Biomedical Image Classification with Random Subwindows and Decision Trees

Raphaël Marée, Pierre Geurts, Justus Piater, and Louis Wehenkel

GIGA Bioinformatics Platform / CBIG, Department of EE & CS,
Institut Montefiore, University of Liège, B-4000 Liège, Belgium

Abstract. In this paper, we address a problem of biomedical image classification that involves the automatic classification of x-ray images in 57 predefined classes with large intra-class variability. To achieve that goal, we apply and slightly adapt a recent generic method for image classification based on ensemble of decision trees and random subwindows. We obtain classification results close to the state of the art on a publicly available database of 10000 x-ray images. We also provide some clues to interpret the classification of each image in terms of subwindow relevance.

1 Introduction

Image classification is an important problem which appears in many application domains. Manual classification of images is time-consuming, repetitive, and could not always be considered reliable. Therefore, there is an important need for automatic image classification tools. Given a set of training images labelled into a finite number of classes, the goal of an automatic image classification method is to build a model that will be able to predict accurately the class of new, unseen images.

In biomedical applications, such automatic techniques could help to organize large-scale image databases into image categories before further retrieval or diagnostic [LGD+05]. A class could for example denote a code [LSK+03] corresponding to an imaging modality and direction, a body part, and a biological system examined, in order to organize images in a general way without limitation to a specific diagnostic study. The annotation of such images is usually done automatically by medical equipments and/or manually adapted by physicians or radiologists. However, [GKK+02] has examined reliability of the encoded information from past clinical routines and the authors observed that some entries are missing, or are false, or do not describe the anatomic region precisely. Automatic image classification systems are thus an important, complementary, first step in medical imaging.

Image classification methods have also been proposed to setup clinical diagnosis tools based on functional magnetic resonance [ZST+05,LTC+04] or optical tomography [BH05] images. Another interesting application is the study of the

International Workshop on Computer Vision for Biomedical Image Applications: Current Techniques and Future Trends (at ICCV 2005).

phenotypic effects of drugs in human cells [KDN01] where a class could for example denote stomatocytes, discocytes, or echinocytes. It is also desirable to setup high-throughput cell phenotype screening [CEW+04] where the goal of an automatic image classification method could be to identify classes of subcellular phenotypes, for example cytoplasm, mitochondria, nucleoli, ... In histological image classification [ZCC05], a class could represent a tissue from an organ and part of the body: pancreas, lung, thyroid, ...

1.1 Related work

Till recently, image classification systems usually relied on a pre-processing step, specific to the particular problem and application domain, which aims at extracting a reduced set of "interesting" features from the initially huge number of pixels. This reduced set is then used as new input variables (or "signatures") for traditional learning algorithms (for example a nearest neighbor or neural network classifiers), possibly tuned for the specific application ([GMT+96], [AZC01], [ALTS03]).

The limitation of this approach is clear: When considering a new application or when new image classes appear, it is often necessary to manually adapt the pre-processing step by taking into account the specific characteristics of the new task. However, a more recent trend is to consider combining several different types of features that describe different aspects of an image. For example, in [LGD+05], image recognition rates are improved by combining global texture measures and local pixel neighborhood information. In [CEW+04], 448 different image features are extracted corresponding to textures descriptions, intensity distributions, edges, ... Other recent computer vision studies [MTS+05] suggest that current feature detectors are complementary (some being more adapted to structured scenes while others to textures) and that all of them should ideally be used in parallel, what would likely increase robustness to different types of image transformations.

In [MGPW05b], we have proposed a generic approach to image classification. Indeed, as we generally don't know in advance what is useful in images to classify them, we proposed to describe images by the combination of a large number of square patches randomly extracted from images ("Random Subwindows"). This process has the advantage to provide a rich representation of images corresponding to various overlapping regions both local and global, whatever the task and content of images. Moreover, to avoid discarding useful information and to be able to classify a large number of classes, we proposed to use a highly informative representation that is basically the pixel values of these subwindows. This representation is also normalized to improve robustness to scale changes. To handle this high-dimensional data and to extract useful information, we rely on recent tree-based machine learning ensemble methods [MGPW05a]. These methods are indeed able to handle more and more complex problems (high-dimensional data) without requiring any a priori information about the application. This approach has been evaluated on various image classification datasets involving the classification of digits, faces, objects, buildings, photographs, ...

1.2 This work

In this paper, we propose to apply and slightly adapt our general approach [MGPW05b] to a specific biomedical application: the classification of a database of 10000 grey-level x-ray images into 57 categories. [1] For image classification methods, this is a very challenging task because of the intra-class variability of such images. But thanks to the generality of the approach, only some minor adaptations have been required to tackle specific issues of this dataset and to obtain results comparable to the state of the art. In this paper, we also propose an interpretation of the classification results by focusing on the test subwindows that contribute to the classification of one image.

The biomedical dataset we used is described in Section 2. The main steps of the approach and its adaptation are described in Section 3. Our empirical results, and the interpretation of the results are given in Section 4.

2 IRMA X-ray Dataset

The IRMA dataset contains 10000 anonymous x-ray images, which have been arbitrarly selected from clinical routine at Aachen University of Technology Hospital (RWTH), Germany. These images were acquired using different imaging techniques and modalities (plain radiography, fluoroscopy, angiography) with different relative directions of the device and the patient (coronal, sagittal, axial, other). They represent different anatomic body parts (cranium, spine, arm, chest, abdomen, leg, pelvis, breast, hand, ...) and biological systems (musculoscetal, uropoietic, gastrointestinal, reproductive, cardiovascular) of patients of various ages, genders, and pathologies. All images are in grey levels and were downscaled to fit into a 512×512 bounding box maintaining the original aspect ratio. All images were classified according to the IRMA code [LSK+03]. Based on this code, 57 classes were defined. As mentionned by [PG04], in addition to natural variations between different patients, the intra-class variability is high for that kind of task, caused in particular by varying orientation and alignment, and/or by the presence of cloths, jewels, artificial-implants and medical instruments, and/or because images are characterized with contrast variation, non-uniform intensity background and various sources of noise. The task is thus non-trivial for image classification methods. Figure 2 exhibits some images of the dataset with intra-class variability, but also images from different classes that may look similar for non-experts.

3 Method

In this section, we briefly describe the framework that we proposed in [MGPW05b] with an emphasis on minor changes for the purpose of the task considered in this paper.

[1] IRMA database courtesy of TM Lehmann, Dept. of Medical Informatics, RWTH Aachen, Germany, `http://www.irma-project.org`.

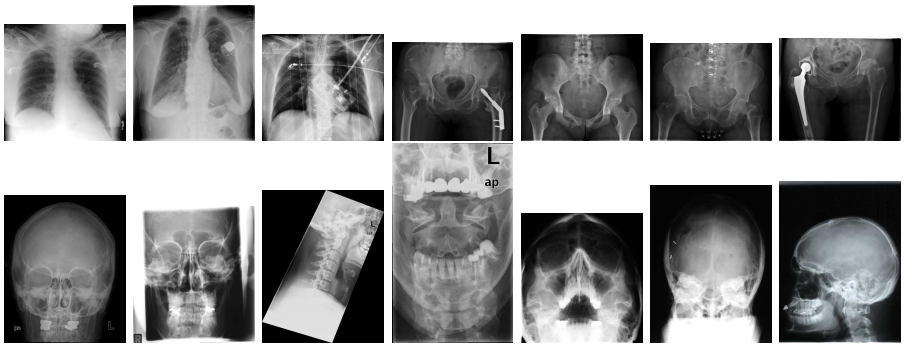

Fig. 1. Some plain radiographies. On the first line, the three first images are from the same class (chest), the four following ones are from another class (pelvis). On the second line, all images are from 7 different classes related to cranium or cervical spine. Note that all these images are correclty classified by our method.

During the training phase, subwindows are randomly extracted from training images (3.1), and a model is constructed by machine learning (3.2). Classification of a new test image (3.3) similarly entails extraction and description of subwindows, and the application of the learned model to these subwindows. Aggregation of subwindow predictions is then performed to classify the test image, as illustrated in Figure 2.

3.1 Subwindows

The method extracts a large number (N_w) of possibly overlapping, square subwindows of random sizes and at random positions from training images. Subwindows are resized to a fixed scale (16×16 pixels) and encoded by their grey-value pixels. Each subwindow is thus described by a feature vector of 256 integer values. The same random process and descriptors are used for test images.

For the specific x-ray task, the method has to cope with a learning dataset with very unbalanced class distributions. For example, a "chest" class is represented by more than 2500 images while several other classes have less than 20 images. For example, an "abdomen" class has only 9 training images. In order to have an equal number of subwindows in each class, we extract from each training image a number of subwindows inversely proportional to the number of images in its class. More precisely, from each training image of class c, we extract $N_w/(m * nb_c)$ subwindows where m is the number of classes and nb_c the number of training images of class c. For testing, we extract a fixed number of subwindows in each image, $N_{w,test}$, as in the original method.

3.2 Learning

At the learning phase, a model is automatically built using subwindows extracted from training images. First, each subwindow is labelled with the class

of its parent image. Then, any supervised machine learning algorithm can be applied to build a subwindow classification model. Here, the input of a machine learning algorithm is thus a training sample of N_w subwindows, each of which is described by 256 integer input variables and a discrete output class. The learning algorithm should consequently be able to deal efficiently with a large amount of data, first in terms of the number of subwindows and classes of images in the training set, but more importantly in terms of the number of values describing these subwindows. In this context, and following our previous comparative study [MGPW05a], we use the Tree Boosting [FRS96] and Extra-Trees [GEW05] algorithms. Their advantages are the computational efficiency (especially Extra-Trees) and their good accuracy.

3.3 Recognition

In this approach, the learned model is used to classify subwindows of a test image. To make a prediction for a test image with an ensemble of trees grown from subwindows, each subwindow is simply propagated into each tree of the ensemble. Each tree outputs conditional class probability estimates for each subwindow. Each subwindow thus receives T class probability estimate vectors where T denotes the number of trees in the ensemble. All the predictions are then averaged and the class corresponding to the largest aggregated probability estimate is assigned to the image.

4 Experiments

4.1 Protocol and Method Parameters

To be able to directly compare our results with other methods, we used the standard protocol defined in the context of the ImageCLEF 2005 Automatic Annotation Task [iCS05]. The learning set is composed of 9000 images and the remaining 1000 images are used for testing.

The parameters of our method were fixed to $N_w = 800000$ learning subwindows (which corresponds approximately to 14000 subwindows per class), $T = 25$ trees, and $N_{w,test} = 500$ subwindows are randomly extracted from each test image. To minimize contrast variations between images, we have applied the contrast enhancement technique of ImageMagick[2] (`-normalize` option) to each image which transforms it to span the full range of grey values.

For each machine learning method within the framework, the values of several parameters need to be fixed.

Extra-Trees With Extra-Trees, the only parameter is the number K of attributes randomly selected at each test node. To fix its value, we used an internal cross validation in the learning sample to determine the best of $K \in \{1, 16, 128, 256\}$. The best result was obtained with $K = 256$ and then we used this value to build a model on the whole learning set.

[2] `http://www.imagemagick.org/`

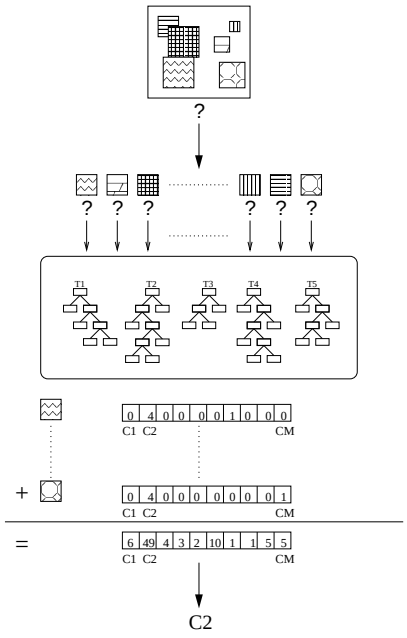

Fig. 2. Recognition: randomly-extracted subwindows are propagated through the trees (here $T = 5$). Votes are aggregated and the majority class is assigned to the image.

Tree Boosting Boosting requires that the learning algorithm does not give perfect models on the learning sample (so as to provide some misclassified instances). Hence, with this method, we used with decision trees the stop-splitting criterion described by [Weh97]. It uses a hypothesis test based on the G^2 statistic to determine the significance of a test. In our experiments, we fixed the nondetection risk α to 0.005, as in our previous study for object recognition [MGPW05a].

4.2 Accuracy results

Our results using the standard protocol are reported in Table 1 in terms of misclassification error rates on the independent test set, as well as several results obtained by 11 other research units. [3] Among the 46 methods evaluated on this dataset, a majority of them yield more than 20% error rate where each 0.1% corresponds to 1 misclassification. The Random Subwindows combined with Tree Boosting method yields 140 errors among 1000 test images (14%) and compares very well with the best result on this dataset (12.6%). Combining Random Subwindows with Extra-Trees also yields good results with 14.7% error

[3] All results are available on `http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef05_aat_results.html`.

rate. Table 2 also gives results when the correct class occurs among the first r classes, and we observe that the error rate could be reduced downto 6.6% with Tree Boosting if the correct class is only required to be within the top three classes.

Method	*error rate*
1-NN + IDM [KGN04]	12.6%
1-NN + CCF + IDM + Tamura	13.3%
Discriminative patches [DKN05]	13.9%
Random Subwindows + Tree Boosting	14.0 %
MI1 Confidence	14.6%
Random Subwindows + Extra-Trees	14.7%
Gift 5NN8g	20.6%
...	...
Nearest Neighbor, 32×32, Euclidian	36.8%
...	...
Texture directionality	73.3%

Table 1. Error rates on IRMA dataset (our results *italic*) [iCS05].

Method	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$r = 5$
RSw + Tree Boosting	14.0 %	9.0%	6.6%	6.5%	6.5%
RSw + Extra-Trees	14.7%	9.4%	7.0%	7.0%	7.0%

Table 2. Error rates for rank $r = 1, ..., 5$

4.3 Computational efficiency

Even though our current implementation has not been fully optimized, some indications can be given about computational requirements. For example, on a standard 2.4Ghz computer it took us about 75 hours to build a tree ensemble by boosting, and 18 hours to build an ensemble of 25 Extra-Trees (without counting the pre-processing tasks [4] such as subwindow sampling, extraction and normalization). With these models, the CPU time needed to classify a single new image was of about 1.125s (without taking into account the time needed to extract the 500 subwindows).

Notice that the computational complexity of the training algorithm is on the order of $TN_w \log N_w$, and that of the testing stage is essentially proportional to $TN_{w,test} \log N_w$. These numbers could thus be adjusted in order to comply with the desired requirements. For example, with Tree Boosting, using $N_{w,test} = 100$ instead of 500 reduced classification time to 0.225s, at the price of a negligible increase in error rate (which increased from 14% to 14.1%).

[4] For the pre-processing tasks, we used ImageMagick.

4.4 Relevant Subwindows

Beyond misclassification error rates, it could be interesting to observe how the proposed method classifies images, or in other words which subwindows contribute to the correct classification of one image. As mentionned before and illustrated by Figure 2, for one test image, each subwindow is propagated into the T trees of the ensemble and thus receives T votes. Consequently, we have for each subwindow the distribution of votes for all classes. The subwindows that receive the highest number of votes for a given class can then be considered as the most specific ones for that class and their visualization on the top of the image can bring potentially useful information about that class.

This functionality could be very helpful if the goal of the biomedical image classication task is for example to detect and classify diseased regions. The most relevant regions for a given class could then be shown to experts for further analysis.

In Figure 3, we simply provide some examples of subwindows receiving a high number of correct votes when using Random Subwindows combined with Tree Boosting. We observe that both small and large regions are among the best classified subwindows of many test images. It seems to confirm that it is not straigthforward to determine in advance what is useful in images to classify them and thus both local and global regions should be used by image classification methods.

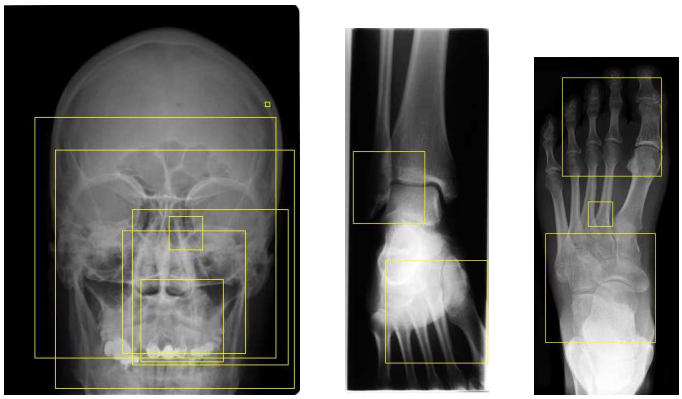

Fig. 3. Subwindows with the highest number of correct votes for three test images (from classes cranium, ankle joint, and foot).

5 Conclusions

In this paper we evaluated the applicability of our image classification method based on ensemble of decision trees and random subwindows [MGPW05b], for a specific biomedical task: x-ray image classification. We obtained results comparable to the state of the art with around 14% error rate on a 57 class dataset

involving large intra-class variability and very unbalanced class distributions. For the task of retrieving the correct class within the third best classes, our method yields 6.6% error rate. We also provide a novel simple way to understand how the method classifies images which may be of high interest for biomedical applications.

Our results confirm the potential of the approach for a wide range of applications. Many biomedical applications could benefit from this approach especially since it is directly applicable without tedious adaptation. We plan to apply the method to other biomedical applications as soon as such image datasets will be publicly available. Possible applications are for example the classification of pharmaceutical powders, human cells, histological tissues, ...

6 Acknowledgments

Raphaël Marée is supported by the GIGA-Interdisciplinary Cluster for Applied Genoproteomics, hosted by the University of Liège. Pierre Geurts is a Postdoctoral Researcher at the National Fund for Scientific Research (FNRS, Belgium). IRMA database courtesy of TM Lehmann, Dept. of Medical Informatics, RWTH Aachen, Germany.

References

[ALTS03] S. Antani, LR. Long, G. Thoma, and RJ. Stanley. Vertebra shape classification using mlp for content-based image retrieval. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pages 160–165, July 2003.

[AZC01] M.-L. Antonie, O. R. Zaïane, and A. Coman. Application of data mining techniques for medical image classification. In *Proc. of Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001) in conjunction with Seventh ACM SIGKDD*, pages 94–101, August 2001.

[BH05] V. Balasubramanyam and A. H. Hielscher. Classification of optical tomographic images of rheumatoid finger joints with support vector machines. In *Proc. SPIE Advanced Biomedical and Clinical Diagnostic Systems III*, volume 5692, pages 37–43, April 2005.

[CEW+04] C. Conrad, H. Erfle, P. Warnat, N. Daigle, T. Lörch, J. Ellenberg, R. Pepperkok, and R. Eils. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research*, 14:1130–1136, 2004.

[DKN05] T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using image patches. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 157–162, June 2005.

[FRS96] Y. Freund and E. Robert Schapire. Experiments with a new boosting algorithm. In *Proc. Thirteenth International Conference on Machine Learning*, pages 148–156, 1996.

[GEW05] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Submitted*, 2005.

[GKK+02] MO. Güld, M. Kohnen, D. Keysers, H. Schubert, BB. Wein, J. Bredno, and TM. Lehmann. Quality of dicom header information for image categorization. In *Proceedings SPIE*, pages 280–287, 2002.

[GMT+96] M. Goldbaum, S. Moezzi, A. Taylor, S. Chatterjee, J. Boyd, E. Hunter, and R. Jain. Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. In *Proc. IEEE International Conference on Image Processing (ICIP)*, volume 3, pages 695–698, 1996.

[iCS05] Springer Lecture Notes in Computer Science, editor. *Proc. of Cross Language Evaluation Forum (CLEF), to appear*, 2005.

[KDN01] D. Keysers, J. Dahmen, and H. Ney. Invariant classification of red blood cells. In *Proc. Bildverarbeitung für die Medezin (BVM)*, pages 367–371, March 2001.

[KGN04] D. Keysers, C. Gollan, and H. Ney. Classification of medical images using non-linear distortion models. In *Bildverarbeitung für die Medizin (BVM)*, pages 366–370, March 2004.

[LGD+05] TM. Lehmann, MO. Güld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and BB Wein. Automatic categorization of medical images for content-based retrieval and data mining. *Computerized Medical Imaging and Graphics*, 29(2):143–155, 2005.

[LSK+03] TM. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and BB. Wein. The irma code for unique classification of medical images. In *SPIE*, volume 5033, pages 109–117, 2003.

[LTC+04] Y. Liu, L. Teverovskiy, O. Carmichael, R. Kikinis, M. Shenton, C.S. Carter, V.A. Stenger, S. Davis, H. Aizenstein, J. Becker, O. Lopez, , and C. Meltzer. Discriminative MR image feature analysis for automatic schizophrenia and alzheimer's disease classification. In *Proc. of the 7th International Conference on Medical Image Computing and Computer Aided Intervention (MICCAI '04)*, pages 393–401, October 2004.

[MGPW05a] R. Marée, P. Geurts, J. Piater, and L. Wehenkel. Decision trees and random subwindows for object recognition. In *ICML workshop on Machine Learning Techniques for Processing Multimedia Content (MLMM2005)*, 2005.

[MGPW05b] R. Marée, P. Geurts, J. Piater, and L. Wehenkel. Random subwindows for robust image classification. In Cordelia Schmid, Stefano Soatto, and Carlo Tomasi, editors, *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 34–40. IEEE, June 2005.

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision, to appear*, 2005.

[PG04] A. Pinhas and H. Greenspan. A continuous and probabilistic framework for medical image representation and categorization. In *Proc. of SPIE Medical Imaging*, volume 5371, 2004.

[Weh97] L. Wehenkel. *Automatic Learning Techniques in Power Systems*. Kluwer Academic Publishers, Boston, November 1997.

[ZCC05] D. Zhao, Y. Chen, and H. Correa. Statistical categorization of human histological images. In *Proc. of Internation Conference on Image Processing (ICIP)*, 2005.

[ZST+05] L. Zhang, D. Samaras, D. Tomasi, N. Volkow, , and R. Goldstein. Machine learning for clinical diagnosis from functional magnetic resonance imaging. In *Proc. International Conference on Computer Vision (CVPR)*, volume 2, pages 1211–1217, 2005.