A MEDICAL X-RAY IMAGE CLASSIFICATION AND RETRIEVAL SYSTEM

Mohammad Reza Zare, School of Information Technology, Monash University, Malaysia, mohammad.reza@monash.edu

Henning Müller, University of Applied Sciences Western Switzerland (HES-SO) Valais, Sierre, Switzerland, henning.mueller@hevs.ch

Abstract

Medical image retrieval systems have gained high interest in the scientific community due to the advances in medical imaging technologies. The semantic gap is one of the biggest challenges in retrieval from large medical databases. This paper presents a retrieval system that aims at addressing this challenge by learning the main concept of every image in the medical database. The proposed system contains two modules: a classification/annotation and a retrieval module. The first module aims at classifying and subsequently annotating all medical images automatically. SIFT (Scale Invariant Feature Transform) and LBP (Local Binary Patterns) are two descriptors used in this process. Image-based and patch-based features are used as approaches to build a bag of words (BoW) using these descriptors. The impact on the classification performance is also evaluated. The results show that the classification accuracy obtained incorporating image-based integration techniques is higher than the accuracy obtained by other techniques. The retrieval module enables the search based on text, visual and multimodal queries. The text-based query supports retrieval of medical images based on categories, as it is carried out via the category that the images were annotated with, within the classification module. The multimodal query applies a late fusion technique on the retrieval results obtained from text-based and image-based queries. This fusion is used to enhance the retrieval performance by incorporating the advantages of both text-based and content-based image retrieval.

Keywords: medical x-ray images, bag of visual words, medical image classification, medical image annotation, medical image retrieval.

1 INTRODUCTION

There is an increase of digital information in the medical domain where medical images of different modalities are produced every day in massive numbers. Medical imaging is a key component in diagnosis and treatment planning. As medical image databases show a wealth of information, these databases also bring problems in retrieving the desired images for specific information needs. As a result, searching and indexing large medical image databases efficiently and effectively has become challenging. Thus, the need for effective retrieval tools is present (Aggarwa, *et. al*., 2009). A large amount of research has been carried out on medical image retrieval in the last two decades. Generally, there are three types of approaches in medical image retrieval.

The first approach is traditional text-based image retrieval where text such as file names or attached keywords are used to retrieve images for a specific information need. Currently, most hospitals and radiology departments are equipped with Picture Archiving and Communications System (PACS). Such traditional systems have limitations, as they often do not allow for free text search in the radiology reports that are often stored in the RIS (Radiology Information System), but only allow to search for limited metadata such as the DICOM headers. This approach can be unreliable in this case, as some of the information might be missing or incorrect. Even though DICOM headers contain much important information, this still remains suboptimal due to errors in up to 30% of the headers (Güld, *et. al*., 2002), particularly in fields such as anatomy. Manual annotation of images is an alternative to header information but it is time consuming and error-prone because it is a subjective task.

The second approach is based on medical Content-Based Image Retrieval (CBIR). Medical CBIR complements traditional text based search for large image databases. Medical CBIR deals with the analysis of image content and the development of tools to represent visual content in a way that it can be efficiently searched and compared. However, there is a semantic gap between low level visual feature used to represent images and high level semantic concept used by human to interpret images. Normally humans recognize objects by using prior knowledge and experience of similar situations, which can be based on personal experience. This kind of information is hard to incorporate in medical CBIR systems. Another constraint of medical CBIR systems is that they are impractical for general users to use because users need to provide a query image in order to retrieve similar images, which works if a new case is concerned but does not work for more general information needs, as for example in teaching. More on the requirements of physicians regarding a retrieval system can also be found in (Markonis *et al.,* 2012). The article describes a survey among several radiologists regarding image search. In (Markonis *et al.* 2015) a user study of a prototype system is shown, again highlighting the important functionality of restricting search to specific modalities and anatomic regions and the importance to combined visual features and text.

In order to reduce the semantic gap in medical CBIR systems and to have an effective and accurate medical image retrieval application, new trends for image retrieval using automatic image classification and annotation have been investigated for the past few years. Such new approaches for medical image retrieval are named Automatic Medical Image Annotation (AMIA) where medical images can be retrieved in similar ways as text documents. The main goal of the AMIA approach is to learn the semantic concept of every image in the database and use the concept model accordingly to label new images. Once medical images are annotated with such labels, they can be retrieved by keywords, similar to text document retrieval. AMIA can be defined as medical image classification. This approach is mapping a new image into pre-defined categories and annotates them by propagating the corresponding words of that category. Automatic medical image classification provides a solution to address some of the challenges raised by text-related and DICOM-related systems. A successful classification can result in better retrieval performance, which can be beneficial in teaching, research and clinical decision making.

The classification task described in this paper begins with extracting appropriate visual features of the image, which is one of the most important factors in the design process of such a system. These features can be handcrafted for a task or be derived from convolutional neural networks, which is a more recent approach. Moreover, the feature extraction step affects all other subsequent processes. Various feature extraction techniques have been used in studies (Zare *et al*., 2014; Kurtz *et al*.,2014; Srinivas *et al.* 2015; Jiang *et al*.,2015; Cao *et al*.,2014; Garcia *et al.* 2015; Kesorn *et al*., 2012; Dimitrovski *et al*., 2011; Alba *et al*.,2012; Lazebnik *et al*., 2006). Among them, the scale invariant feature transform (SIFT) and subsequently Bag of Words (BoW) generated by applying clustering technique on SIFT, were successfully exploited in many medical image classification and retrieval tasks.

Local binary patterns (LBP) introduced by (Ojala *et al*., 2002) are also considered as one of the effective texture features due to their robustness and computational simplicity. Empirical studies show that LBP and BoW are powerful descriptors for local features in medical x-ray images. Zare *et al.* (2013) used various image representation techniques for classification of medical images. The highest accuracy was obtained using LBP and BoW compared to other representation techniques.

In recent years, feature fusion algorithms have also been used in various studies (Li *et al*.,2015; Garcia & Muller, 2014). For example, Zare *et al.* (2013) combined textual and visual features to create a medical x-ray image representation. Then PLSA is applied on these features to generate high level representations of the images. Cao *et al.* (2011) combined features from various modalities. They are then incorporated into an extended latent semantic analysis (LSA). Hong & Lele (2015) combined SIFT, LBP, Gabor texture and Tamura texture feature with tf-idf textual feature extraction from image description text in automatic modality classification of medical images. The system in this article consists of two sub-modules:

- Image classification/annotation: The aim of this sub-system is to automatically classify any new inserted medical image into one of the pre-defined categories. Images are then annotated with the corresponding keywords of each category.
- Image retrieval module: The aim of this module is to retrieve similar images to the user's query using different modalities such as text and image.

Inspired by the fact that fusion of information from textual and visual sources can improve the retrieval performance (Li *et al*.,2015; Garcia & Muller, 2014), the proposed system supports multimodal fusion in both classification and retrieval module. This paper confirms the positive influence of the fusion of textual and visual cues for the retrieval. The main contribution is of experimental nature applying several visual and automatically extracted textual features to the retrieval of medical images.

The rest of the paper is organized as follow. Section 2 discusses image representation techniques, two feature integration approaches and the classification results obtained using the extracted features. Section 3 presents the retrieval module followed by discussion in section 4. Finally, the overall conclusion of this experiment is presented in section 5.

2 IMAGE CLASSIFICATION AND ANNOTATION SUB-SYSTEM

In this research experiment, a classification module is designed and tested on the ImageCLEF 2007 medical database (Müller *et al*., 2007). This database consists of 11,000 training images and 1000 test images from 116 categories. The database is publicly available and thus allows for full reproducibility.

2.1 Feature Extraction

The first module is feature extraction. SIFT was employed as one of the feature representation technique in this system. The process of SIFT starts with detecting local interest points. Local interest point detectors extract specific points and areas based on saliency, so usually linked to higher grandients. One of the popular approaches for the detection of local interest point is the Difference of Gaussians (DoG), which is used in this experiment. This detector was chosen since it showed to perform well for the task of wide-baseline matching when compared to other detectors. DoG was built to be invariant to translation, scale, rotation, and illumination changes. Next, the detected key points are represented using SIFT. The image gradient is sampled and its orientation and quantized. Using a grid division of the local interest area, local gradient orientation histograms are created where the gradient magnitude is accumulated. The final feature is the concatenation of all the local gradient orientation histograms. The final descriptor proposed in this paper contains 8 orientations and 4 × 4 blocks, resulting in a descriptor of 128 dimensions.

Subsequently, the LBP descriptor is extracted from the images. The LBP operator assigns a label to every pixel of an image by thresholding the 3 × 3 neighborhood of each pixel with the center pixel value and considering the result a binary number. In the computation of the LBP histogram, the histogram has a separate bin for every uniform pattern, and all non-uniform patterns are assigned to a single bin. Using uniform patterns, the length of the feature vector for a 3 × 3 window reduces from 256 to 59.

Next step in the implementation of the bag of visual words is the codebook construction where the local image features have to be quantized into discrete visual words. This task is performed using clustering or vector quantization algorithms. This step usually uses the k-means clustering method, which clusters the keypoint descriptors in their feature space into a defined number of clusters and encodes each keypoint by the index of the cluster to which it belongs. We conceive each cluster as a visual word that represents a specific local pattern shared by the keypoints in the cluster.

Thus, the clustering process generates a visual-word vocabulary describing local patterns that are present in the images. The number of clusters determines the size of the vocabulary, which can vary from hundreds to over tens of thousands based on the exact requirements. Mapping the key points to visual words, we can represent each image as a "bag of visual words". In this task, there are two independent extracted descriptors: SIFT and LBP. Inspired by (Jing *et al*., 2013), prior to BoW construction two techniques are employed to combine the descriptors at the patch and the image level.

In the patch-based integration approach, the extracted SIFT features are concatenated with LBP features simply by adding LBP at the end of SIFT, which results in a 187 dimensional vector. This is followed by a clustering algorithm to generate $\mathrm{BoW}_{SIFT-LBP}$. The optimal size of the codebook is selected empirically as presented in the results section.

In the image-based integration approach, the BoW is constructed independently on SIFT and LBP. Thus, there is a possibility that LBP takes a smaller number of cluster centers as the number of LBP keys is smaller than the SIFT keys. To balance the importance between these two sets of features, a weighted parameter w $(0 \leq w \leq 1)$ is defined in this algorithm to compute a balanced number of cluster centers for SIFT and LBP using the following formulae:

$$\mathrm{K}_{SIFT} = w.\mathrm{K} \qquad (1)$$

$$\mathrm{K}_{LBP} = (1 - w).\mathrm{K} \qquad (2)$$

Where *K* is the size of codebook or number of cluster centers and K_{SIFT} and K_{LBP} represent the balanced number of cluster center. There are two phases in the classification process: training and testing. In the training (offline), the selected features are extracted from all the training images, and a classifier is trained on the extracted features to create a model. The selection of optimal parameters such as codebook size and classifier kernel is done during the training phase. The constructed classification model is then used to classify the query images into a pre-defined class and then extract the corresponding keywords that are assigned to the query image as an annotation. Support Vector Machines (SVMs) are chosen as a classifier based on their performance compared to other classifiers.

2.2 Classification Results and Discussion

A set of experiments were run to evaluate the performance of the classification algorithm on the ImageCLEF 2007 medical dataset. In these experiments, the classification results obtained by employing LBP, BoW, two integration approaches of $\mathrm{BoW}_{SIFT-LBP}$with various parameters are evaluated separately.

As shown in Figure 1, the best classification result obtained is 92% by integration of LBP and SIFT using an image-based approach. The classification performance is also analyzed for each individual category. In Table 1, the number of categories with high accuracy rate (> 80%) and low accuracy rate (<60%) is presented.

Feature Extraction	**Accuracy Obtained**	**# Classes (Accuracy > 80 %)**	**# Classes (80 > accuracy > 60)**	**# Classes (Accuracy < 60 %)**
LBP	90.7 %	80	7	29
BoW	90.0 %	77	7	32
$\text{BoW}_{SIFT-LBP}$ $Patch\ Based\ Integration$	91.0 %	85	3	28
$\text{BoW}_{SIFT-LBP}$ $Image\ Based\ Integration$	92.0 %	87	2	27

Table 1: Classification result on 116 classes with various feature extraction techniques.

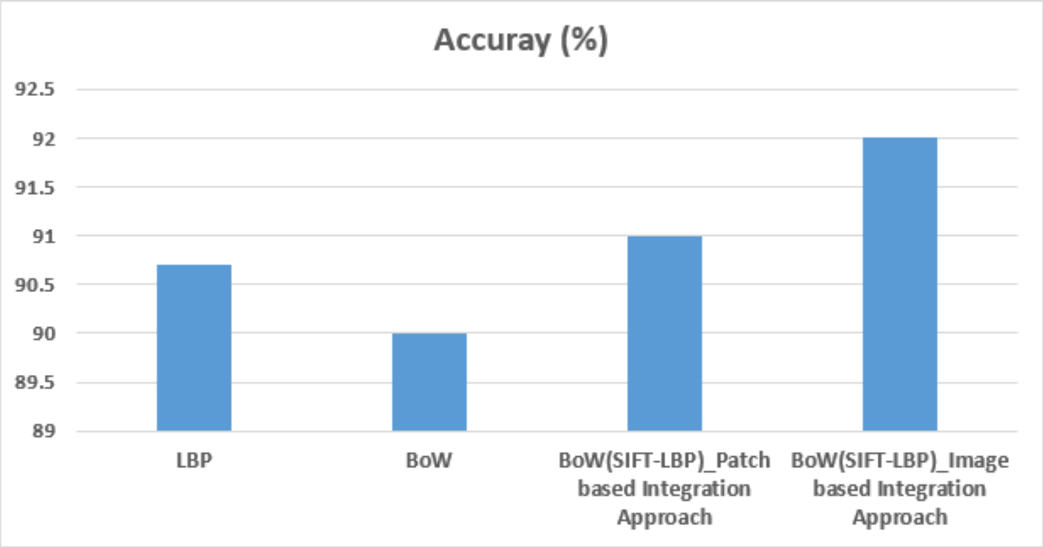


Figure 1: Classification accuracy obtained from 2 different feature extraction techniques & 2 proposed feature integration approaches.

The codebook size is an important parameter in the BoW process. To evaluate the classification performance, the optimal codebook size is chosen empirically on the training data. In this experiment, the best classification performance is achieved at K=500 as compared to other codebook sizes. The same value is used in a patch-based and an image-based process of BoW construction. Figure 2 shows the classification performance using various vocabulary sizes in the patch based integration approach.

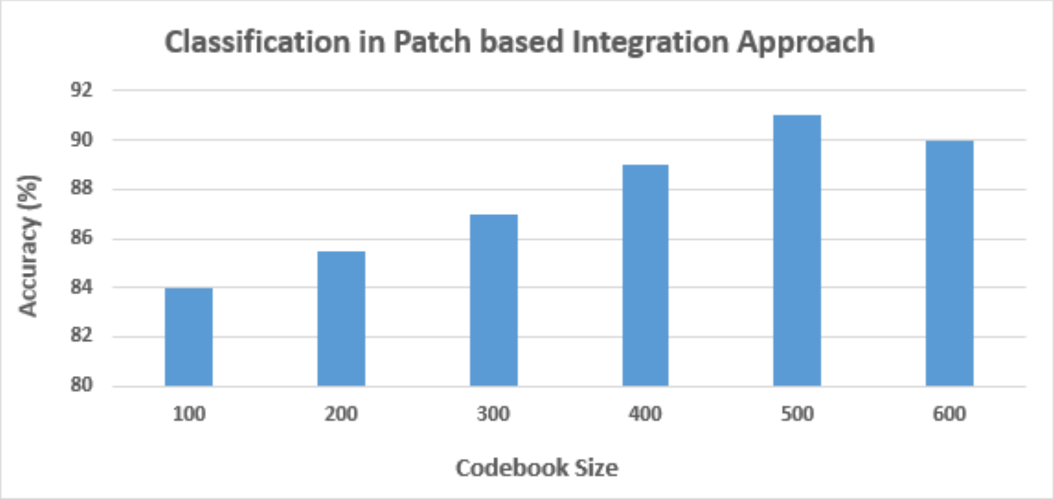


Figure 2: comparison of retrieval results obtained using a varying number of visual words.

In the image-based integration approach, possible choices of the k-means weight parameter (*w*), starting from 0.2 to 0.9, were considered and evaluated. The classification results presented in Figure 3 indicate that *w*=0.6 outperforms all the other weights on our data.

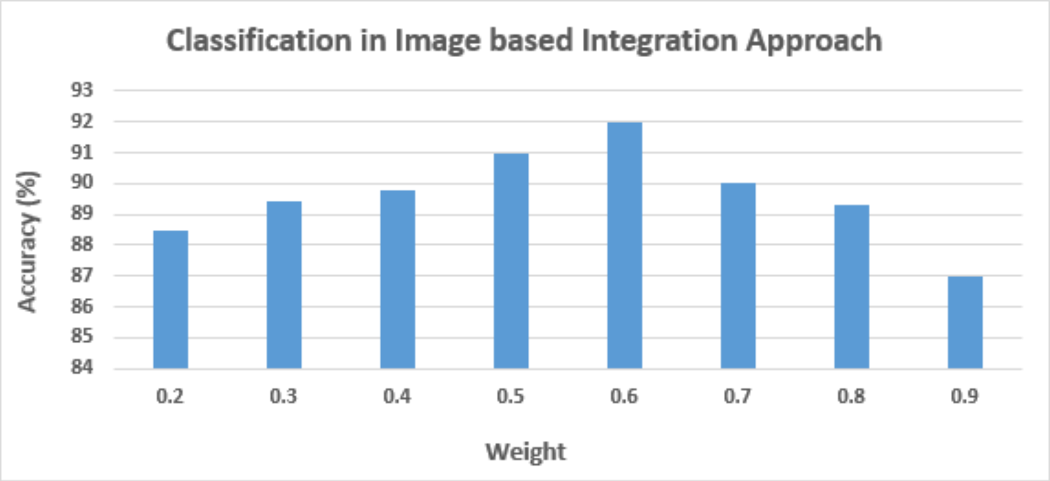


Figure 3: Comparison of retrieval results obtained with several choices of the k-means weight in the image-based integration.

As can be seen in Table 1, the number of classes with accuracy between 60% and 80% was 7 using single feature extraction. This number was reduced to 3 and 2 by employing the integration approach of the extracted features. Thus, a significant effect of the proposed integration approach is on classes with accuracy between 60 and 80% whereas there is not much difference on the number of classes with accuracy below 60%. This is mainly due to the complexity existing in medical databases such as inter-class similarities, intra-class variability and inconsistent distribution of training images.

Detailed analysis on the classification results indicates that classes with a higher number of training images generally obtain a high accuracy compared to those with a smaller number of images. The classification result represents the fact that classes with a smaller number of training images have the potential to be misclassified. Further analysis on the misclassified images indicates that they are classified within their sub-body region. For instance, “forearm”, the sub-body region of “Arm”, consists of eight distinct classes where the images are visually similar and they vary from one another in terms of the number of images in each classes. This observation shows SVMs or any other discriminative classification technique is biased to the category with a higher number of training images.

Similar to what was done in a previous study (Zare *et al*.,2013), the experimental results prove that the classification performance of databases with such complexities can be increased by exploiting a generative-discriminative approach. The unsupervised Probabilistic Latent Semantic Analysis (PLSA) (Hoffman, 2001) was employed in order to get a more stable representation of the images upon construction of $\mathrm{BoW}_{SIFT-LBP}$ $Image\ Based\ Integration$. It was then used as an input to the supervised SVM classifier to build the classification model. At the end, upon automatic classification of an unseen test image into a predefined category, the ground truth keywords representing that specific category are assigned to the new test image as an annotation.

3 IMAGE RETRIEVAL SUB-SYSTEM

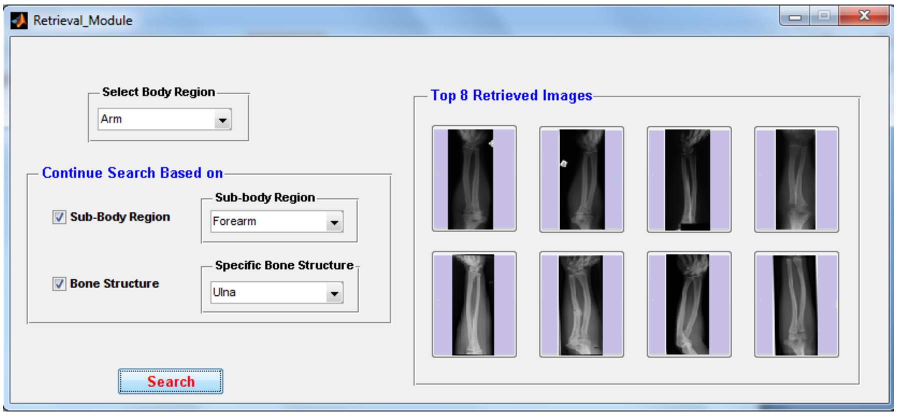
Typically, a retrieval session starts by providing a request or query. From a system perspective, there are two general categories of queries; query by example and query by text. Queries can also be multimodal, containing text and images and in a larger context case-based retrieval contains all information on an entire case (Garcia *et al*.,2015; Mourão *et al.*,2015; Jing *et al*., 2013).

3.1 Query by Text

In query by text, users provide keywords as a query that describe the information need of a clinicians. The system automatically annotates the images based on their content as discussed in the classification sub-system where medical images are inserted into pre-defined categories followed by assigning related keywords to the images. These keywords can then be used to form a query for text-based image retrieval as described below.

In the proposed retrieval system, an interface for this module is designed to enable users to filter the query from general to more specific by providing the list of body regions, corresponding sub-body regions and bone structures.

For example, if the user selects “arm”, then the system displays an option for the user to select sub-body regions; upon selecting sub-body region, then all the specific bone structure under that sub-body region will be loaded into the drop down menu as shown in Figure 4. This process assists a user to filter out his/her search to be specific in order to retrieve images related to the query.



Select Body Region: Arm

☑ Sub-Body Region — Sub-body Region: Forearm

☑ Bone Structure — Specific Bone Structure: Ulna

Figure 4: Screenshot for Image Retrieval using query by category.

3.2 Query by Image

Another type of search is when the user provides an image as a query. Upon extraction of visual features of the query image, the system allows to use the extracted features to retrieve images in the database that are visually similar to the query image.

3.3 Query by Multimodal Fusion

Fusion techniques are used to integrate visual and textual search. It is believed that such a combination can lead to better retrieval performance than single modalities. However, the challenge is to choose the right fusion technique and parameters. In general, there are two main fusion approaches: early and late fusion. In early fusion, features of different modalities (textual and visual features) are integrated into one vector before making any decision. In late fusion, the independent results obtained from various modalities (here: textual and visual) are combined to form a final retrieval result. The late fusion technique is used in this study.

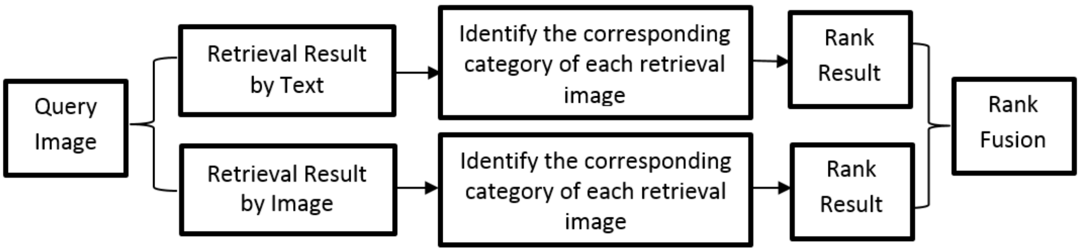

Query Image
Retrieval Result by Text
Retrieval Result by Image
Identify the corresponding category of each retrieval image
Identify the corresponding category of each retrieval image
Rank Result
Rank Result
Rank Fusion

Figure 5: General Scheme for the Proposed Multimodal Fusion Technique.

As shown in Figure 5 upon retrieval of each modality, the corresponding category for every retrieved image is identified. The result contains two sets of numbers, each consists of 10 values. Each value in every set is representing a category/class. A rank-based method in the late fusion technique is then used to compute the final retrieval result.

The results indicate that the retrieval performance for little complex categories such as intra-class variability and inter-class similarities are relatively good with both retrieval options. Subsequently the fusion techniques provide the same retrieval performance. A challenge remains for complex classes. For example, for the query by text such as “Arm, forearm, wrist joint, elbow joint, radius, ulna”, the top 10 images retrieved are scattered within the forearm sub-body region where in the retrieval module using query by image, the top 6 retrieved images are from the right class. The analysis on the retrieval results show that retrieval performance using fusion techniques is better than the results obtained from the single modality.

4 CONCLUSION

In this paper, a novel learning-based classification and retrieval system is presented for medical x-ray images from diverse medical image collections. The major module of the system is the classification part. It starts with extracting SIFT and LBP from the entire training data. Two integration approaches are then proposed on these descriptors to construct the bag of words model. Next, an SVM classifier is trained on these features to create a classification model. This model is then employed to classify the newly inserted image to the system into one of the pre-defined categories followed by assigning the ground truth annotated keywords of the category to the image. The retrieval module is exploiting fusion techniques by employing the advantages of both text-based annotation and visual retrieval approaches.

The current work has some shortcomings as the automatic classification only uses images labelled with modality, anatomy and view aspects and nothing is said on pathology. Most of the techniques are standard visual features, machine learning and fusing techniques. This combination and adaptation of the techniques is part of the novelty. The good results obtained highlight that the chosen techniques make sense for the given application.

Future work will concentrate on the evaluation of more medically relevant databases that contain pathology and not only anatomy information. It is also important to extend the retrieval to a variety of imaging modalities other than only x-rays. With an extremely large number of approaches using deep learning it is also important to compare the approach to convolutional neural networks in the future.

References

Aggarwal, P.l., Sardana, H., and Jindal, H. (2009). Content based medical image retrieval: Theory, gaps and future directions, ICGST Journal of Graphics, Vision and Image Processing, (9), 27-37.

Cao, Y., Steffey, S., He, J., Xiao, D., Tao, C., Chen, P., Müller, H. (2014). Medical Image Retrieval: A Multimodal Approach. Cancer Informatics, 13(3), 125–136.

Cao, Y., Li, Y., Müller, H., Kahn, C.E., Munson, E. (2011) Multi-modal medical image retrieval, SPIE Medical Imaging.

Dimitrovski, I., Kocev, D., Loskovska, S., & Džeroski, S. (2011). Hierarchical annotation of medical images. Pattern Recognition, 44(10–11), 2436-2449.

Garcia Seco de Herrera, A., Müller, H. (2014). Fusion Techniques in Biomedical Information Retrieval, 209–228. Springer, (2014)

Garcia Seco de Herrera, A., Markonis, D., Eggel, I., Müller, H. (2012). The medGIFT group in ImageCLEFmed 2012. In: Working Notes of CLEF 2012.

Garcia Seco de Herrera, A., Schaer, R, Markonis, D., Müller, H. (2015). Comparing fusion techniques for the ImageCLEF 2013 medical case retrieval task, Computerized Medical Imaging and Graphics, 39, 46-54.

Güld, M. O., Michael, K., Daniel, K., Henning, S., Berthold, B. W., Jörg, B., & Lehman, T.M., (2002). Quality of DICOM header Information for Image Categorization. International Symposium on Medical Imaging

Hoffman, T. (2011). Unsupervised learning by probabilistic latent semantic analysis, Machine Learning,42(1-2):177-196.

Hong, W., Lele, H., (2015). Combining visual and textual features for medical image modality classification with ℓp−norm multiple kernel learning. Neurocomputing, 147, 387-394.

Jiang, M., Zhang, S., Li, H., Metaxas, D.N. (2015) Computer-Aided Diagnosis of Mammographic Masses Using Scalable Image Retrieval. IEEE Transactions on Biomedical Engineering, , 62(2), 783-792.

Jing, Y., Zengchang, Q., Tao, W., Xi, Z. (2013). Feature integration analysis of bag-of-features model for image retrieval, Neurocomputing, 120, 355-364.

Kesorn, K., and Poslad, S. (2012). An Enhanced Bag-of-Visual Word Vector Space Model to Represent Visual Content in Athletics Images. IEEE Transaction on Multimedia, 14(1), 211-222.

Kurtz, C., Depeursinge, A., Napel, S., Beaulieu, C.F., Rubin, D. (2014). On combining image-based and ontological semantic dissimilarities for medical image retrieval applications. Medical Image Analysis, 18 (7), 1082-1100.

Lazebnik, S., Schmid, C., Ponce, J. (2006). Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. IEEE conference on computer vision and pattern recognition, Washington, DC, USA, pp 2169–2178.

Li, Y., Shi, N., Frank, D.H. (2014). Fusion analysis of information retrieval models on biomedical collections. International conference on information fusion, p. 1-8

Markonis, D., Holzer, M., Baroz, F., Ruiz De Castaneda, R.L., Boyer, C., Langs, G., Müller, H. (2015). User-oriented evaluation of a medical image retrieval system for radiologists, International Journal of Medical Informatics, 84(10), 774-783.

Markonis, D., Holzer, M., Dungs, S., Vargas, A., Langs, G., Kriewel, S., Müller, H. (2012). A survey on visual information search behavior and requirements of radiologists, Methods of information in Medicine, 51(6), 539-548.

Mourão, A., Martins, F., Magalhães, J. (2015). Multimodal medical information retrieval with unsupervised rank fusion, Computerized Medical Imaging and Graphics, 39, 35-45.

Müller, H., Deselaers, T., Kim, E., Kalpathy-Cramer, J., Deserno, TM, Clough, P., Hersh, W. (2007). Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working Notes of the 2007 CLEF Workshop.

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 24(7), 971-987.

Srinivas, M., Ramu, R., Sastry, C.S., Krishna C.M. (2015) Content based medical image retrieval using dictionary learning. Neurocomputing, 168, 880-895

Zare, M.R., Mueen, A., Woo, C.S., (2014). Automatic Medical X-ray Image Classification using Annotation. Journal of Digital Imaging (27), 77–89

Zare, M.R., Mueen, A., Awedh, M. H., Woo, C.S. (2013). Automatic Classification of Medical X-Ray Images: A Hybrid Generative-Discriminative Approach. IET Image Processing, 7(5), 523-532.