Improved patch based automated liver lesion classification by separate analysis of the interior and boundary regions

Idit Diamant[1]*, Assaf Hoogi[2]*, Christopher F. Beaulieu[3], Mustafa Safdari[2], Eyal Klang[4],

Michal Amitai[4], Hayit Greenspan[1]**, Daniel L. Rubin[2]**

***Abstract*—The bag-of-visual-words (BoVW) method with construction of a single dictionary of visual words has been used previously for a variety of classification tasks in medical imaging, including the diagnosis of liver lesions. In this paper, we describe a novel method for automated diagnosis of liver lesions in portal-phase computed tomography (CT) images that improves over single-dictionary BoVW methods by using an image patch representation of the interior and boundary regions of the lesions. Our approach captures characteristics of the lesion margin and of the lesion interior by creating two separate dictionaries for the margin and the interior regions of lesions ("dual dictionaries" of visual words). Based on these dictionaries, visual word histograms are generated for each region of interest (ROI) within the lesion and its margin. For validation of our approach, we used two datasets from two different institutions, containing CT images of 194 liver lesions (61 cysts, 80 metastasis and 53 hemangiomas). The final diagnosis of each lesion was established by radiologists. The classification accuracy for the images from the two institutions was 99% and 88%, respectively, and 93% for a combined dataset. Our new BoVW approach that uses dual dictionaries shows promising results. We believe the benefits of our approach may generalize to other application domains within radiology.**

***Index Terms*— Automated diagnosis, computed tomography, classification, focal liver lesions, image patch analysis, visual words.**

I. Introduction

CANCER is one of the major causes of death, and according to data from WHO (World Health Organization) in 2012, it accounts for 8.2 million deaths worldwide [1]. The liver is one of the three most common sites of metastases, while the other two are bone and lung [2]. Focal liver lesions are a common medical problem, and determining whether a liver lesion is a cancer as opposed to a benign lesion can be very challenging. Early diagnosis and treatment is the most useful way to reduce cancer deaths.

Computed tomography (CT) images are widely used by clinicians for detection, diagnosis, and monitoring of liver lesions [3]. The imaging characteristics of liver lesions are used to classify them as being benign (such as hepatic cysts and hemangiomas) or malignant (such as metastases and hepatocellular carcinoma). Diagnosis of focal liver lesions can be a challenging task due to variability in their appearance (e.g., differences in lesion shape, size, margin sharpness and internal texture). Moreover, the imaging appearance of different lesions overlaps, leading to considerable inter-reader variation. Thus, there is an interest in and need for automated diagnostic tools to assist radiologists in evaluating liver lesions.

A number of researchers have developed methods for automated detection, segmentation or diagnosis of focal liver lesions (FLL) in CT [4-18]. Additional modalities can be used for FLL analysis, such as contrast-enhanced Ultrasound (e.g. [19-21]). A complete overview of all FLL analysis techniques and modalities is beyond the scope of this paper. Hereon we focus on CT related works: In Gletsos et al [4], texture features are used to classify lesions as representing normal liver parenchyma, cyst, hemangiomas and hepatocellular carcinomas. A hierarchical classifier is used with a separate neural network (NN) at each level. The best results were obtained when a feature selection technique was used. Quatrehomme et al [5] classified five types of liver lesions using multi-phase CT images. They performed classification using visual features, such as histogram statistics and law measures and using the Support vector machine (SVM) classifier. They concluded that there is a significant improvement using multi-phase images compared to single-phase. In Bilello et al [15], a lesion detection step was performed, followed by lesion classification. A Gaussian filter was used to obtain the weighted average of various parameters. The features used were the weighted average,

Manuscript received March 14, 2015; accepted September 3, 2015. This work was supported in part by a grant from the National Cancer Institute, National Institutes of Health, grant No. U01-CA-142555 and U01-CA-190214. * indicates equal contributors. ** indicates equal contributors and corresponding authors.

I. Diamant and H. Greenspan are with Department of Biomedical Engineering, Tel Aviv University, Tel Aviv (correspondence e-mail: hayit@eng.tau.ac.il).

A. Hoogi, M. Safdari and D. L. Rubin are with Departments of Radiology and Medicine (Biomedical Informatics Research), Stanford University, Stanford, CA, USA (correspondence e-mail: dlrubin@stanford.edu).

C. F. Beaulieu is with Department of Radiology and, by courtesy, Orthopedic Surgery, Stanford University, Stanford, CA, USA.

E. Klang and M. Amitai are with Department of Diagnostic Imaging, The Chaim Sheba Medical Center, Tel-Hashomer, Israel.

minimum, standard deviation of the intensity values and frequency filters. Using SVM for classification showed high accuracy in separating the hemangiomas and cysts, whereas separating hemangiomas from metastasis and separating metastasis from cysts was more challenging.

The machine vision community has recently developed BoVW models as an approach to obviate the need to pre-specify an exhaustive list of image features. This method, adapted from the text retrieval domain [22] to the visual analysis domain, has been successfully applied for different classification tasks [23-27]. BoVW methods have been shown to recognize objects in images with a variety of appearances due to rotation, scaling and illumination change. BoVW methods have been shown to be successful in medical classification tasks, such as labelling X-ray images on the organ and pathology levels [25], retrieval of similar-appearing liver lesions [6] and classification of breast tissue in mammograms [26-27]. BoVW methods have also been shown to enable automated diagnosis of liver lesions, circumventing the challenge of their varying appearance [6]. Though the results are encouraging, classification accuracy was at best 91% (using single-phase images), leaving room for improvement in the approach.

We present an adaptation of the BoVW method for the automatic classification and diagnosis of liver lesions in portal-phase CT images - that improves prior results. The proposed methodology focuses on characterizing the interior lesion region and the lesion margin region separately. The relative importance of each lesion component to the characterization of liver lesions is explored. The rest of the paper is organized as follows: The proposed method is described in detail in Section 2, Section 3 presents results and a discussion concludes the paper in Section 4.

II. Methods

A. Data

Two datasets are used in this work. One dataset, from (institution name withheld), “Dataset 1”, contains cases selected by searching the medical record for typical cases of hemangioma, cysts, and metastases. Cases were collected with approval of the institution’s Institutional Review Board and de-identified for compliance with the Health Insurance Portability and Accountability Act (HIPAA). Cases were acquired between 2007 and 2008 using either General Electric (GE) Healthcare or Siemens Medical Systems scanners with the following parameters: 120 kVp, 140–400 mAs, and 2.5–5.0-mm section thickness. The second dataset, from (institution name withheld), “Dataset 2”, was constructed as another collection of cases of hemangioma, cysts, and metastases. Cases were acquired between 2011 and 2013 using GE Medical Systems scanner with the following parameters: 120 kVp, 140–400 mAs, and 2.5–5.0-mm section thickness. Cases were collected with approval of the institution’s Institutional Review Board.

Dataset 1 (Fig. 1) contains portal-phase CT images of 109 liver lesions: 39 cysts, 46 metastases, 24 hemangiomas. Dataset 2 (Fig. 2) contains portal-phase CT images of 85 liver lesions: 22 cysts, 34 metastases and 29 hemangiomas. In both image sets, a radiologist circumscribed each lesion margin and provided its diagnosis which was established either by biopsy or clinical follow-up (ground truth). Lesion diameters are between 10-130 mm. Figs. 1 and 2 show examples of marked lesions from each dataset. Cysts are non-enhancing water-attenuation circumscribed lesions. Hemangiomas show typical features of discontinuous nodular peripheral enhancement, with fill-in on delayed images. Metastases are hypo-attenuating, show enhancement with contrast material administration, and have less well-defined margins than cysts [28].

Due to institutional differences in imaging equipment and scanning techniques, the appearance of liver lesions is different in the two datasets. Two expert radiologists who reviewed both datasets noted that Dataset 2 contains larger intra-category variability and lower inter-category variability. The lesions in Dataset 2 can be seen to contain larger variability than the lesions in Dataset 1. For example, some of the metastases in Dataset 2 contain areas of higher density, possibly calcifications or prominent blood vessels. Such characteristics can make them look more like hemangiomas. The difference between the datasets could be due to the fact that the data is collected from two different institutions, with Dataset 1 including a retrospective choice of more typical lesions (with a typical appearance), whereas Dataset 2 was derived prospectively on sequential cases.

The input ROI to our algorithm is the lesion area with surrounding parenchyma liver tissue (see examples in Fig. 3). The lesion ROI was cropped using the radiologist markings, and adding liver parenchyma tissue using a morphological dilation operation (six pixels distance from lesion boundaries).

B. Extracting Local Descriptors

The first step of the BOW approach is feature extraction. For this process, we first extract fixed size patches from each lesion. The patch size needs to be larger than a few pixels across, in order to capture higher-level semantics such as edges or corners. At the same time, the patch size should not be too large if it is aimed to serve as a common building block for many images. Once patches are defined, each patch is represented with a set of feature descriptors. Raw image data, normalized raw data, or other descriptors can be used. The intensity of liver lesions in CT images is important for diagnosis. Thus, in our implementation we use the raw image data without normalization as the patch descriptor. PCA was applied in order to reduce features dimensionality [29].

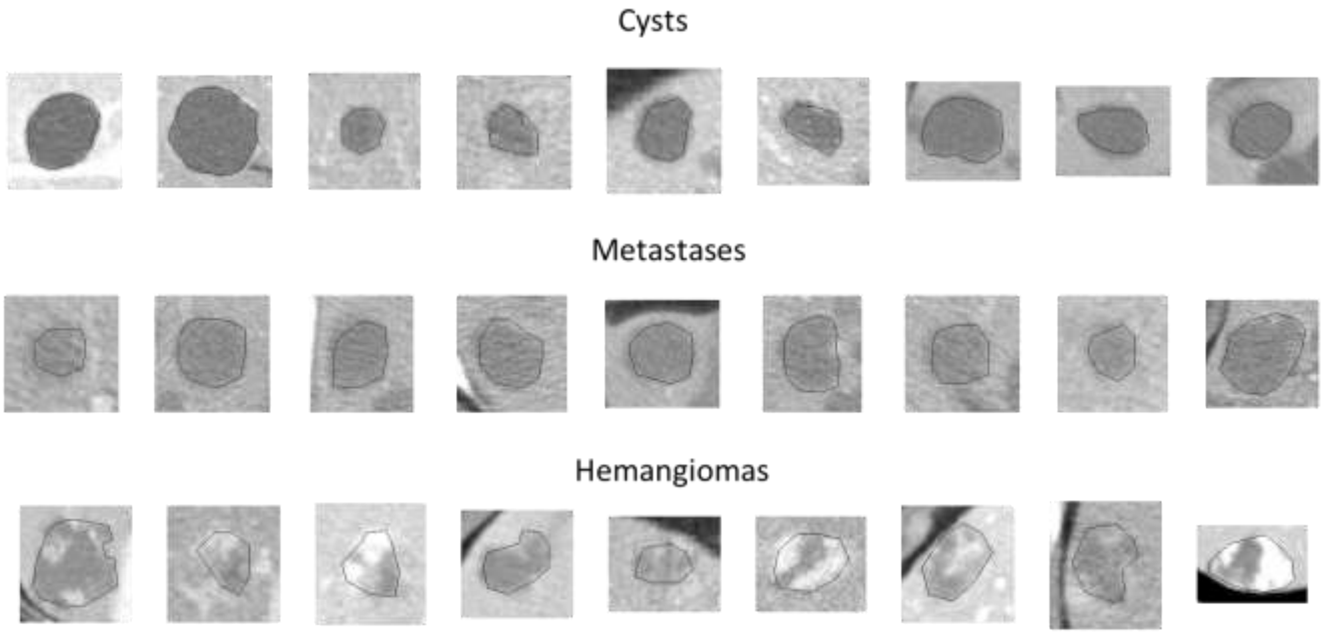


Fig. 1. Dataset 1: Typical examples of Cysts, Hemangiomas and Metastases. For each lesion, the manual segmentation that was done by an expert clinician (C.F.B) is shown

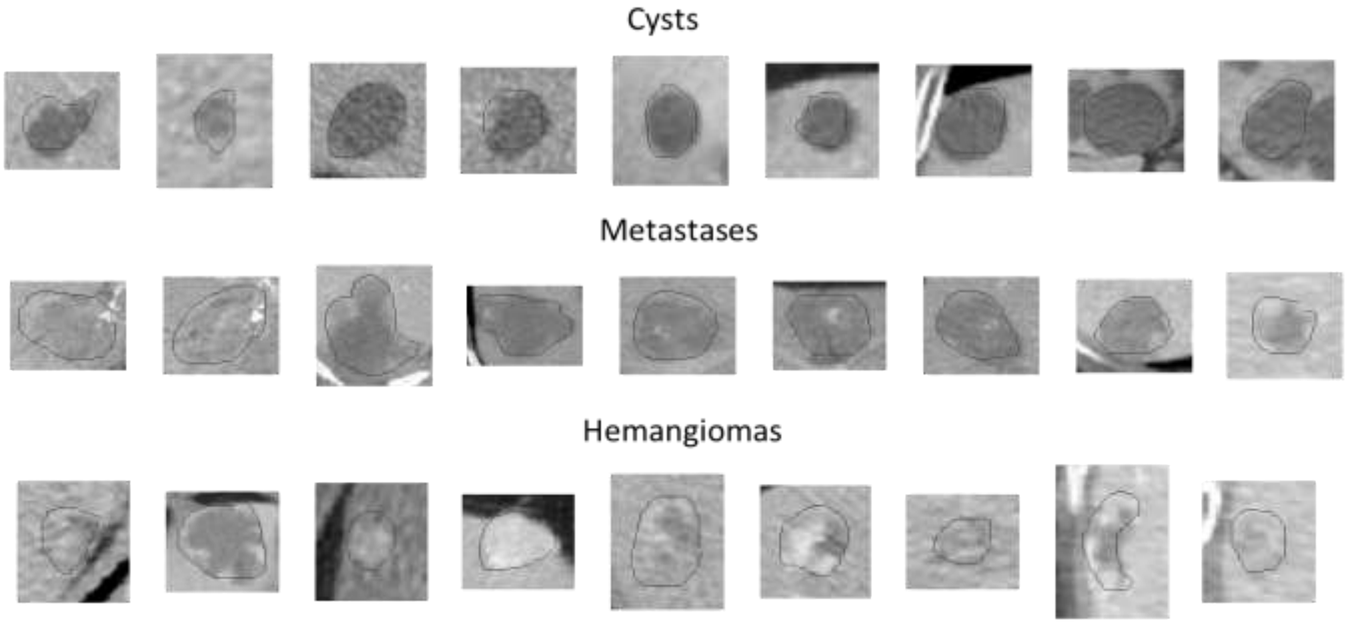


Fig. 2. Dataset 2: Typical examples of Cysts, Hemangiomas and Metastases. For each lesion, the manual segmentation that was done by an expert clinician (E.K) is shown

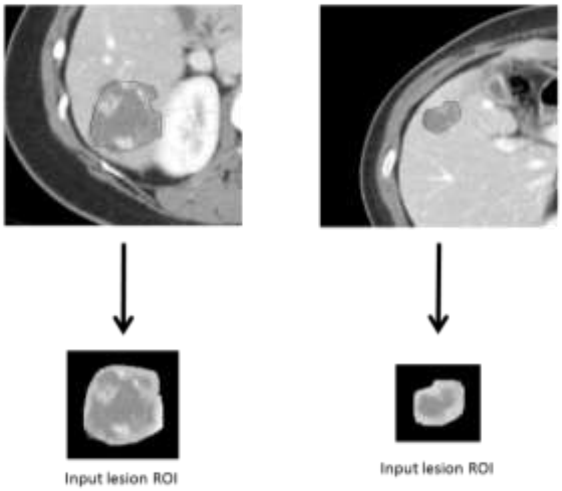


Fig. 3. Input lesion ROI examples. The upper row presents the CT liver images with the clinicians' manual marking. In the 2nd row, the cropped ROI for each lesion is demonstrated. The classification process which is obtained below is applied only over the cropped ROI

C. Dictionary of visual words

After obtaining the features vectors from all patches inside the ROI, they are clustered by using k-means algorithm. The centroids' centers are not chosen randomly. Similar to [30], the centers are chosen to be as far as they can from each other. This ensures the stability of the clustering procedure. Each evaluated cluster represents a single visual word. The visual words are constructed from all training images then assemble into a dictionary of visual words [23-26].

Let X be a set of D-dimensional local features that are extracted from different patches in the image, $X = [x_1, x_2......x_N] \in R^{D \times N}$. Given also a dictionary B with K visual words $B = [b_1, b_2.....b_N] \in R^{D \times K}$. Vector quantization (VQ) solves the following least square problem and finds the optimal code for input x with B [31]:

$$\arg\min_{c} \sum_{i=1}^{N} \| x_i - Bc_i \|^2 \ ,$$

$$\| c_i \|_{l} 0 = 1, \ \| c_i \|_{l} 1 = 1, \ c_i \geq 0, \ \forall i \qquad (1)$$

where $C = [c_1, c_2......c_N]$ is the set of codes for X. The constraint $\| c_i \|_{l} 0 = 1$ means that there will be only one non-zero element in each code c_i, corresponding to the VQ. The second constraint $\| c_i \|_{l} 1 = 1, \ c_i \geq 0, \ \forall i$ means that the coding weight for x is 1. The single non-zero element is found by searching the nearest neighbor. After the VQ is applied, sum-pooling is done in order to get adequate characterization of the examined area.

In the current work we propose a variation of the BoVW methodology. We create two dictionaries of visual words for liver lesions by generating separate dictionaries using image patches obtained from inside the lesion and from the boundaries of lesions. Thus, the interior and margin of a lesion are characterized separately when creating image feature vectors. The motivation for this approach is based on the fact that lesions differ from each other by their boundary as well as by their interior characteristics. By constructing separate dictionaries for the boundary and the interior regions, both regions contribute to the task. In the more customary approach, a single dictionary is used, including words from both the lesion's boundary and interior. In this scenario, the boundary contribution is reduced due to the smaller number of patches that it contains.

The input lesion ROI is first divided into boundary area and lesion interior area (see Fig. 4). The boundary region is extracted by applying a dilation operation around the initial radiologist markings. Interior patches are chosen to be located fully inside the lesion. Otherwise, they are considered as boundary patches.

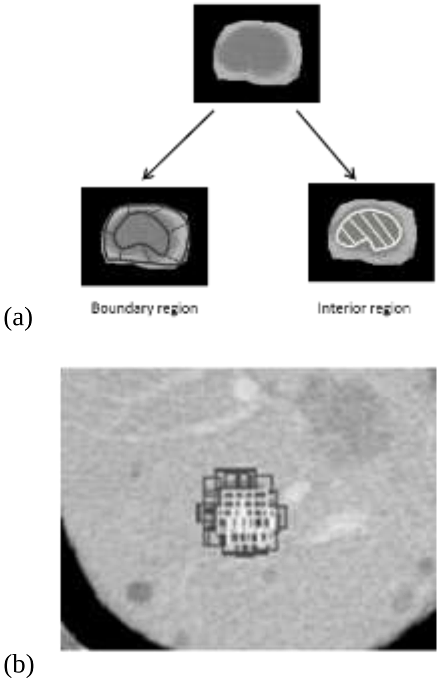


Fig. 4. Lesion boundary and interior regions. (a) Pre-definition of a lesion's interior area and its boundaries according to the clinicians' manual marking, (b) Boundary patches (gray); Interior patches (white)

Fig. 5a displays a block diagram of the two dictionaries generation. Given an input image (training or testing image), two histograms are built, one for the boundary region using the boundary dictionary and one for the interior region using the interior dictionary. These two histograms are then concatenated, creating the new feature vector that represents the lesion. Fig. 5b shows the algorithm flowchart for creating the dual dictionary representation, which is the concatenated histogram, for a given input ROI.

D. Classification using SVM with Histogram Intersection kernel

In the proposed method, we use non-linear SVM with Histogram Intersection (HI) kernel. Let $p=\{p_1 ...., p_k\}$ and $q=\{q_1 ...., q_k\}$ be two histograms of K visual words, the intersection kernel of those histograms is calculated by:

$$HI\{p,q\}=\sum_{i=1}^{k}\min\{p(i),q(i)\} \quad , \quad p,q\geq 0 \qquad (2)$$

We selected the HI kernel due to the fact that it was found to outperform other possible SVM kernels (Barla et al. [32]). In addition, the histogram intersection has no free kernel parameters, which makes it convenient for fast parameter evaluation.

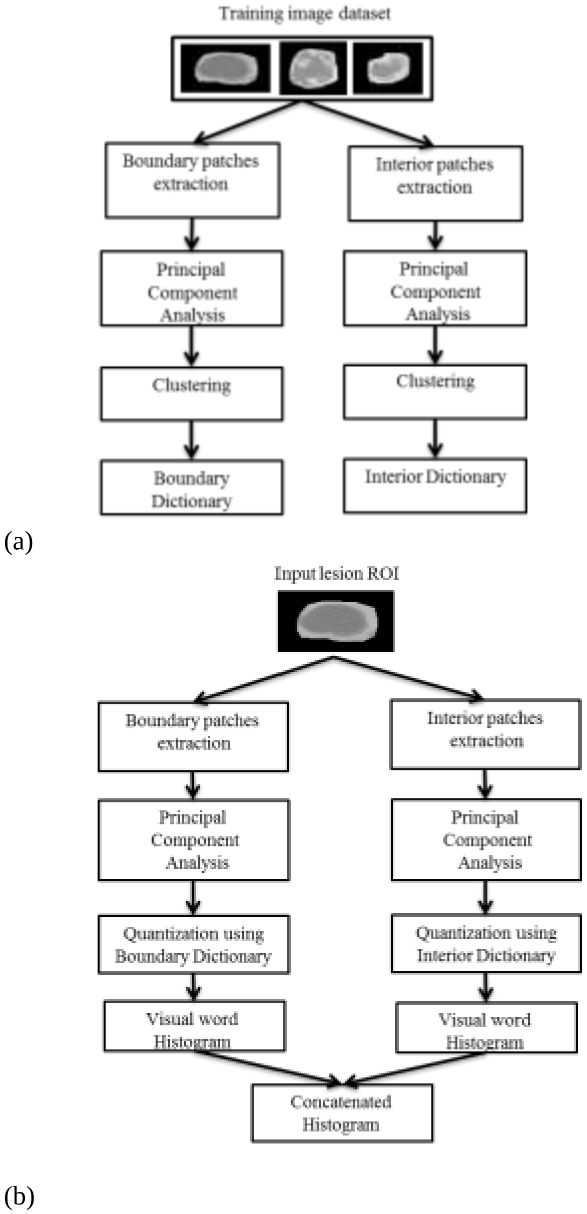

Training image dataset
Boundary patches extraction
Interior patches extraction
Principal Component Analysis
Principal Component Analysis
Clustering
Clustering
Boundary Dictionary
Interior Dictionary
(a)
Input lesion ROI
Boundary patches extraction
Interior patches extraction
Principal Component Analysis
Principal Component Analysis
Quantization using Boundary Dictionary
Quantization using Interior Dictionary
Visual word Histogram
Visual word Histogram
Concatenated Histogram
(b)

Fig. 5. Dual dictionary algorithm flowchart – (a) Boundary and interior, (b) ROI representation using dual dictionary method

E. Evaluation

We evaluated the algorithm performance using our two datasets, separately and together (termed here on "Combined dataset"). Evaluation for each dataset was achieved by using a leave-one-out cross-validation method. Additionally, we used training and testing sets from different sources, separately using each for training and testing ("Dataset 1 trained on Dataset 2", and "Dataset 2 trained on Dataset 1"). We tested the algorithm performance for different parameter values: patch sizes varying between 5*5 and 13*13 pixels, visual word size (amount of PCA coefficients) between 10 and 20, and visual word amount in dictionary (dictionary size) between 80 and 320. We performed grid search optimization to obtain a combination of parameters which achieves best performance (see results section, Fig. 6).

Additionally, patches were extracted from every second pixel in the ROI. This step size was found to be the optimal in terms of classification accuracy and computation time (see results section, Fig. 7). In cases of small lesions, relatively small amount of patches can be extracted. The clustering, as well as the reconstructed histogram, will thus not be able to characterize optimally the lesion. Using the 2-pixel step size results in an overlap between adjacent patches, thus improving the lesion's characterization.

We evaluated the algorithm performance using a total classification accuracy measure according to the following equation:

$$Total\ Accuracy = \frac{\sum_{all\ lesion types} TP}{Amount\ of\ lesions} \qquad (3)$$

Additionally, we calculated confusion matrix and sensitivity and specificity measures for each category, using the following equations:

$$Sensitivity=\frac{(TP)}{(TP+FN)} \qquad (4)$$

$$Specificit\,y=\frac{(TN)}{(TN+FP)} \qquad (5)$$

For each category, positives (P) are the specific category and negatives (N) are the other two classes.

III. Results

A. Performance evaluation

Table I shows the results that were obtained using different BoVW dictionaries for each dataset. Using only interior or only boundary regions yielded poorer results than using both regions ("dual dictionary"). Using only the interior dictionary produces better performance than using the boundary dictionary alone. This suggests that there is more discriminative information within the interior region. Table I also presents the performance of the methods by using training and testing sets from different sources ("Dataset 1 trained on Dataset 2", "Dataset 2 trained on 1"). This is an even more challenging task due to the substantial differences between the datasets.

TABLE I
CLASSIFICATION ACCURACY (OPTIMAL RESULTS) ACROSS ALL DATASETS.

Dataset	Dual dictionary	Only Interior dictionary	Only Boundary dictionary
Dataset 1 trained on Dataset 1	99.08%	96.33%	88.99%
Dataset 1 trained on Dataset 2	92.66%	84.40%	75.23%
Dataset 2 trained on Dataset 2	88.24%	82.35%	70.59%
Dataset 2 trained on Dataset 1	81.18%	77.65%	72.94%
Combined dataset	92.78%	88.14%	77.32%

TABLE II
CONFUSION MATRIX FOR DATASET 1 (TRAINED ON 1)

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	39	0	0	**100%**
Met	0	46	0	**100%**
Hem	0	1	23	**95.8%**
Specificity	**100%**	**98.4%**	**100%**	

TABLE III
CONFUSION MATRIX FOR DATASET 2 (TRAINED ON 2)

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	21	1	0	**95.5%**
Met	1	30	3	**88.2%**
Hem	0	5	24	**82.8%**
Specificity	**98.4%**	**88.2%**	**94.6%**	

TABLE IV
CONFUSION MATRIX FOR COMBINED DATASET

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	59	2	0	**96.7%**
Met	0	77	3	**96.3%**
Hem	1	9	43	**81.1%**
Specificity	**99.2%**	**90.4%**	**97.9%**	

TABLE V
CONFUSION MATRIX FOR DATASET 1 (TRAINED ON 2)

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	37	2	0	**94.9%**
Met	0	41	5	**89.1%**
Hem	0	1	23	**95.8%**
Specificity	**100%**	**95.2%**	**94.1%**	

TABLE VI
CONFUSION MATRIX FOR DATASET 2 (TRAINED ON 1)

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	20	1	1	**90.9%**
Met	1	30	3	**88.2%**
Hem	0	10	19	**65.5%**
Specificity	**98.4%**	**78.4%**	**92.9%**	

The discrimination between hemangiomas and metastases is the most challenging task, as shown in the confusion matrices (Tables II-VI). It was the least accurate among all three pair-wise classifications and yielded the highest amount of miss-classifications. Separating hemangiomas and cysts is shown to be the easiest task among the three. The highest sensitivity and specificity values were obtained for cysts for all datasets. For hemangiomas and cysts, specificity was higher than sensitivity, meaning more false-negatives (FN) than false-positives (FP), while for metastases, sensitivity was higher than specificity. From a clinical point of view, it is preferable to have greater FP than FN for metastases, since it is a malignant type of disease.

Better results (on the order of ~10%) were obtained using Dataset 1 compared with using Dataset 2 throughout the entire set of experiments. This may be due to the fact that Dataset 2 is more difficult for human expert classification than Dataset 1, since it contains fewer typical lesions.

B. *Sensitivity to parameters*

We next present a robustness analysis of the presented system. The algorithm performance for different parameters values was compared: using patch sizes between 5*5 and 13*13, visual word size between 10 and 20 and number of visual words in the dictionary between 80 and 320. Figure 6 shows the total accuracy obtained using Dataset 1 (trained on Dataset 1) for each parameter combination. We can see that for intermediate patches having width of 7*7 and 9*9 pixels, we obtain higher accuracy levels than for small or large patches (5*5 or 13*13 pixels, respectively). The best performance was obtained using the following parameter combinations: 1) patch size of 7*7 pixels, dictionary size of 160 words and word size of 14 coefficients 2) patch size of 7*7 pixels, dictionary size of 240 words and word size of 14 coefficients, 3) patch size of 9*9 pixels, dictionary size of 200 words, and word size of 12 coefficients (as shown by white areas in Fig. 6).

Figure 7 shows the sensitivity to step size, which represents the spacing in pixels between two adjacent patch centers. As shown in Figure 7a, a 2-pixel step size supplied the highest classification accuracy with a reasonable computation time. This is the case for the training phase (as shown in Fig. 7b), as well as the testing phase. In the testing phase, an average of 0.06 seconds was achieved when using the 2-pixel step and 0.18 seconds with a 1-pixel step.

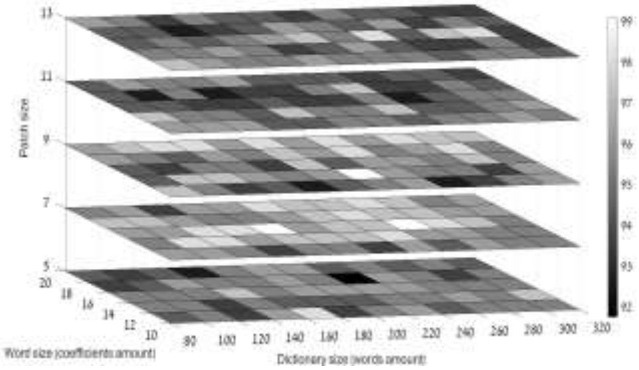


Fig. 6. Total accuracy with changing dictionary size, patch size and word size using the dual dictionary method for Dataset 1 trained on Dataset 1 (color bar shows accuracy).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2478255, IEEE Journal of Biomedical and Health Informatics

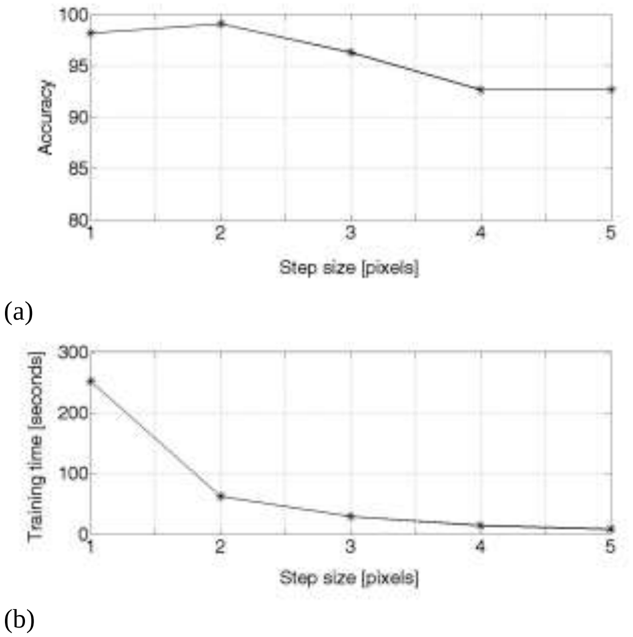


Fig. 7. Sensitivity to step size: a) accuracy versus step size for using optimal parameter combination, b) training phase running time (dictionary generation) for each step size.

Figure 8 presents a plot of the amount of variance per PCA coefficient amount (word size) from 1 to 20 for each patch size (5*5 to 13*13). This figure shows that for our range of word size (10-20 coefficients) we obtain above 90% amount of variance. It also shows that the first 14 principal components account for or "explain" 98.5% of the overall variability for patch size 7*7 pixels (which is best parameter combination).

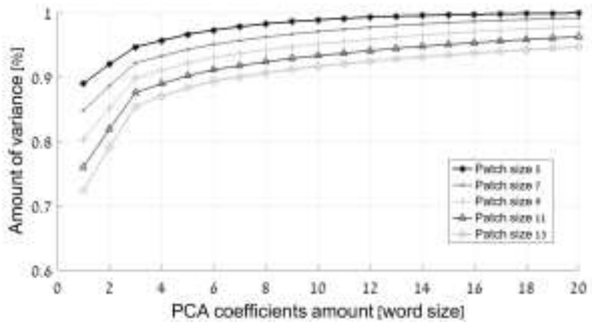


Fig.8. Amount of variance per PCA coefficient amount (word size) for each patch size (5*5 to 13*13).

Finally, we checked the algorithm sensitivity to changes in expert's manual markings. We changed each contour point of the original manual markings randomly by 3 pixels moving closer or further from the lesion center. Twenty different initializations have been tested. For each, we applied our algorithm with combinations of parameters which yielded best performance for the original markings. Results are shown in Fig. 9. It can be seen that the results are robust to the variations in given manual markings.

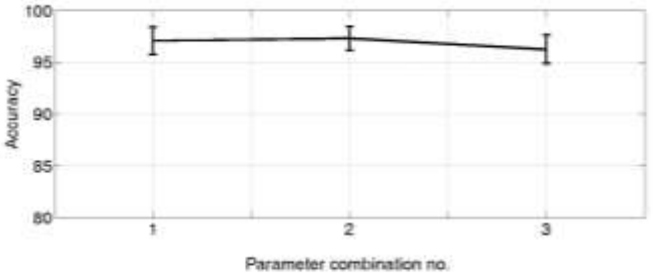


Fig. 9. Sensitivity to manual markings. For specific parameter combination (chosen from the best combination in Fig.6), the error bar in the graph shows mean and the standard deviation of the classification over 20 different manual markings of the ROI.

C. Comparison to an Independent observer classification

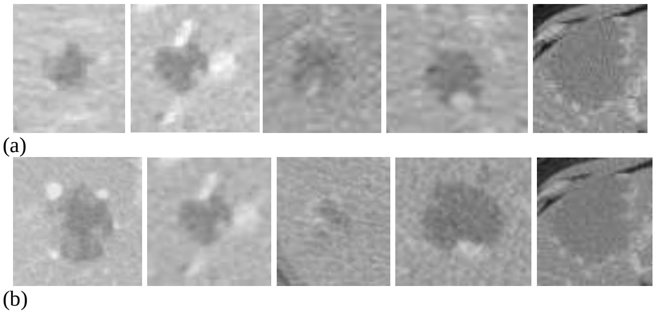
We compared our system's performance with the diagnosis of an independent clinician observer for the analysis of Dataset 2. Both our algorithm and the analysis of the independent observer were compared with the ground truth classification. Our method produced a diagnostic accuracy of 88.2% while the diagnosis accuracy of the independent observer was 85.9%. Table VII shows the confusion matrix obtained by an independent observer and the ground truth. Table VIII shows performance comparison between our system and the independent observer's diagnosis. By comparing this result with the confusion matrices in Table III and VI, we can see that the expert observer had similar miss-classifications with two additional hemangiomas being misclassified as metastases. Fig. 10 shows miss-classification examples of our system and that of the independent observer.

TABLE VII
CONFUSION MATRIX FOR DATASET 2 FOR AN INDEPENDENT OBSERVER DIAGNOSIS

True\Auto	Cyst	Met	Hem	Sensitivity
Cyst	21	1	0	**95.5%**
Met	1	30	3	**88.2%**
Hem	0	7	22	**75.9%**
Specificity	**98.4%**	**84.3%**	**94.6%**	

TABLE VIII
PERFORMANCE COMPARISON FOR DATASET 2 BETWEEN OUR SYSTEM AND AN INDEPENDENT OBSERVER

	Our method – Sensitivity	Our method – Specificity	Observer - Sensitivity	Observer – Specificity
Cyst	95.5%	98.4%	95.5%	98.4%
Met	88.2%	**88.2%**	88.2%	84.3%
Hem	**82.8%**	94.6%	75.9%	94.6%



(a)

(b)

Fig. 10. Miss-classification examples: a) miss-classifications of our system, b) miss-classifications of an independent observer. Two left images in each row are metastases misclassified as hemangiomas, and three right images are hemangiomas misclassified as metastases.

D. Comparison with other classification methods

We compared our results with other alternative approaches, such as ones using a selected feature set and a strong state-of-the-art classifier. Several common feature sets were used, including gray-level co-occurrence matrix (GLCM) and Gabor features, along with the SVM classifier. For GLCM, we obtained all Haralick features [33] and optimized accuracy levels using different distances and angles. For Gabor, we obtained the mean, standard deviation, energy and entropy for each scale and orientation. We performed grid search optimization to choose the scale, orientation, and filter size obtaining best performance (Table IX).

TABLE IX
COMPARISON WITH OTHER TEXTURE FEATURES, USING THE OPTIMAL PARAMETER SET.

Method	Combined Dataset
Dual dictionary	***92.78%***
Single dictionary	91.24%
GLCM	80.41%
Gabor	82.47%

In addition, comparison with a single-dictionary BoVW has been applied, constructing a single dictionary for interior and boundary patches. We performed parameter optimization (grid search optimization) in order to get the optimal results using single dictionary. We can see that the dual dictionary method achieved best performance.

Fig. 11 shows the stability of dual dictionary and its sensitivity to different parameter sets. The dual dictionary had relatively narrow distribution for all datasets, thus was found stable. This idea can be seen in each of the analysed datasets but mostly in Dataset 2 (Fig. 11b) and in the combined dataset (Fig. 11c). For some sets, the single dictionary resulted in classification accuracy lower than 75% and 85% for Dataset 2 and the combined dataset, respectively. For all datasets, the dual dictionary showed higher accuracy.

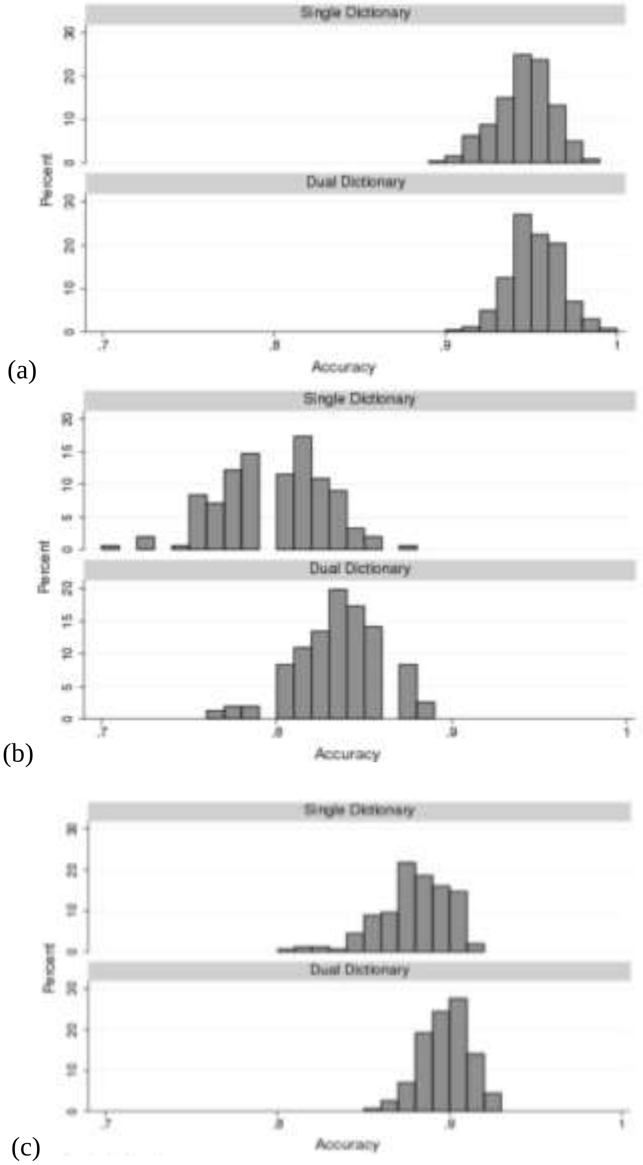


Fig. 11. Distribution of the classification accuracy for (a) Dataset 1, (b) Dataset 2, (c) Combined dataset.

E. Statistical tests

Both algorithms, single and dual dictionaries, were tested with different parameter settings on 109 cases from Dataset 1, and 85 cases from Dataset 2, as well as the combined datasets. Accuracy was assessed as the proportion of cases correctly determined. Since proportions are beta-distributed, beta distributions were fit to each algorithm's performance. Rather than the usual alpha/beta shape parameterization, the beta distribution was parameterized by μ (the mean) and φ (the precision, i.e., a larger value of φ is associated with a smaller variance). Algorithms were compared with a beta regression [38]. Algorithm type was tested simultaneously as a predictor for both μ and φ, and the resulting odds ratio between algorithms' performance was estimated (Odds defined as the ratio of the probability of success and the probability of failure). All statistical analyses were performed using Stata Release 13.1 (StataCorp LP, College Station, TX).

TABLE X
BETA DISTRIBUTION COMPARISON BASED ON TOTAL ACCURACY MEASURE BETWEEN SINGLE AND DUAL DICTIONARY METHODS FOR ALL DATASETS. THE VALUES HAVE BEEN AVERAGED FOR ALL TESTED SETS OF PARAMETERS.

Method (N – number of samples)	μ	95% CI	ϕ	95% CI	Odds Ratio for μ
Dataset 1 – Single (N=240)	0.946	0.944 – 0.948	212	177 – 254	1.13*
Dataset 1 – Dual (N=240)	0.952	0.95 – 0.953	201	168 – 240	
Dataset 2 – Single (N=156)	0.798	0.793 – 0.802	180	144 – 224	1.28*
Dataset 2 – Dual (N=156)	0.835	0.831 – 0.839	226	181 – 282	
Combined dataset – Single (N=156)	0.878	0.875 – 0.881	259	208 – 324	1.21*
Combined dataset – Dual (N=156)	0.897	0.895 – 0.899	459	368 - 573	

* P<0.0001

Table X shows the statistical analysis for both single and dual dictionaries. It presents the μ and φ parameters, their 95% confidence interval (CI) and the odds ratio for μ. Those are the average values that have been calculated for all tested sets of parameters. The dual dictionary was significantly more accurate than the single one for both datasets and combined (p<0.0001), as indicated by a higher mean accuracy and an odds ratio for μ greater than unity.

We also compared the classification results of the automated algorithm with the equivalent values of the 2 independent observers (Table XI). Significant improvement was found between our presented method and each observer (unweighted Kappa, p<0.001) [35].

TABLE XI
UNWEIGHTED KAPPA DISTRIBUTION THAT WAS CALCULATED FOR DIFFERENT COMPARISONS OF OUR PRESENTED METHOD AND EACH OF THE 2 OBSERVERS. 95% CI IS ALSO PRESENTED (P<0.001).

Method	Observed agreement	Kappa [95% CI]
Our - Ground	0.894	0.838 [0.738, 0.938]
Our - Observer1	0.823	0.729 [0.604, 0.854]
Our – Observer2	0.8	0.696 [0.566, 0.825]
Observer1 – Observer2	0.858	0.785 [0.673, 0.898]
Observer1 - Ground	0.858	0.784 [0.671 0.897]
Observer2 - Ground	0.847	0.768 [0.651, 0.884]

IV. DISCUSSION AND CONCLUSION

This work focuses on automated liver lesion diagnosis for three specific liver lesion diagnoses of Cyst, Metastasis and Hemangioma. We present a new approach for the liver lesion analysis based on a BoVW model that analyses the interior and boundary lesion regions separately. This work was conducted on CT data.

In recent work by Huang [36], a similar concept of separating the interior and boundary for individual representation is presented. However, our work differs from Huang’s method in two key ideas. First, Huang et al. analyze MRI brain tumors, while in this work we analyze CT liver lesions. The latter are characterized by much noisier regions and by lower contrast lesions, especially hemangiomas. Therefore, applying Gaussian filters as in Haung will decrease the contrast between the lesion and its surrounding even more, thus it will make the classification process more complicated. In addition, Haung et al. uses a normal to the boundary at every point and extracts intensity profiles along this normal. The authors pointed out that this step may be sensitive to noise, and thus may not be optimal in our case of characterizing liver lesions. Second, our method differs from Huang’s method in terms of the overall representation of the internal region and the boundary; their spatial descriptor is more elaborate, having 6 regions (tumor region, tumor-surrounding area and 4 additional sub-regions which they find to be useful to represent the boundary). Therefore, each boundary point is considered twice – once for level 1 and once for level 2 (for more local analysis), while the internal part is considered only in level 1 – thus only once. From Huang’s conclusions of the results [36], it seems that the boundary in the MR brain tumor application is of major significance. In our work, we show that for the CT liver lesion categorization task, using interior information contributes more than the boundary (Table I).

In the current work we found that a dual dictionary BoVW approach yields high accuracy for liver lesion classification. The intuition behind a dual dictionary approach is that the interior and boundary of the lesions have distinct visual characteristics, and the complexity in lesion appearance can be better captured through separate visual dictionaries for these lesion components than by using a single dictionary. For the three lesion types that were investigated, the lesion interior information was more relevant for classification than the boundary information, as shown in Table I. Further improvement is achieved by using a dual dictionary that takes into account both interior and boundary information, while preserving the contribution of each (Table I, X). In this case, additional information about the relative location of a visual word (as a lesion interior or boundary representative feature) is preserved. When we used only a boundary dictionary or only an interior dictionary, we obtained lower accuracy (e.g., in the combined dataset, 88% accuracy using interior dictionary, and 77% accuracy using boundary dictionary), compared to using dual dictionary (93% accuracy). Table I also shows that regardless of our training and testing sets—even if the training and testing datasets are substantially different from each other—the dual dictionary is superior to using only interior or boundary dictionary. Interestingly, we found there to be variability in classification results between our two datasets; there was approximately a 10% difference in classification performance when applying our method to the two datasets. This difference could be explained by the different appearance of the lesions in ach set. Neither of these initial datasets were controlled carefully for the original tumor type leading to metastasis, which may also help explain the results.

Although our results pertain to specific datasets and a specific algorithm, we believe that such variability could be expected among classification systems in general. Further building the training and test databases from both institutions and including other types of liver lesions will be helpful in better understanding the sources of variability and differing accuracy.

The automated system aims to help radiologists and reduce the number of biopsies and imaging follow up studies. Our goal was to build a system that has similar performance to an expert radiologist. Therefore, an independent observer diagnosed the cases in Dataset 2, which is a more challenging dataset, and his results were compared to our algorithm's performance. Our system produced higher accuracy compared to the expert observer (Table VIII) obtaining similar miss-classifications for the most difficult cases as shown in Fig. 10.

We compared our results with a single dictionary BoVW. The results suggest that using two dictionaries is superior to using a single dictionary in every comparison that was conducted. The dual dictionary approach may perform better than single dictionary, since it provides a better representation of the lesion boundary and interior characteristics by using separate visual words. Additionally, Table X shows that for more typical cases ('Dataset1'), both methods showed similar performance. However, for less typical cases ('Dataset2'), substantial differences in the performance of the methods were found. Moreover, by examining the combined dataset, the dual dictionary showed a significantly higher accuracy (p=0.004) than the single dictionary.

We also performed additional comparisons to other methods using Gabor and GLCM features and showed that our method is more discriminative than a selection of pre-defined features (Table IX).

Our experiments showed that each dataset had different optimal parameter settings, e.g., Dataset 2 applied a 13*13 patch size, while a 7*7 patch was used for the optimal analysis of Dataset 1. However, when using a combined dataset, both the 7*7 and 13*13 patch sizes resulted in optimal performance. These results indicate two possible scenarios for making the system usable in a clinical setting: One scenario can be to tune the system parameters based on a training set per given clinical site. A second scenario is to use a training set that is taken from many different medical centers, and learn an optimal global parameter set. Using such a combined dataset will take into account the variability among datasets and will enable strong performance overall.

There are several directions for future work. We plan to explore additional features, such as scale invariant and rotation invariant features extracted from the image, such as SIFT [25,39]. Alternative methods for dimensionality reduction can be used, such as Independent Component Analysis [38], which impose dependency requirements (data along reduced dimensions are statistically independent). Finally, to reduce the sensitivity to outliers in using PCA, we could employ a Weighted PCA approach [39], where we rank each image ROI with a weighted term representing its quality.

There are several limitations of our current work. First, focusing on a single phase instead of multi-phase data is a limitation that could be addressed in future work. Another limitation of our method is that it was applied only to 2D images. Extension of the two-dimensional analysis to 3D volumetric data could be done in the future and offer advantages. Tables II-IV show that the most challenging classification occurs between Hemangiomas and Metastasis cases; additional 3D information may be informative and reduce the frequency of misclassifications. Finally, we intend to extend our dataset using more lesions from each type of diagnosis, as well as apply our methods to a wider range of lesion types.

In conclusion, our new method of using dual dictionaries for BoVW shows promising results, with a statistically significant improvement in performance over that achieved using a single dictionary. We believe the benefits of our approach may generalize to other application domains within radiology, and that these methods may help radiologists make more accurate diagnoses.

Acknowledgment

We would like to thank Prof. Jacob Goldberger from Bar-Ilan University for providing us K-means algorithm implementation.

References

[1] WHO, "Globocan 20120: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012," Available from: http://globocan.iarc.fr/Default.aspx.

[2] National Cancer Institute at the NIH, "FactSheet," Available from: http://www.cancer.gov/cancertopics/factsheet/Sites-Types/metastatic.

[3] H. M. Taylor, and P. R. Ros, "Hepatic imaging: An overview," Radiologic Clinics Vol. 36(2), pp. 237-245, 1998.

[4] M. Gletsos, S. G. Mougiakakou, G. K. Matsopoulos, K. S. Nikita, A. S. Nikita, D. Kelekis, "A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier," IEEE Transaction on Information Technology in Biomedical, vol. 7(3), pp. 153–62, 2003.

[5] A. Quatrehomme, I. Millet, D. Hoa, G. Subsol, W. Puech, "Assessing the Classification of Liver Focal Lesions by Using Multi-phase Computer Tomography Scans," in Medical Content-Based Retrieval for Clinical Decision Support, MCBR-CDS 2012, pp. 80-91.

[6] W. Yang, Z. Lu, M. Yu, M. Huang, Q. Feng, W. Chen, "Content-Based Retrieval of Focal Liver Lesions Using Bag-of-Visual-Words Representations of Single- and Multiphase Contrast-Enhanced CT Images," Journal of Digital Imaging, vol. 25(6), pp. 708-719, 2012.

[7] S. Upadhyay, M. Papadakis, S. Jain, G. Gladish, I.A. Kakadiaris, R. Azencott, "Semi-Automatic Discrimination of Normal Tissue and Liver Cancer Lesions in Contrast Enhanced X-Ray CT-Scans," in Medical Image Computing and Computer-Assisted Intervention, MICCAI 2012, edited by H. Yoshida, D. Hawkes, M.W. Vannier (Springer-Verlag Berlin, Heidelberg, 2012), pp. 158-167.

[8] M. Freiman, O. Cooper, D. Lischinski, L. Joskowicz, "Liver tumors segmentation from CTA images using voxels classification and affinity constraint propagation," International Journal of Computer Assisted Radiology and Surgery, vol. 6, pp. 247–55, 2011.

[9] J. Zhou, W. Huang, W. Xiong, W. Chen, S.K. Venkatesh, Q. Tian, "Delineation of Liver Tumors from CT Scans Using Spectral Clustering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2478255, IEEE Journal of Biomedical and Health Informatics

with Out-of-Sample Extension and Multi-windowing," in Medical Image Computing and Computer-Assisted Intervention, MICCAI 2012.

[10] X. Zhang, J. Tian, D. Xiang, X. Li, K. Deng, "Interactive liver tumor segmentation from CT scans using support vector classification with watershed," in proceedings of the 33rd International Conference of the IEEE Engineering Medical Biology Society (EMBS), pp. 6005-6008 2011.

[11] Y. Masuda, T. Tateyama, W. Xiong, J. Zhou, M. Wakamiya, S. Kanasaki, A. Furukawa, Y.W. Chen, "Liver tumor detection in CT images by adaptive contrast enhancement and the EM/MPM algorithm," IEEE International Conference on Image Processing (ICIP), pp. 1421–1424, 2011.

[12] J. H. Moltz, L. Bornemann, J. M. Kuhnigk, V. Dicken, E. Peitgen, S. Meier, H. Bolte, , M. Fabel, H.C. Bauknecht, M. Hittinger, A. Kiessling, M. Pusken, H.O. Peitgen, “Advanced segmentation techniques for lung nodules, liver metastases, and enlarged lymph nodes in CT scans,” IEEE Journal of Selected Topics in Signal Processing, vol. 3(1), pp. 122–134, 2009.

[13] M. Schwier, J. Hendrik, H. Peitgen, “Object-based analysis of CT images for automatic detection and segmentation of hypodense liver lesions,” International Journal of Computer Assisted Radiology and Surgery, vol. 6(6), pp. 737–747 2011.

[14] A. Militzer, T. Hager, F. Jager, C. Tietjen, J. Hornegger, “Automatic Detection and Segmentation of Focal Liver Lesions in Contrast Enhanced CT Images,” International Conference on Pattern Recognition, pp. 2524–2527, 2010.

[15] M. Bilello, S.B. Gokturk, T. Desser, S. Napel, R.B. Jeffrey, C.F. Beaulieu, “Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT,” Medical Physics, vol. 31(9), pp. 2584-2593, 2004.

[16] A. Shimizu, T. Narihira, D. Furukawa, “Ensemble segmentation using AdaBoost with application to liver lesion extraction from a CT volume,” in Proceedings MICCAI Workshop on 3D Segmentation in the Clinic: A Grand Challenge II, 2008.

[17] Y. Chi, J. Zhou, S.K. Venkatesh, S. Huang, Q. Tian, T. Hennedige, J. Liu, “Computer-aided focal liver lesion detection,” International Journal of Computer Assisted Radiology and Surgery, vol. 8(4), pp. 511-525, 2013.

[18] J. Ye, Y. Sun, S. Wang, L. Gu, L. Qian, J. Xu, “Multi-Phase CT Image Based Hepatic Lesion Diagnosis by SVM,” International Conference on Biomedical Engineering and Informatics (BMEI), 1-5, 2009.

[19] S. Bakasa, K. Chatzimichailb, G. Huntera, B. Labbéc, P. S. Sidhud and D. Makrisa, “Fast semi-automatic segmentation of focal liver lesions in contrast-enhanced ultrasound, based on a probabilistic model,” Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2015.

[20] A Anaye, G Perrenoud, N Rognin, M Arditi, L Mercier, P Frinking, C Ruffieux, P Peetrons, R Meuli, JY Meuwly, “Differentiation of focal liver lesions: usefulness of parametric imaging with contrast-enhanced US”, Radiology, 261,pp. 300-310, 2011.

[21] S. Bakas, A Hoppe, K. Chatzimichail, V. Galariotis, G. Hunter, D. Makris, “Focal Liver Lesion Tracking in CEUS for Characterisation Based on Dynamic Behaviour”, Advances in Visual Computing Lecture Notes in Computer Science 7431, pp 32-41, 2012.

[22] C.D. Manning, P. Raghavan, H. Schutze, “Introduction to information Retrieval,” Cambridge University Press, 1 edition, 2008.

[23] J. Yang, Y.G. Jiang, A.G. Hauptmann, and C.W. Ngo, “Evaluating bag-of-visual- words representations in scene classification,” in Proceedings of the International Workshop on Multimedia Information Retrieval, pp. 197-206, 2007.

[24] G. Csurka, C. Dance, C. Bray, L Fan, “Visual categorization with bags of keypoints,“ in Proceedings Workshop on Statistical Learning in Computer Vision, 2004.

[25] 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 Transactions on Medical Imaging, vol. 30(3), pp. 733–746 2011.

[26] I. Diamant, J. Goldberger, H. Greenspan, “Visual words based approach for tissue classification in mammograms”, in Proceedings SPIE Medical Imaging, vol. 8670, 2013.

[27] J. Wang, Y. Li, Y. Zhang, H. Xie, “Bag-of-Features Based Classification of Breast Parenchymal Tissue in the Mammogram via Jointly Selecting and Weighting Visual Words,” in Proceeding of the International Conference on Image and Graphics (ICIG), pp. 622-627, 2011.

[28] S. Napel, C.F. Beaulieu, C. Rodriguez, J. Cui, J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. Ma, D.L. Rubin, “Automated retrieval of CT images of liver lesions on the basis of image similarity: method and preliminary results,” Radiology, vol. 256(1), pp. 243–252, 2010.

[29] J.E.A. Jackson, User’s Guide to Principal Components, (Wiley, New York, 1991).

[30] S.S. Khan, A. Ahmad, “Cluster center initialization algorithm for K-mean clustering”, Pattern Recognition Letters, vol. 25, pp. 1293–1302, 2004.

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. “Locality-constrained linear coding for image classification”. Proc. CVPR, 2010.

[32] A. Barla, F. Odone, A. Verri, “Histogram intersection kernel for image classification,” International Conference on Image Processing (ICIP), vol. 3, pp. 513-516, 2003.

[33] R. Haralick, K. Shanmugam, and I. Dinstein, “Textural Features for Image Classification,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 3(6), pp. 610–621, 1973.

[34] S. Ferrari and F. Cribari-Neto, “Beta regression for modelling rates and proportions,” Journal of Applied Statistics, vol. 31(7), pp. 799-815, 2004.

[35] J. A Cohen, “Coefficient of agreement for nominal scales”. Educational and Psychological Measurement, vol. 20, pp. ,37-46, 1960

[36] M. Huang, W. Yang, M. Yu, Z. Lu, Q. Feng, and W. Chen, “Retrieval of Brain Tumors with Region-Specific Bag-of-Visual-Words Representations in Contrast-Enhanced MRI Images,” Computational and Mathematical Methods in Medicine, 2012.

[37] D. Lowe, “Object recognition from local scale-invariant features,” in Proceedings of the International Conference on Computer Vision (ICCV), 2, pp. 1150-1157, 1999.

[38] A. Hyvärinen , J. Karhunen, E. Oja, Independent Component Analysis (John Wiley & Sons, New York, 2001).

[39] H. P. Kriegel, P. Kröger, E. Schubert, A. Zimek, "A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms," in Proceedings of the 20th international conference on Scientific and Statistical Database Management (SSDBM), edited by B. Ludascher and N. Mamoulis (Springer-Verlag Berlin, Heidelberg, 2008), pp. 418-435.