracy*, which refers to the percentage of query ROIs that are correctly classified. The classification accuracy achieved by two methods at top K=1, 5, and 10 retrievals is summarized in Table 2. Once again, our method consistently outperforms the baseline system. In addition, the classification accuracy is even better than the retrieval precision, since irrelevant retrievals would not cause a misclassification as long as they remain a minority of the retrieval set. Especially, a classification accuracy as high as 87.0% is obtained at K=5, which is pretty satisfactory.

4. CONCLUSION

In this paper, we propose to use scalable CBIR for the automatic diagnosis of mammographic masses. To retrieve efficiently from a large database, which leads to better retrieval precision and diagnostic accuracy, we employ the vocabulary tree framework to hierarchically quantize and index SIFT descriptors. Furthermore, contextual information in the vocabulary tree is incorporated into TF-IDF weighting scheme to improve the discriminative power of tree nodes. Query mammographic ROIs are classified using a weighted majority vote of its best matched database ROIs. Experiments are conducted on a database including 2174 mass ROIs and 2831 CAD-generated false positive ROIs, which is the largest dataset to the best of our knowledge. Excellent results demonstrate the retrieval precision and diagnostic accuracy of our method. Future endeavors will be devoted to refine retrieval set using spatial contextual information of SIFT features. Diagnostic information can also be taken into consideration using feature selection and fusion methods [16, 17].

5. REFERENCES

[1] H. Alto, R. M. Rangayyan, and J. E. L. Desautels. Content-based retrieval and analysis of mammographic masses. *J. Electron. Imaging*, 14(2):023016–1–023016–17, 2005.

[2] J. C. Caicedo, A. Cruz, and F. A. Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Proc. AIME*, pages 126–135, 2009.

[3] 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. *Pattern Recognit.*, 39(4):646–668, 2006.

[4] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K.-H. Ng. Computer-aided breast cancer detection using mammograms: A review. *IEEE Rev. Biomed. Eng.*, 6:77–98, 2013.

[5] M. Heath, K. Bowyer, D. Kopans, W. P. Kegelmeyer Jr, R. Moore, K. Chang, and S. Munishkumaran. Current status of the digital database for screening mammography. In *Digital Mammography*, pages 457–460. Springer Netherlands, 1998.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[7] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. IEEE CVPR*, pages 2161–2168, 2006.

[8] 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. *Med. Image Anal.*, 14(2):87–110, 2010.

[9] S. C. Park, R. Sukthankar, L. Mummert, M. Satyanarayanan, and B. Zheng. Optimization of reference library used in content-based medical image retrieval scheme. *Med. Phys.*, 34(11):4331–4339, 2007.

[10] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. IEEE ICCV*, pages 1470–1477, 2003.

[11] 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. Inf. Technol. Biomed.*, 13(2):236–251, 2009.

[12] G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd. Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms. *Med. Phys.*, 34(1):140–150, 2007.

[13] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious Jr, and C. E. Floyd Jr. Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. *Med. Phys.*, 30(8):2123–2130, 2003.

[14] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE CVPR*, pages I–511–I–518, 2001.

[15] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han. Contextual weighting for vocabulary tree based image retrieval. In *Proc. IEEE ICCV*, pages 209–216, 2011.

[16] S. Zhang, J. Huang, H. Li, and D. N. Metaxas. Automatic image annotation and retrieval using group sparsity. *IEEE Trans. Syst. Man Cybern. B Cybern.*, 42(3):838–849, 2012.

[17] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas. Query specific fusion for image retrieval. In *Proc. ECCV*, pages 660–673, 2012.

[18] B. Zheng, A. Lu, L. 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. *Med. Phys.*, 33(1):111–117, 2006.