BONE TEXTURE CHARACTERIZATION WITH FISHER ENCODING OF LOCAL DESCRIPTORS

Yang Song[1]*, Weidong Cai*[1]*, Fan Zhang*[1]*, Heng Huang*[2]*, Yun Zhou*[3]*, David Dagan Feng*[1]

[1]BMIT Research Group, School of IT, University of Sydney, Australia
[2]Department of Computer Science and Engineering, University of Texas, Arlington, USA
[3]Russell H. Morgan Department of Radiology and Radiological Science,
Johns Hopkins University School of Medicine, USA

ABSTRACT

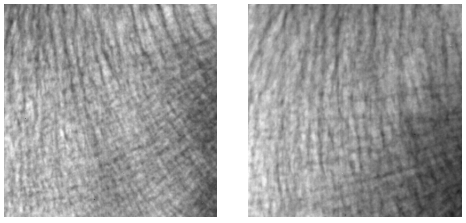
Bone texture characterization is important for differentiating osteoporotic and healthy subjects. Automated classification is however very challenging due to the high degree of visual similarity between the two types of images. In this paper, we propose to describe the bone textures by extracting dense sets of local descriptors and encoding them with the improved Fisher vector (IFV). Compared to the standard bag-of-visual-words (BoW) model, Fisher encoding is more discriminative by representing the distribution of local descriptors in addition to the occurrence frequencies. Our method is evaluated on the ISBI 2014 challenge dataset of bone texture characterization, and we demonstrate excellent classification performance compared to the challenge entries and large improvement over the BoW model.

Index Terms— Bone texture, classification, feature encoding, Fisher vector

1. INTRODUCTION

Osteoporosis, which is a bone disease, causes bone fragility and increases the risk of fracture. It is widely recognized that the analysis of microarchitectural alterations would lead to better diagnosis of this disease [1]. Recently 2D texture analysis using X-ray imaging has shown its potential in providing a cost-effective and efficient way to detect and evaluate osteoporosis [2]. Realizing automated bone texture characterization is however very challenging, since the healthy and osteoporotic images exhibit very similar visual patterns (examples shown in Fig. 1). The current research thus focuses on designing and evaluating various feature descriptors for characterizing the bone textures [3, 1, 4, 5].

In our study, we are interested in understanding how the popular bag-of-visual-words (BoW) model works for bone texture characterization. BoW basically quantizes the densely extracted local descriptors into a set of visual words using k-means clustering, and computes a histogram descriptor for an image by counting the occurrence frequencies of visual words. It has been successfully applied to many medical imaging applications [6, 7, 8, 9]. We observe that at the finer scales, there are large inhomogeneities in the bone textures. This thus prompts us to consider dividing the images into small patches, extracting the local descriptors, and combining/encoding them into an overall descriptor. The BoW model conveniently fits this purpose.



Healthy Osteoporotic

Fig. 1. Examples images.

To achieve more accurate classification, in general computer vision, more advanced encoding schemes have recently been incorporated into the BoW model [10]. In particular, the improved Fisher vector (IFV) [11] has been proposed to quantize the local descriptors based on the Gaussian mixture model (GMM) rather than k-means clustering. The advantage of IFV is that it encodes information about the distribution of local descriptors in addition to the occurrence frequencies, and it is designed to work with linear-kernel support vector machine (SVM) to facilitate efficient computation. It has been highly effective for various classification problems on a number of challenging datasets in general computer vision [10, 12]. Despite these nice properties, IFV has rarely been explored for medical imaging. To the best of our knowledge, only a few studies investigated the earlier version of IFV [13, 14, 15].

We thus incorporated the IFV encoding into the BoW model for bone texture characterization. While the original IFV is developed with the scale-invariant feature transform

This work was supported in part by ARC grants.

(SIFT) [16] descriptor, we have extended it to the local binary patterns (LBP) [17] as well. This extension is motivated by our previous finding that LBP provides highly representative texture description [18, 19]. Our method is used to classify healthy and osteoporotic images, and evaluated on the ISBI 2014 challenge dataset on Texture Characterization of Bone radiograph images (TCB) [20]. We have observed large performance improvement with the IFV encoding, compared to the standard BoW model.

2. METHODS

2.1. The BoW Model

Based on a set of local descriptors extracted from an image, the BoW model represents the occurrence frequencies of visual words in the image. Formally, given a set of N images each with M local descriptors extracted, we denote the entire set of local descriptors as $\{x_i : i = 1, ..., MN\}$, where $x_i \in \mathbb{R}^D$ and D is the feature dimension. A visual vocabulary $\{\mu_k : k = 1, ..., K\}$ of K visual words is trained by quantizing these local descriptors into K bins using k-means clustering. For a certain image I, each of its local descriptor x_i is assigned to its nearest visual word μ_k. The BoW representation of the image I is then the histogram $f \in \mathbb{R}^K$ with each element $f(k)$ containing the number of local descriptors assigned to the visual word μ_k.

We used the BoW model as our baseline for characterizing bone textures. Two types of local descriptors are extracted: SIFT and LBP. These are chosen due to their popularity in both medical imaging and general computer vision. For SIFT, we extract the 128-dimensional local descriptors densely at multiple scales with spatial bins of 4, 6, 8 and 10 pixels, which are sampled every three pixels. The local descriptors are then reduced to 64-dimensional using the principal component analysis (PCA). For LBP, a local feature is computed for every 3×3 pixel neighborhood and aggregated as a 58-dimensional local descriptor for every non-overlapping 80×80 pixel cell. The VLFeat toolbox [21] is used to extract the dense SIFT and LBP features. Apart from the above mentioned parameters, the default settings are applied. Based on the training images, visual vocabularies are then constructed separately for the SIFT and LBP features using k-means clustering. An image I is then represented by two BoW descriptors f_{SIFT} and f_{LBP}, based on the SIFT and LBP local descriptors and learned visual vocabularies.

2.2. Fisher Encoding

With IFV, the visual vocabulary is generated using GMM with K components, with visual words represented by the mean vector $\mu_k \in \mathbb{R}^D$, covariance matrix $\Sigma_k \in \mathbb{R}^{D \times D}$ and the prior probability ω_k of each Gaussian component. The expectation maximization (EM) algorithm is used to train these parameters based on the training set of N images. The covariance matrix Σ_k is assumed to be diagonal, and hence can be represented by the covariance vector $\sigma_k^2 \in \mathbb{R}^D$. For a local descriptor x_i, soft assignments to the visual words are computed by:

$$q_{ki} = \frac{p(x_i|\mu_k, \sigma_k^2)\omega_k}{\sum_{j=1}^{K} p(x_i|\mu_j, \sigma_j^2)\omega_j}, \quad \forall k = 1, ..., K. \tag{1}$$

Then rather than summarizing the soft assignments into a histogram similar to the standard BoW, the average first and second order differences, $u_k \in \mathbb{R}^D$ and $v_k \in \mathbb{R}^D$, between the M local descriptors from image I and the visual words are computed:

$$u_k = \frac{1}{M\sqrt{\omega_k}} \sum_{i=1}^{M} q_{ki} \frac{x_i - \mu_k}{\sigma_k}, \tag{2}$$

$$v_k = \frac{1}{M\sqrt{2\omega_k}} \sum_{i=1}^{M} q_{ki} \left[\frac{(x_i - \mu_k)^2}{\sigma_k^2} - 1 \right]. \tag{3}$$

The Fisher encoding of the image I is then the concatenation of all u_k and v_k for all K visual words:

$$f = [u_1^T, v_1^T, ..., u_K^T, v_K^T]^T. \tag{4}$$

The feature dimension is thus $2KD$. Finally, f is power and L2 normalized to improve its discriminative capability. The power normalization is performed by transforming each feature dimension $d \in \{1, ..., 2KD\}$ with the following:

$$f(d) = sign(f(d))|f(d)|^{0.5} \tag{5}$$

This obtained IFV descriptor is denoted as f_{IFV}.

Different from the original IFV, we didn't incorporate the spatial pyramid matching (SPM) [22] pooling. This is because spatial information is less important for textured images compared to natural scenes, and SPM pooling would increase the feature dimension by a factor of the number of regions. The local descriptors (SIFT and LBP) are extracted in the same way as mentioned in the previous section, and two sets of IFV descriptors are generated separately with SIFT and LBP as the local descriptors. The feature encoding toolbox [10] is modified to implement our feature extraction and encoding methods.

2.3. Classification

We used the dataset with released ground truth from the ISBI 2014 TCB challenge [20], which comprises 58 healthy and 58 osteoporotic X-ray images. Each image is of 400×400 pixels showing the region of bone textures only. Based on the f_{IFV} descriptors, linear-kernel binary SVM classifiers (using LIBSVM [23]) were applied to categorize the images with three-fold cross validation. Specifically, images of each class

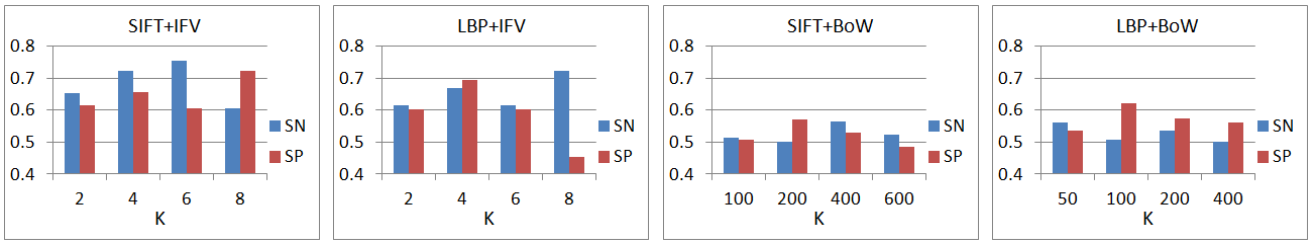


Fig. 2. Sensitivity (SN) and specificity (SP) of bone texture characterization, comparing the various approaches (SIFT with IFV, LBP with IFV, SIFT with BoW, and LBP with BoW), with different settings of K.

were divided sequentially into three groups of 20, 20, and 18 images, respectively. Training and testing were then conducted in three splits. During each split, two groups were used for training and one for testing. The regularization parameter C was set with training and validation on the training set, and $C = 2.6$ produced the highest classification accuracy. Note that IFV was originally designed to work with linear-kernel SVM, and we also found that the linear kernel indeed provided better classification accuracy compared to the other well-known nonlinear kernels.

2.4. Evaluation Metrics

We followed the requirement of the TCB challenge that sensitivity (SN) and specificity (SP) were used to evaluate the characterization performance, and were derived based on the hard binary classification results of SVM. Osteoporosis is the positive class and a correctly identified osteoporotic image is a true positive (TP). A true negative (TN) is a correctly identified healthy image. A false positive (FP) or false negative (FN) is an image misclassified as the osteoporotic or healthy case. Then, SP = TP/(TP+FN), and SN = TN/(TN+FP). The receiver operating characteristic (ROC) curve was also plotted by varying the classification threshold based on the probability estimates of the classification outputs. The area under the curve (AUC) was computed to quantify the characterization performance. We compared our SIFT/LBP descriptor encoding with IFV against the standard BoW model with SIFT/LBP local descriptors. We also compared with the winning entry [5] of the TCB challenge.

3. RESULTS

As shown in Fig. 2, the parameter K, which is the number of visual words, is important for the characterization results. For BoW, the descriptor dimension is the same as K, while for IFV, the descriptor dimension is $2KD$ with D denoting the dimension of the local descriptors (after dimension reduction). We thus prefer a small K for IFV and a large K for BoW to keep a similar descriptor dimension between the two encoding methods. The results show that with $K = 4$, we

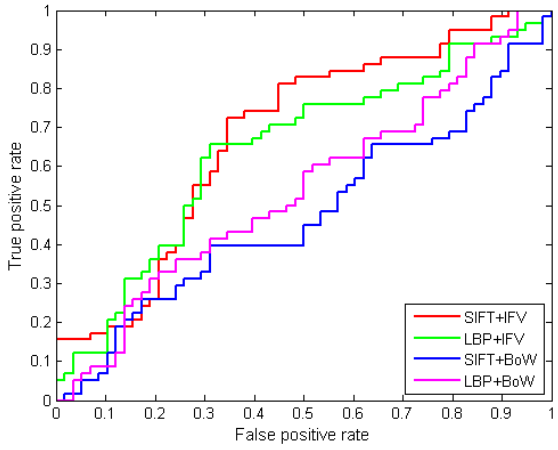


Fig. 3. The ROC curves of classification results, comparing SIFT with IFV, LBP with IFV, SIFT with BoW, and LBP with BoW, all with the best K settings.

obtained the highest SN and SP for SIFT/LBP with IFV; and for BoW, the highest results were obtained when $K = 400$ with SIFT and $K = 200$ with LBP.

Overall, the IFV encoding achieved about 19% and 13% improvement in SN and SP over the BoW model when SIFT was used as the local descriptor, with the optimal settings of K for both methods. With LBP, the improvement was about 13% and 12% in SN and SP. The advantage of IFV over BoW is further demonstrated by the ROC curves as shown in Fig. 3. Larger AUCs were derived with the IFV encoding, i.e. 0.68 vs. 0.48 for IFV and BoW with SIFT, and 0.64 vs. 0.54 for IFV and BoW with LBP. The benefit of IFV is more obvious when coupled with SIFT than LBP, suggesting IFV is more effective with more descriptiveness features like SIFT.

SIFT with IFV and LBP with IFV produced comparable results, with SIFT providing a higher SN while LBP providing a higher SP (at $K = 4$). By analyzing the ROC curves of the two descriptors (Fig. 3), we can see that SIFT achieved slightly higher overall performance than LBP, with the AUCs of 0.68 vs. 0.64. The computation with LBP was however

Table 1. The classification results compared to [5].

	SIFT+IFV	LBP+IFV	[5] Haar
SN	0.72	0.67	0.62
SP	0.66	0.70	0.66

much faster than using SIFT. The time taken for training the visual vocabulary and descriptor generation was on average 0.05 s per image for LBP with IFV, while that for SIFT with IFV was on average 4.3 s. In addition, the IFV encoding was more computationally efficient than the standard k-means used in BoW. For example, with SIFT as the local descriptors, the BoW approach took on average 11.3 s per image, which was about 7 s more than the IFV approach.

Table 1 shows the comparison with the winning entry [5] of the TCB challenge, which computes the marginals of Haar wavelet decompositions as the feature descriptor. Our SIFT/LBP with IFV methods achieved higher classification SN and SP over the compared approach. Note that both our results and the results from [5] listed in the table were obtained from the initial dataset with released ground truth. The approach [5] achieved consistent results on the blind dataset of 58 additional images. Among the 15 competition entries of the TCB challenge, some entries reported higher performance than [5] on the initial dataset but lower performance on the blind dataset. Our comparison with [5] is thus inconclusive at the present stage, and we will investigate our method performance with the blind dataset in future work.

4. CONCLUSION

In this paper, we present a method for characterizing bone textures in X-ray images. We describe the image features with an enhanced bag-of-visual-words (BoW) model, by encoding SIFT and LBP local descriptors with the improved Fisher vector (IFV). Linear-kernel SVM is then used to classify the image descriptors into healthy and osteoporotic cases. Our method has been evaluated on the ISBI 2014 TCB challenge dataset of 58 healthy and 58 osteoporotic images. Promising performance has been demonstrated with on average 16% and 13% improvement in classification sensitivity and specificity over the standard BoW model.

5. REFERENCES

[1] L. Houam, A. Hafiane, A. Boukrouche, E. Lespessailles, and R. Jennane, "One dimensional local binary pattern for bone texture characterization," *Pattern Anal. Applic.*, vol. 17, no. 1, pp. 179–193, 2014.

[2] E. Lespessailles, C. Gadois, I. Kousignian, J. P. Neveu, P. Fardellone, S. Kolta, C. Roux, J. P. Do-Huu, and C. M. Benhamou, "Clinical interest of bone texture analysis in osteoporosis: a case control multicenter study," *Ost. Int.*, vol. 19, no. 7, pp. 1019–1028, 2008.

[3] A. S. El Hassani, M. El Hassouni, L. Houam, M. Rziza, E. Lespessailles, and R. Jennane, "Texture analysis using dual tree m-band and renyi entropy, application to osteoporosis diagnosis on bone radiographs," *ISBI*, pp. 1487–1490, 2012.

[4] R. Jennane, J. Touvier, M. Bergounioux, and E. Lespessailles, "A variational model for trabecular bone radiograph characterization," *ISBI*, pp. 1283–1286, 2014.

[5] F. Yger, "Challenge IEEE-ISBI/TCB: application of covariance matrices and wavelet marginals," *arXiv preprint arXiv:1410.2663*, 2014.

[6] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, "X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words," *IEEE Trans. Med. Imaging*, vol. 30, no. 3, pp. 733–746, 2011.

[7] Y. Song, W. Cai, Y. Zhou, L. Wen, and D. Feng, "Pathology-centric medical image retrieval with hierarchical contextual spatial descriptor," *ISBI*, pp. 202–205, 2013.

[8] X. Zhang, L. Yang, W. Liu, H. Su, and S. Zhang, "Mining histopathological images via composite hashing and online learning," *MICCAI*, pp. 479–486, 2014.

[9] F. Zhang, Y. Song, S. Liu, S. Pujol, R. Kikinis, D. Feng, and W. Cai, "Latent semantic association for medical image retrieval," *DICTA*, pp. 50–55, 2014.

[10] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," *BMVC*, pp. 1–12, 2011.

[11] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *ECCV*, pp. 143–156, 2010.

[12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," *CVPR*, pp. 3606–3613, 2014.

[13] S. Manivannan, R. Wang, and E. Trucco, "Inter-cluster features for medical image classification," *MICCAI*, pp. 345–352, 2014.

[14] A. P. Twinanda, M. De Mathelin, and N. Padoy, "Fisher kernel based task boundary retrieval in laparoscopic database with single video query," *MICCAI*, pp. 409–416, 2014.

[15] R. Kwitt, S. Hegenbart, N. Rasiwasia, A. Vecsei, and A. Uhl, "Do we need annotation experts? a case study in celiac disease classification," *MICCAI*, pp. 454–461, 2014.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[18] Y. Song, W. Cai, Y. Zhou, and D. Feng, "Feature-based image patch approximation for lung tissue classification," *IEEE Trans. Med. Imaging*, vol. 32, no. 4, pp. 797–808, 2013.

[19] Y. Song, W. Cai, D. Feng, and M. Chen, "Cell nuclei segmentation in fluorescence microscopy images using inter- and intra-region discriminative information," *EMBC*, pp. 6087–6090, 2013.

[20] "Challenge IEEE-ISBI: bone texture characterization," http://www.univ-orleans.fr/i3mto/challenge-ieee-isbi-bone-texture-characterization.

[21] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," *ACM Int. Conf. Multimedia*, pp. 1469–1472, 2010.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *CVPR*, pp. 2169–2178, 2006.

[23] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.