IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

# Efficient Image Retrieval approach for Large-scale Chest X Ray data using Hand-Crafted Features and Machine Learning Algorithms

**Irene Getzi S[1*], D. Christopher Durairaj[2], V Joseph Raj[3]**

[1] Dept. of Computer Science, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, India
[2] School of Computer Science, V.H.N.S.N College, (Madurai Kamaraj University), Virudhunagar, India
[3] Computer Science, Kamaraj College, (Manonmaniam Sundaranar University), Thoothukudi, India

[*]*Corresponding Author: igetzis@gmail.com, Tel.: +91-94801-44908*

**Available online at: www.ijcseonline.org**

Accepted: 06/Nov/2018, Published: 30/Nov/2018

***Abstract*—** The rapid growth in digital imaging techniques have resulted in the generation of large volume of diverse medical images. Most of these image corpus is either unlabeled or partially annotated. To ex-tract relevant information from such large-scale image corpus, it is necessary to have an efficient and scalable image retrieval techniques. In this article, we present an effective approach for retrieving images from large-scale Chest X-Ray dataset that have the similar disease conditions or severity as that of the query image. We tested our approach on NIH chest x-ray image dataset, that contains images of pneumonia affected patients. The Histogram of Gradients (HoG) features are found to give better results in classifying the disease. The dimensionality of dense HoG features is reduced by using level decomposition of Haar wavelet and using random projection. The performance degradation happened due to the feature reduction is rectified by using a hybrid approach. The proposed features are compact and capable of conveniently outperforming several existing approaches in image retrieval. To find the nearest match to the query image, the feature space is reduced further by applying k-means clustering. The implementation results are presented to test efficacy of the proposed approach.

***Keywords*—**Medical image retrieval, pneumonia detection, hand-crafted features, classification, Histogram of Gradient, feature reduction, clustering

## I. INTRODUCTION

Content-based image retrieval has been an active area of research for more than two decades. However, there is still no widely adopted method in medical domain, as the domain requires high accuracy. The medical dataset consists of highly sensitive and valuable information. It includes images taken using different modalities and diagnosis reports of the patients with various diseases.

There is an exponential growth in the generation of medical imaging data in the recent years. The massive amount of data can be used in clinical diagnosis, preventive medicine as well as a training tool for medical students. Retrieving relevant images from large and varied collections of medical image databases is a challenging and important problem.

The large repositories consist mostly of unlabeled images as it requires trained medical experts to annotate the radiological images. The manual classification of a large number of images is labour intensive, repetitive and may not always produce reliable results due to variations in the experimental conditions, image quality and human subjectivity.

Hence, the traditional image processing methods may not be suitable to deduce meaningful information from these heterogeneous, high dimensional and complex collection of medical images [1]. The processing of huge volume of data can be enhanced by framing patterns and associations for extracting the specific information from the large dataset. Therefore, machine learning and data mining techniques, and automated clustering algorithms plays an important role for efficient retrieval. The clustering approach can be used to reduce the number of comparisons while searching the image database.

In this work, we propose a two-fold approach using machine learning algorithms, for efficient retrieval of similar images from a collection of Chest X-Ray (CXR) images. The different hand-crafted texture features suitable for CXR images are identified. The scheme uses both supervised and unsupervised methods for fast retrieval from large-scale image corpus. The rest of the paper is organized as follows.

 

Section II provides the background study in this area. The proposed work is discussed in detail in section III. The implementation and results are discussed in section IV and finally concluded in section V.

## II. RELATED WORK

Chest radiography is the reference standard for the diagnosis of various thorax diseases. The different methods presented in the literature for CBIR varies with the type of features extracted, choice of feature selection or reduction, learning algorithms used for classification and clustering. Due to the complex nature of the medical images, the success of a retrieval algorithm depends on the correct features extracted. Statistical features on texture, shape-based features and local binary features are commonly used in the discrimination of Lung abnormalities. Recently pixel level deep learning features are used to produce better accuracy.

Haralick's texture features computed from Gray-level Co-occurrence Matrix (GLCM) is a commonly used technique [2, 3, 4] to detect lung abnormalities. Local binary patterns (LBP) are used in [5] to efficiently detect pneumothorax. Modified LBP descriptors, namely centre symmetric LBP extracted from local grid wavelets proposed by B. Ko, S. Kim, and J. Nam, improved image classification performance as well as computation time [6]. K. Vo, A. Sowmya, employed wavelet-based texture extraction using generalized Gaussian density to classify four diffuse lung diseases: normal, emphysema, ground glass opacity and honey-combing [7].

Hybrid approaches are proposed [8, 9] by combining different kinds of feature descriptors with the intention of enhancing the performance. Haralick features and shape descriptors extracted globally as well as from local grids are combined with pixel level features [8] to identify X Ray images of different body parts. Three different histogram-based descriptors representing color, edge, and texture features are combined to retrieve the best matches for query images in heterogeneous datasets [9].

Binary-valued features, such as BRIEF, ORB, and BRISK as described in [10], and deep features [11] are used for efficient local feature matching. Segmentation of the lung region, and dimensionality reduction techniques attempt to minimize the computation time by reducing the size of the feature descriptors.

To retrieve similar images from the corpus, the feature vector of the query image is compared with the database collection using brute force method to find the nearest match. This procedure takes a large amount of time and thus not suitable for retrieval from large-scale image corpus with high dimensional data. Hashing based methods offered better results for fast image retrieval [12], and locality sensitive hashing (LSH) methods are suitable for scalable large-scale datasets [13,14]. But, LSH based implementation requires significantly more memory. Recently, supervised learning algorithms and neural networks are being used to identify the diseases and to select the nearest match.

M. Bruijne, in [15], discussed the main challenges in approaching diagnosis with machine learning techniques. While supervised learning techniques have shown much promise in prediction, the schemes require the data to be properly annotated. Semi-supervised methods are proposed to work with unlabelled data, in which the labels are predicted from available labels [16, 17]. Unsupervised clustering-based methods are highly relevant and suitable for large-scale unannotated data. In [18], unsupervised clustering algorithms are used to extract disease specific features to classify different lung diseases from CXR images. The work of [19, 20] compares the performance of widely used clustering algorithms namely K-means, Fuzzy c-means, Medical Storage Platform for data Mining, and Homogeneity Similarity based Hierarchical clustering.

In this work, we study the best feature that characterize the pneumonia clouds. Further, the k-fold approach for classification enable us to select the suitable classifier. The feature space reduction by clustering pave way to efficient retrieval. Through experimental evaluations, it shown that the proposed method outperforms several existing hand-engineered feature extraction methods on a challenging dataset. The detailed description of the proposed work is presented in the following section.

## III. METHODOLOGY

Retrieval of relevant information from CXR images is a non-trivial task, due to the complex nature of the information in the images. Moreover, there is a great variation within a class, caused by different doses of X Ray, varying orientation, alignment and pathology. Hence, hand-crafted features are used to improve the accuracy of the image retrieval. This article proposed a three-fold approach, namely feature extraction, image classification, and clustering for efficient image retrieval. The schematic diagram of the proposed work is depicted in Figure 1.

**Feature Extraction:** Initially, for each given CXR image, features have been extracted of varied types such as texture-based statistical features, local binary features, gradient-based, wavelet-based and histogram-based features. Haralick texture features are derived from the GLCM matrix that

 

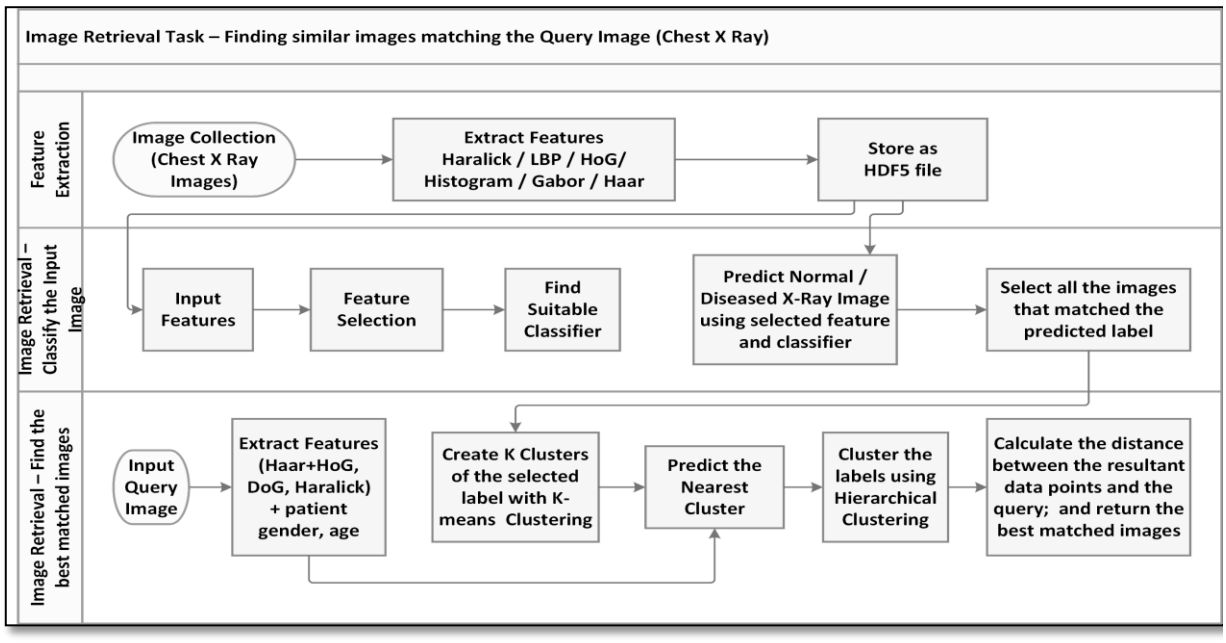


**Figure 1: Schematic diagram of finding similar images matching the query image**

records how many times two gray-level pixels adjacent to each other appear in an image. Thirteen statistical indicators such as Energy, Correlation, Inertia, Entropy, Inverse difference moment, Sum average, Sum variance, Sum entropy, Difference average, Difference variance, Difference entropy, Information measure of correlation 1 and Information measure of correlation 2 are used to quantify the texture.

LBP is an important local feature descriptor used for texture matching. Unlike Haralick texture features it computes a local representation of texture by comparing each pixel with its surrounding neighborhood of pixels.

The histogram of oriented gradients (HoG) is a gradient-based dense global feature descriptor that tries to capture the shape of structures in a region by capturing information about occurrences of gradient orientation in localized portions of an image. The Difference of Gaussians (DoG) is used to detect blobs that is to find out the regions that differ in properties compared to surrounding regions. The wavelet transform can capture and represent effectively low frequency components such as image backgrounds as well as high frequency transients such as image edges. Gabor feature is the global feature that includes more detailed information of frequency and orientation while Haar feature is used to show the variation in the pixel.

**Feature Selection and Image Classification:** Feature selection process optimizes the feature set in classifying the image. These feature selection is implemented to optimize the feature set and to achieve a better accuracy. A wrapper approach for feature selection is employed by training different classifiers in which, the discriminative power of attribute collections is evaluated based on their impact on performance of a given data classification algorithm. The extracted features are fed to different classifiers independently as well as in combination with other descriptors. Various classifiers used in this work are Linear Regression(LR), Latent Dirichlet Allocation (LDA), k Nearest Neighbor (kNN), Support Vector Machine (SVM), and Random Forests (RF). K-fold validation method with k as 10 is used for feature selection. The best classifier suitable for each descriptor is identified.

**Image Retrieval:** One of the key challenges with large-scale medical image retrieval systems, is to develop a fast solution for indexing high-dimensional image contents. Moreover, in medical domain, the dataset has a multi-level hierarchical similarity structure. Retrieval of images from the same disease class as the query image was not sufficiently accurate; the retrieved images should belong to the same subclass according to image structure, disease stage, and/or severity. In addition, the image features that work best to discriminate among different classes are different from the features needed to retrieve similar images. i.e., images belonging to the same subclass within each class. Clustering

 

is a natural choice for large-scale datasets that can find the natural groupings of the data.

When a query image is input, a classifier that uses HoG feature is employed to predict if the patient has pneumonia. HoG feature descriptor is calculated using a 5-step process as mentioned in Figure 2. The gradients in X, Y directions are extracted from normalized input image. The gradient image is divided into 8*8 cells and histogram of gradients are computed. The resultant histogram is block normalized and the feature vectors are computed.

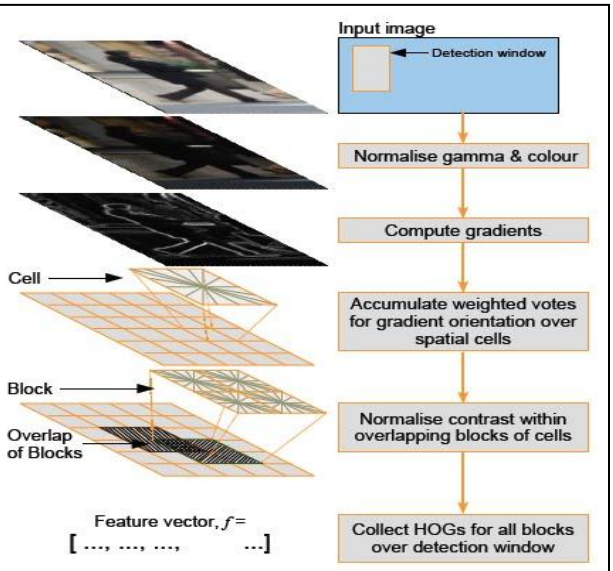


**Figure 2: Histogram of HoG Computational steps**

To retrieve the more relevant images that matched the query image, we use clustering approach. The output labels from the classifier are identified. A fusion of hand-crafted features such as the number of blobs identified using DoG local minima, Haralick features such as ASM, standard deviation, correlation, inverse difference moment and variance, along with patient age and gender are used for initial clustering with k-means clustering algorithm. The k-means algorithm aims at minimizing the objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||X_i^{(j)} - C_j||^2$$

where, $X_i^{(j)}$ is a data point, $C_i$ is the cluster center, *n* is the number of data points, k is the number of cluster, and $|| X_i^{(j)} - C_i ||^2$ is distance between a data point , $X_i^{(j)}$ and cluster centre $C_j$ .

The choice of k, depends on the feature subset as the clustering algorithms have different behaviours depending on the features of the data set. And, using a fixed number of clusters for all subspaces does not model the data in each respective subspace correctly. Thus, we need to find the number of clusters while clustering each candidate feature subset. The best cluster size is identified by using various evaluation parameters such as Homogeneity score, Silhouette Coefficient, and Davies-Bouldin Index as well as manually analyzing the output results.

A clustering result satisfies homogeneity(H) if all its clusters contain only data points which are members of a single class. The score is between 0.0 and 1.0. and 1.0 stands for perfectly homogeneous labeling. The Silhouette Coefficient (S), is calculated as follows:

$$S = (b - a) / \max(a, b)$$

where '**a**' is the mean distance between a sample and all other points in the same class and '**b**' is the mean distance between a sample and all other points in the next nearest cluster. The best score is 1.

The Davies-Bouldin Index (DB) is defined as the average similarity between each cluster $C_i$ for i=1,..,k and its most similar one Cj. The similarity is defined as a measure of $R_{ij}$ defined as

$$R_{ij} = S_i + S_j / d_{ij}$$

where $S_i$ is the average distance between each point of cluster and the centroid of that cluster, and $d_{ij}$ is the distance between cluster centroids. Now DB can be calculated as below. The DB score closer to zero indicate a better partition of clusters.

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

The query image is compared with the centroids of the k clusters, and the nearest cluster is selected. To further refine the result, the clustered labels of the predicted cluster are sent to hierarchical clustering. The data points inside the selected cluster are then compared with the features of the query image by comparing the distances using brute-force match and the top five images that produced best match are retrieved. The efficiency of the algorithm is tested against the labels created manually from the clustering dataset. The experimental results and analysis are presented in the following section.

## IV. RESULTS AND DISCUSSION

The proposed scheme has been successfully implemented using Python as a tool. The details of the implementation and the results are discussed in the following section.

The proposed scheme includes processing at three levels: feature extraction and selection, classification of the query image and finding the nearest match for classified label. The

 

**Table 1: Classification results of CXR images varying hand-crafted features and classifiers**

| | LR | LDA | KNN | RF | NB | SVM |
|---|---|---|---|---|---|---|
| **Haralick** | 89.39 | 90.52 | 90.89 | **92.01** | 87.03 | 88.27 |
| **LBP** | 84.27 | **87.52** | 85.27 | 85.90 | 79.53 | 77.66 |
| **Gabor** | 71.66 | **83.53** | 74.16 | 76.53 | 73.15 | 69.17 |
| **Histogram** | 82.03 | 81.78 | 81.90 | **94.26** | 81.90 | 65.67 |
| **All Above** | 92.76 | 91.89 | 86.52 | **94.01** | 83.27 | 70.79 |
| **HoG** | **98.00** | 96.88 | 92.51 | 94.63 | 92.51 | 84.65 |
| **LBP + Gabor** | 86.52 | **91.27** | 86.52 | 87.89 | 80.90 | 79.90 |
| **LBP + Hist** | 86.27 | 89.64 | 85.52 | **93.89** | 82.40 | 68.05 |
| **Haar + Hist** | 92.14 | 90.27 | 85.02 | **94.26** | 82.64 | 68.42 |
| **LBP + Har** | 92.76 | **94.14** | 92.14 | 93.01 | 85.65 | 87.15 |

performance of the proposed method is analyzed by conducting experiments using DICOM images from National Institute of Health (NIH) chest X-Ray dataset. The original dataset consists of X Ray images of patients affected by pneumonia, other lung diseases as well as normal images. A subset of 2000 images of pneumonia patients and normal images are used to validate the efficacy of the present work. The experiment has been implemented in Python and run on DELL system, equipped with Intel Core i5-6402P @ 2.80 GHz x 4 computer with 8GB RAM running Ubuntu 14.04.

The accuracy of image classification and clustering mainly depends on image feature extraction. More discriminated features lead to better result. The initial experiments are conducted on different classifiers by varying different kinds of features and predicted accuracy using k-fold approach are tabulated in Table 1.

The scheme uses a two-fold approach using machine learning algorithms and hand-crafted texture features. The different texture features such as Haralick, Local Binary Patterns (LBP), Histogram of Gradients (HoG) and histogram-based features are extracted from the test images. Histogram of Oriented Gradients (HoG) is found to be suitable due to the complex nature of the image.

The feature analysis of HoG is presented in Table 2. It is found that HoG descriptor is giving better classification results with highest level of accuracy over 96% with LR and LDA. Since HoG is a dense feature of large dimension with more than 80000 features, it requires several minutes to extract features from the dataset as well to train the classifier. This will scale down the performance and requires large amount of memory for large-scale datasets. To reduce the dimension and to preserve the features, Haar wavelets with level-1 decomposition and random projection were employed and first 5 components were selected. As shown in the column 3 of table 2, the accuracy level dropped drastically.

**Table 2: Feature Analysis of HoG**

| Classifier | HoG | HoG + Haar | Haar + HoG_RP | DoG + Age HoG_RP + Haralick + Gender |
|---|---|---|---|---|
| **LR** | 98.00 | 96.50 | 62.17 | 78.40 |
| **LDA** | 96.80 | 95.00 | 62.67 | 91.39 |
| **RF** | 94.63 | 93.39 | 60.55 | 92.26 |

X ray images show variations with patient's age and gender. When a combination of hand-crafted features such as DoG, Haralick, patient age and gender are fused with dimensionality reduced HoG the result improved as shown in column 4 of table 2.

To find the relevant images that describe the similar medical conditions as that of the query image, the predicted label from the classification are chosen. Initial, clustering of these labels has been done by using k-means method as the algorithm is scalable and suitable for high dimensional feature vector. The data exploration done on the dataset revealed that the dataset contains images with different disease categories as described in Figure 3.

 

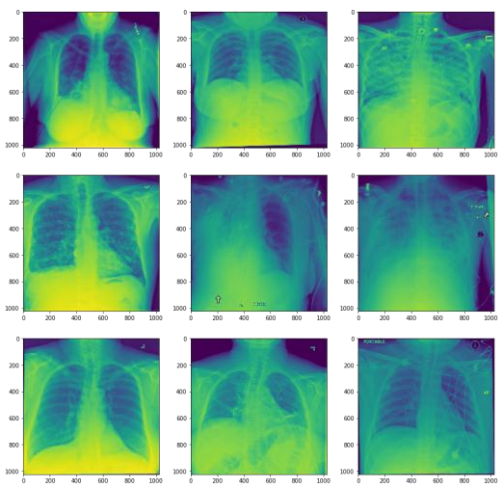

**Figure 3: Sample images of different disease categories**

The initial clustering using k-means is used to assign the query image to the nearest cluster by calculating the shortest distance between the query vector and the cluster centroid. The resultant labels are fed to hierarchical algorithm to further reduce the search space and find closer match. To find the best match, the resultant cluster labels from hierarchical clustering are compared with query vector using brute force method. The Figure 4 illustrates the result. The manual verification of the result shows 82% of accuracy. Thus, our scheme is suitable for feature extraction for large scale datasets. More analysis is required to validate the scheme on datasets that has multi-level labels based on the disease severity.

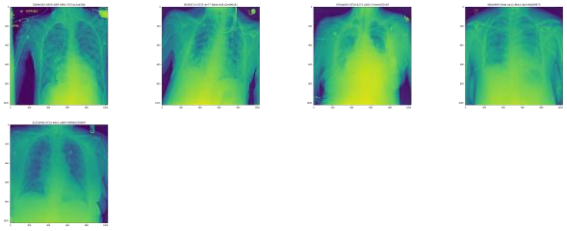

**Figure 4: The resultant images that matches the query image**

## V. CONCLUSION AND FUTURE SCOPE

A successful content-based image retrieval system must stand on two pillars: effective image-processing technique to achieve high-quality retrieval, and efficient search techniques to make the system work in real time and on a large scale. We address the scalability issue when it comes to image retrieval in large image archives in the medical domain. The proposed approach using HoG and its variants produce practical solution for large scale image datasets when combined with machine learning approaches.

## REFERENCES


[1] B. Venkataramana, L. Padmasree, M. Srinivasa Rao, G. Ganesan, K. Rama Krishna, "Implementation of Clustering Algorithms for real datasets in Medical Diagnostics using MATLAB", Journal of Soft Computing and Applications, Vol.2017, Issue.1, pp.53-66, 2017.

[2] N. Zayed, H. A. Elnemr, "Statistical Analysis of Haralick Texture Features to Discriminate Lung Abnormalities", International Journal of Biomedical Imaging, Hindawi, Volume 2015.

[3] N. R. Ratnasari, A. Susanto, Indah Soesanti, Maesadji, "Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes", In the Proceedings of 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME) Bandung, Indonesia, Nov. 2013.

[4] P. M. Gordaliza, J.J. Vaquero, S. Sharpe, M. Desco, A. M. Barrutia, "Towards an Informational Model for Tuberculosis Leision Discrimination on X-RAY CT Images", In the Proceedings of 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, April 2018.

[5] Y. Chan, Y. Zeng, H. Wu M. Wu, H. Sun, "Effective Pneumothorax Detection for Chest X-Ray Images Using Local Binary Pattern and Support Vector Machine", Journal of Healthcare Engineering, Hindawi, Volume 2018, April 2018.

[6] B. C. Ko, S. H. Kim, J. Y. Nam. "X-ray image classification using random forests with local wavelet-based CS-local binary patterns", Journal of Digital Imaging, Vol. 24, Issue. 6, pp. 1141-1151, 2011.

[7] K. T. Vo, A. Sowmya, "Diffuse lung disease classification in HRCT lung images using generalized Gaussian density modeling of wavelets coefficients", In the Proceedings of 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, November 2009.

[8] C. Ray, K. Sasmal, "A New Approach for Clustering of X-ray Images", International Journal of Computer Science Issues, Vol.7, Issue.4, No.8, July 2010.

[9] C. Reta, I. S. Moreno, J. A. C. Ceballos, "Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors", Multimedia Tools and Applications, Springer, Vol. 77, Issue. 7, pp. 8163–8193, April 2018.

[10] M. Muja, D. G. Lowe, "Fast Matching of Binary Features", In the Proceedings of Ninth Conference on Computer and Robot Vision, CRV '12, pp. 404-410, 2012.

[11] J. Ahmad, K. Muhammad, S.W. J. Baik, "Medical Image Retrieval with Compact Binary Codes Generated in Frequency Domain Using Highly Reactive Convolutional Features", Journal of Medical Systems, Vol. 42, Issue 24, February 2018.

[12] Z. Li, X. Zhang, H. Müller, S. Zhang, "Large-scale retrieval for Medical Image Analytics: A Comprehensive Review", Medical Image Analysis, Vol.43 , pp. 66–84, 2018.

[13] W. Weihong , W. Song, "A Scalable Content-based Image Retrieval Scheme using Locality-sensitive Hashing", International Conference on Computational Intelligence and Natural Computing, Wuhan, China, 2009.

[14] T. Reato, B. Demir L. Bruzzone, "Primitive cluster sensitive hashing for scalable content-based image retrieval in remote sensing archives", In the proceedings of IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, July 2017.


 

[15] M. Bruijne, "Machine Learning Approaches in Medical Image Analysis: From Detection to Diagnosis", Medical Image Analysis, Vol. 33, 94–97, 2016.

[16] Y. D. Cid et al. "Making Sense of Large Data Sets without Annotations: Analyzing Age–related Correlations from Lung CT Scans", In the Proceedings of Imaging Informatics for Healthcare, Research, and Applications, Medical Imaging, Vol. 10138, Orlando, Florida, United States, March 2017.

[17] I. E. Livieris, A. Kanavos, V. Tampakas, P. Pintelas, "An Ensemble SSL Algorithm for Efficient Chest X-Ray Image Classification", Journal of Imaging, Vol.4, Issue.7, 2018.

[18] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, A. M. Aisen, "Unsupervised Feature Selection Applied to Content-Based Retrieval of Lung Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 3, March 2003.

[19] E. M.F. E. Houby, "Medical Images Retrieval using Clustering Technique", International Journal on Recent and Innovation Trends in Computing and Communication, Vol.3, Issue.5, pp.3134-3141, October 2015.

[20] P. Premalatha, S. Subasree, "Performance Analysis of Clustering Algorithms in Medical Datasets", In the Proceedings of Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 2017.

 