Automatic Classification of Body Parts X-ray Images

Moshe Aboud[1], Assaf B. Spanier[1,2], and Leo. Joskowicz[2]

[1] Department of Software Engineering, Jerusalem College of Engineering
[2] The Selim and Rachel Benin School of Engineering, The Hebrew Univ., Jerusalem, Israel

Abstract. The development of automatic analysis and classification methods for large databases of X-ray images is a pressing need that may have a great impact on clinical practice. To advance this objective the ImageCLEF-2015 clustering of body part X-ray images challenge was created. The aim of the challenge is to group digital X-ray images into five structural groups: head-neck, upper-limb, body, lower-limb, and other. This paper presents the results of an experimental evaluation of X-ray images classification in the ImageCLEF-2015 challenge. We apply state-of-the-art classification and feature extraction methods for image classification and optimize them for the challenge task with emphasis on features indicating bone size and structure. The best classification results were obtained using the intensity, texture and HoG features and the KNN classifier. This combination has an accuracy of 86% and 73% for the 500 training images and 250 test images, respectively.

Keywords: Classification, X-ray images, ImageCLEF-2015

Source code are available at:
https://bitbucket.org/mosheab/classifying-medical-images

1 Introduction

The increasing amount of medical imaging data acquired in clinical practice constitutes a vast database of untapped diagnostically-relevant information, millions of images are acquired worldwide each year. Clinicians are struggling under the burden of diagnosis and follow up of such an immense amount of images. This phenomenon gave rise to a plethora of methods to improve and assist clinicians using efficient search capabilities.

Content-Based Image Retrieval (CBIR) is a popular growing research topic [1]. The goal of CBIR is to assist physicians with diagnosis by finding similar cases to the case at hand. Therefore, CBIR requires efficient search capabilities in a vast database of medical images. The problem is emphasized in X-ray imaging, the most widely used medical imaging modality today as many clinical home health-care centers are equipped with X-ray scanners and maintain their own database of images.

This paper elucidates the problem of classification of digital X-ray image into five groups: head-neck, upper-limb, body, lower-limb and other (Fig 1).

A variety of methods exist for medical image feature extraction and classification, Haralick et al. [14] suggest feature extraction based on gray level co-occurrences matrices, whereas Weszka et al.[23] perform a classification based on local binary patterns (LBP). Another strategy is to combine local and global features presented by Rublee et al. [18], using pixel values and shape features extracted with the Canny edge detection method. The pixel values and shape features are then used as a unique multi-feature vector used for classification.

Advanced methods include image classification based on the IRMA code [16]. In this method, features are extracted from the modality, body orientation, anatomic region and biological system. More recently, the Bag of Visual Words model (BoVW) [4] was used for X-ray images. In the BoVW approach, a visual word vocabulary is created from local image patches to represent an image, which is obtained by extracting feature descriptors around interest points.

Ghofrani et al. recently proposed the classification-based fuzzy set theory [12] They performed a fuzzy set classification with feature extraction based on a combination of shape and texture using the Canny Edge Detector and the Discrete Gabor Transform. Zare et al. [24] present three techniques for image annotation: the probabilistic latent semantic analysis (PLSA) image annotation, binary classification annotation, and annotation based on similar images. In their approach, semantic information is captured from textual and visual modalities and the correlation between them is learned.

This paper presents the results of an experimental evaluation of X-ray images classification [3] in the ImageCLEF-2015 challenge [21]. The goal of the challenge is to group digital X-ray images into five groups: head-neck, upper-limb, body, lower-limb, and other. In the context of the challenge, we apply state-of-the-art classification and feature extraction methods for image classification and optimize them for the challenge task with an emphasis on features indicating bone size and structure.

2 Method

The aim of our method is to group digital X-ray images into five groups: head-neck, upper-limb, body, lower-limb, and other (see Fig 1) . Our objective is to apply state-of-the-art classifiers and feature selection methods and to optimize them for the challenge task with an emphasis on features indicating bone size and structure.

The input to our method is a set of (1) label X-ray images from five groups, (2) features extraction techniques and; (3) classifiers. The output of our method is a combination of 10 features-classifier pairs that achieve the highest classification accuracy on the given five groups classification task.

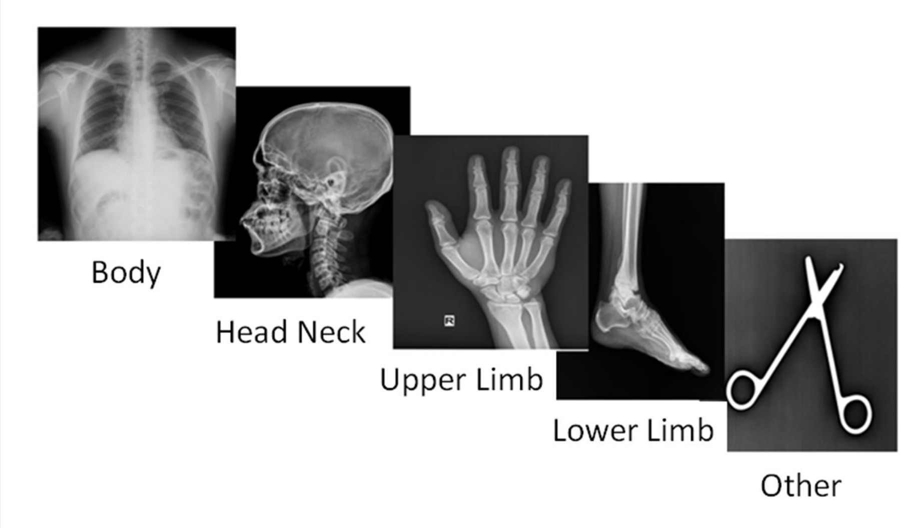

Fig. 1. Examples of the five X-ray images groups: Body, Head-Neck, Upper limb, lower-limb and other.

Our method consists of two steps. (1) a two-class experiment was used in order to select the features-classifier pair that best distinguished between X-ray images containing big and long bones (e.g. skull, arm and leg) against small and short bones (e.g. chest and abdomen bones). (2) The features-classifier pairs that that provide an average accuracy of grater then 90% in the first step evaluate on the five groups of X-ray images (head-neck, upper-limb, body, lower-limb and other Fig 1) set to find 10 best combination of features-classifier. Those 10 best features-classifier pairs were submitted to the challenge evaluation.

Fig 2 illustrates the flow of our method. Next, we describe each step in details.

2.1 Features Extraction

Nine features extraction methods were evaluated in this study. Below we describe the various features that were examined in our study.

- Color Extracting. A gray-scale histogram [22] is used to represent the color distribution of the image. We divide the image into equal patches (7x7) and compute an 8 bit histogram for each region. Then we add all patch based histogram into a single vector that serves as the color feature of our method.
- Texture Extracting. Texture features are examined using Local Binary Pattern (LBP) [13] which provides highly discriminative texture information and is used to provide robust pattern-related information. We divide each image into equal (10x10) patches and extract LBP values for each patch. The patches are then represented in a single vector to serve as the texture feature of our method.

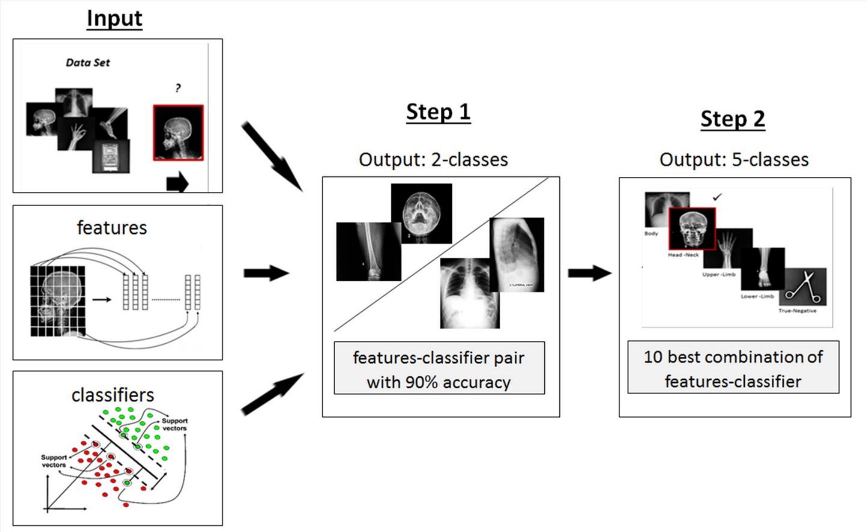

Fig. 2. The input to our method is a set of (1) labeled X-ray images (2) features extraction techniques, and (3) classifiers. Our method consists of two steps. Step 1: 2-class experiment was used in order to select the best features-classifier pair that distinguish between X-ray images containing big and long bones against small and short bones. Step 2: The features-classifier pairs that that provide an average accuracy of grater then 90% in the first step train on the five groups of X-ray images set to find 10 best combination of features-classifier. Those 10 best features-classifier pairs were submitted to the challenge evaluation.

- HoG This is a histogram of neighborhood pixels according to their gradient orientation, weighted by their gradient magnitude. HoG features were shown to be particularly discriminative of people and body shapes. We extract the HoG values for each 10x10 patch in the image. These values are then represented as a single vector to serve as the HoG feature of our method. [8].
- BoVW [9] This method produces a visual vocabulary. The method descriptors were extracted from detected key points using the following algorithms:
  - Scale invariant feature transform (SIFT) [17].
  - Speded up robust features (SURF) [5].
  - Binary robust independent elementary features Brief (BRIEF) [6].
  - Oriented fast and rotated BRIEF (ORB) [20].

  These descriptors were then clustered using the k-means algorithm. The cluster centers act as the BoVW feature of our method. Applying this scheme using the mentioned descriptor algorithms provides four additional methods. Thus, we have 4 different types of BoVW features.
- Color+Texture A combination of the color and texture values represented as as a multi-feature vector.
- Color+Texture+HoG A combination of the color, texture and HoG values represented as a multi-feature vector.

2.2 Classifiers

We tested the following four classification methods:

1. **KNN** assigns a label according to the majority labels of the K-nearest neighbor in space [7].
2. **SVM** is a linear classification of the points in space into two distinct classes [11] .
3. **LR** is a probability model that predicts a binary output based on the model predictor variables [10].
4. **DBN** constructs deep hierarchical layers based on a representation of the training data. The DBN performs an unsupervised pre-training learning and then sets the weights of the network in order to successfully use a supervised learning for classification [2] .

2.3 Model Selection

Given a set of feature extraction and classifier methods our goal is to find the 10 best combinations of feature-classifier that will provide the highest classification accuracy for the five groups of X-ray images.

To reduce the number of combinations and the complexity of the problem, a two-class selection is first applied to distinguish between X-ray images containing big and long bones (e.g., skull, arm and leg) and those with small and short bones (e.g., chest and abdomen bones). An additional motivation is to identify the features that will isolate different bone structures.

Next, we select the feature-classifier pairs that provide and average accuracy of greater than 90% in the two-class experiment and train them on the five groups of X-ray images set to find the 10 best combinations of feature-classifiers. We test each feature-classifiers pair in leave-one-out cross-validation, in which training is learned based on all cases besides a single case that is not part of the training process and used for testing.

The 10 best feature-classifier combination were then submitted to the challenge to be tested on an unlabeled test set of images released by the ImageCLEF organization [21] [3]. We use the OpenCV-Python library [15] for the feature extraction and Python scikit-Learn Machine learning tool [19] to examine the four selected classifiers.

3 Results

The data set released by ImageCLEF[21] [3] consists of 750 X-ray images: (a) 500 images were labeled images (100 images from each group) and were released for training purposes. The five image groups are Head-Neck, Body, Upper-Limb, Lower-Limb and other (Fig 1). and (b) 250 unlabeled images released for the challenge evaluation and benchmarking.

We first present the results of the 500 labeled images training set of X-ray images obtained in our two-step approach. Then, we present the results of the 250 unlabeled images as validated by the challenge organizers.

3.1 Training

In the two-class experiment to we use all four classifiers and nine sets of features, thus creating combination of $36(4*9)$ feature-classifiers. In the first step we select feature-classifier pairs that provide an average accuracy greater than 90%. This selection reduced the number of combination from 36 to 16. The results for all 36 feature-classifiers are shown in Table 1:

Table 1. Results of the first training step: Classify between long and short bones, the nine sets of features were used, each represented as a row in the table. The columns represent the four classifiers. Values in the table are the results of the leave-one-out cross-validation process.

Feature/Classifier	LR	DBN	SVM	KNN
BoVW BRIEF	79.55%	74.06%	80.05%	75.81%
BoVW ORB	79.05%	74.06%	80.30%	78%
BoVW SIFT	82.54%	74.06%	81.30%	76.81%
BoVW SURF	87%	74.31%	87.28%	84.29%
HoG	90.77%	79.55%	92.52%	93.02%
TEXTURE	89.78%	91.27%	92.27%	90.52%
COLOR	89.28%	92.77%	91.02%	92.52%
COLOR+TEXTURE	91.27%	93.27%	91.52%	91.77%
COLOR+TEXTURE+HoG	93.52%	92.77%	92.27%	92%

The BoVW and HoG features exhibit low accuracy regardless of the classifier tested. The combination of color and texture yield high accuracy rate of 89-93%. Using color, texture and HoG features all together yields the highest average classification accuracy in all classifiers.

In the second step, 5-class, classification was preformed on all five group. We select 16 combinations of features-classifier that yield an accuracy greater than 90%. The goal of this second step is to investigate the performance of the methods and to reduce the number of feature-classifiers to the best 10 .

Table 2 presents the results of the second step on the training set.

Table 2. Results of the second training phase for the 5-groups training set classification, the four sets of features were used, each represented as a row in the table. The columns represent the four classifiers. The 10 best combinations are marked in the table with (*) and were sent for the challenge

Feature/Classifier	LR	DBN	SVM	KNN
TEXTURE	73.40%	77.40%	75.00%	78.80(*)%
COLOR	72.40%	79.20%	74.80%	80.80(*)%
COLOR+TEXTURE	75.60(*)%	80.20(*)%	79.80(*)%	83.20(*)%
COLOR+TEXTURE+HoG	80.80(*)%	83.40(*)%	83.80(*)%	86(*)%

To sum-up, the 10 best features-classifier pairs are: **1.** Color+Texture+HoG and KNN **2.** Color+Texture+HoG and SVM **3.** Color+Texture + HoG and DBN **4.** Color+Texture+HoG and LR **5.** Texture+HoG and KNN **6.** Texture+HoG and SVM **7.** Texture+HoG and DBN **8.** Color+Texture and LR **9.** Color and KNN **10.** Texture and KNN. These 10 best combinations are marked in Table 2 with (*) were sent to the challenge organizer for evaluation.

3.2 Challenge Results

Table 3 shows the classification results of the 250 training images with the 10 best combinations of feature-classifier methods which were submitted to the ImageCLEF organization for evaluation. Our best summation ranked 12th out of 30 using the color histogram features and the KNN classifier achieving an accuracy of 73.2%. The challenge results for all 10 methods are presented in the table below:

Table 3. Challenge Results for 250 training images with our 10 best combinations features-classifier methods

Feature/Classifier	LR	DBN	SVM	KNN
TEXTURE				66.4%
COLOR				73.2%
COLOR+TEXTURE	71.2%	68.0%	71.2%	71.2%
COLOR+TEXTURE+HoG	69.2%	69.2%	72.8%	72.4%

Note that our best submission (73.2% accuracy) is lower than the best accuracy obtained in the five group training set (80-85%) using the same combination of color texture and HoG features. The challenge results are similar to the average results achieved on the training set. This may indicate that the variability of the training dataset does not fully reflect the images variability of the challenge dataset.

4 Discussion

In this work we have evaluated four state-of-the-art classifier and nine sets of features, resulting in 36 combinations of feature-classifiers. Surprisingly, the simplest classifier, in terms of implementation and computational complexity, KNN, exhibited the best results. Moreover, despite the major trend of using Deep Belief Network (DBN) methods for many image-based classification problems, the use of DBN in our study exhibited reliable results only when applied to a large-scale dataset. However, suboptimal results were obtained when used on small-scale datasets (see Table 2). This is in line with the theory of DBN, which requires large-scale databases for reliable performance. The BoVW method was the least efficient method among all feature extraction schemes that have been tested.

From all feature extraction methods we evaluated, the color feature yielded the highest accuracy. This is surprising, considering that X-ray images are gray-level based images. This could be explained by the algorithm implementation, which extracts the gray-scale histogram features from different regions of the image and provides more specific information and a better perspective on the distribution of the color. Another advantage of using the color features is low computational complexity as compared to the texture, HoG and BoVW.

5 Conclusions

This paper presents research on medical X-ray image classification. We analyze state-of-the-art classifiers and feature extraction methods for image classification. The image features that have been used include the color, texture, HoG and BoVW, which were used by our tested classifiers: SVM, KNN, LR and DBN. We used the datasets of the ImageCLEF-2015 [19] clustering of body part X-ray challenge [3]: 500 X-ray images were used for training and 250 for testing. The highest classification accuracy results were obtained when using the intensity, texture and HoG features and the KNN classifier. This combination has an accuracy of 86% and 73% for the 500 training images and 250 test images, respectively.

Future work consists of examining an additional set of classifiers and extending the completeness of our algorithm to estimate the partitioning of the initial clusters into sub-clusters. For example, the upper-limb cluster can be further divided into the following categories: clavicle, scapula, humerus, radius, ulna and hand.

Future work consists of examining an additional set of classifiers and extending the completeness of our algorithm to estimate the partitioning of the initial clusters into sub-clusters, for example the upper-limb cluster can be farther divided into: Clavicle, Scapula, Humerus, Radius, Ulna, and Hand.

References

1. Akgül, C.B., Rubin, D.L., Napel, S., Beaulieu, C.F., Greenspan, H., Acar, B.: Content-based image retrieval in radiology: Current status and future directions. Journal of Digital Imaging 24(2), 208–222 (2011)
2. Ali, K.H., Wang, T.: Learning features for action recognition and identity with deep belief networks. In: 2014 International Conference on Audio, Language and Image Processing (ICALIP),. pp. 129–132. IEEE (2014)
3. Amin, M.A., Mohammed, M.K.: Overview of the ImageCLEF 2015 medical clustering task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Toulouse, France (September 8-11 2015)
4. Avni, U., Goldberger, J., Sharon, M., Konen, E., Greenspan, H.: Chest x-ray characterization: from organ identification to pathology categorization. In: Proceedings of the international conference on Multimedia information retrieval. pp. 155–164. ACM (2010)

5. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding 110(3), 346–359 (2008)
6. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: Binary robust independent elementary features. In: Computer Vision–ECCV 2010, pp. 778–792. Springer (2010)
7. Cunningham, P., Delany, S.J.: k-nearest neighbour classifiers. Multiple Classifier Systems pp. 1–17 (2007)
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005. vol. 1, pp. 886–893 (2005)
9. Deselaers, T., Pimenidis, L., Ney, H.: Bag-of-visual-words models for adult image classification and filtering. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. pp. 1–4 (2008)
10. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics 35(5), 352–359 (2002)
11. Fan, Y., Shen, D., Davatzikos, C.: Classification of structural images via high-dimensional image warping, robust feature extraction, and svm. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005, pp. 1–8. Springer (2005)
12. Ghofrani, F., Helfroush, M.S., Rashidpour, M., Kazemi, K.: Fuzzy-based medical x-ray image classification. Journal of medical signals and sensors 2(2), 73 (2012)
13. Guo, Z., Zhang, L., Zhang, D.: Rotation invariant texture classification using lbp variance (lbpv) with global matching. Pattern recognition 43(3), 706–719 (2010)
14. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics (6), 610–621 (1973)
15. Howse, J.: OpenCV Computer Vision with Python. Packt Publishing Ltd (2013)
16. Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B.: The irma code for unique classification of medical images. In: Medical Imaging 2003. pp. 440–451. International Society for Optics and Photonics (2003)
17. Liu, X., Shao, Z., Liu, J.: Ontology-based image retrieval with sift features. In: First International Conference on Pervasive Computing Signal Processing and Applications (PCSPA). pp. 464–467. IEEE (2010)
18. Mueen, A., Zainuddin, R., Baba, M.S.: Automatic multilevel medical image annotation and retrieval. Journal of digital imaging 21(3), 290–295 (2008)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. The Journal of Machine Learning Research 12, 2825–2830 (2011)
20. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an efficient alternative to sift or surf. In: IEEE International Conference on Computer Vision (ICCV). pp. 2564–2571 (2011)
21. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science, Springer International Publishing (2015)
22. Wang, S.L., Liew, A.: Information-based color feature representation for image classification. In: International Conference on Image Processing (ICIP). vol. 6, pp. VI–353 (2007)

23. Weszka, J.S., Dyer, C.R., Rosenfeld, A.: A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man and Cybernetics (4), 269–285 (1976)
24. Zare, M.R., Mueen, A., Seng, W.C.: Automatic medical x-ray image classification using annotation. Journal of digital imaging 27(1), 77–89 (2014)