STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LX**, Number 1, 2015

# DESCRIPTORS FUSION AND GENETIC PROGRAMMING FOR BREAST CANCER DETECTION

ŞTEFANA FRĂTEAN AND LAURA DIOŞAN

Abstract. The detection of tumors in digital images originated from mammograms can be a challenging task. In this paper we investigate a Computer Aided Diagnosis System based on a Genetic Programming classifier. The performance of the considered classifier is evaluated for five of the image descriptors used in literature and we propose a new approach, by utilizing descriptors fusion. Numerical experiments are performed on a sample of the Digital Database for Screening Mammography data set and indicate that descriptors fusion can lead to better classifying performances.

## 1. Introduction

Breast cancer is the second most frequent form of cancer in the world, and, by far, the most frequent form of cancer among women, with over 1.67 million new cases diagnosed in 2012 (25% of all cancer cases) [8]. Even though early, asymptomatic stages of breast cancer can be detected using a non-invasive technique, mammogram examination [16], this approach relies mostly on the expertise of the radiologist and can lead to a high number of erroneous diagnoses. According to [10], about one in 2000 women will have her lifespan increased by ten years due to mammogram examination, while other ten women will be administrated unnecessary treatment. Moreover, 200 women will suffer from significant psychological stress due to false positive results. Another reason for concern is the false negative diagnoses, as it is estimated that about 10% - 30% of the breast cancer cases are never detected using mammograms.

As a solution to this problem, a technique called double reading has been adopted [18]. More precisely, each case is analyzed independently by two radiologists, in an attempt to reduce the rate of wrong diagnoses. However, since this leads to additional costs and workload, Computer Aided Diagnosis

Received by the editors: October 23, 2014.

2010 *Mathematics Subject Classification.* 68T05, 91E45.

1998 *CR Categories and Descriptors.* I.2.6 [**Artificial Intelligence**]: Learning – *Induction.*

*Key words and phrases.* Genetic programming, Image descriptors, Descriptors fusion, Breast cancer.

Systems can assist a single radiologist in interpreting the mammogram, offering him support in establishing a diagnostic.

The purpose of this paper is to investigate a genetic programming (GP) classifier for the learning phase of such a system. The input data for the classifier will consist of image features extracted using several different descriptors, namely Statistical Moments, Gray Level Run Length (GLRL), Haralick features, Gabor filters and the Histogram of Oriented Gradients (HOG). Also, we will assess the performance obtained by combining (through fusion) image descriptors as, to the best of our knowledge, this approach has not been used before for breast cancer detection.

The paper is organized as follows: Section 2 details the breast cancer diagnosis problem and the main steps that must be performed for solving it. Section 3 describes the utilized image descriptors, Section 4 reviews the main components of the proposed learning algorithm, and Section 5 shows the numerical experiments and the obtained results. Conclusion and future work can be found in Section 6.

## 2. Problem of breast cancer diagnosis

The problem of breast cancer diagnosis has been intensively studied as a binary classification problem and has been solved by a supervised learning algorithm by respecting the architecture of a general classifier of two main modules: one for data pre-processing (in fact, image processing) and another one for induction of the classifier.

The input data for such a problem consists in a set of training data examples (in our case features extracted from images) each labeled correspondingly as positive or negative samples depending on which class (healthy patient or sick patient) they belong to. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. The data set is split in two parts: training data and testing data. The learning algorithm will construct the decision model by using both information (image descriptor and class) about all the images from the training set. After the classifier has been constructed, its performance will be verified by using the testing data (the classifier outputs will be compared by the real labels associated to each image from the testing set). Following this scenario, the algorithm will be able to correctly determine the class labels for unseen instances.

Different machine learning classifiers have been trained on features like GLRL [12], Statistical Moments [2], Gabor filters [19], Haralick features [21]. From the previously used classifiers, we remember Support Vector Machines, Random Forests, Logistic Model Trees, K Nearest Neighbors, and Naive Bayes.

In [15], image descriptors have been evaluated in the presence and absence of clinical data and the HOG feature descriptor has been used for the first time for describing breast lesions.

In this paper, we propose a new approach that involves the use of a GP classifier that encodes discriminant functions, and the use of the area under the Receiver Operating Characteristic (ROC) curve as a fitness measure. Furthermore, we show that for the given problem, descriptors fusion outperforms the use of individual descriptors, the system obtaining a maximum accuracy of 0.87 when using the combination of Haralick features, moments and GLRL.

## 3. Image descriptors

The first step in building automated diagnosis systems is the extraction of relevant characteristics from images. In order for a system to be able to classify images in different categories, first these need to be processed, resulting in numerical values that the system can interpret. Therefore, representations of visual characteristics like shape, color and texture, called image descriptors or visual descriptors, need to be extracted from the images.

3.1. **Statistical Moments.** Statistical Moments are statistical measures based on the intensity of the pixels of an image, and are computed from the gray level image histogram. In this paper we include a set of six features: mean value, standard deviation, skewness, kurtosis, the minimum and the maximum intensity value.

3.2. **Grey Level Run Length.** Grey Level Run Length (GLRL) [9] computes the occurrences of sets of correlated pixels, for a given length, direction, and for a particular gray level, and stores these values on a GLRL matrix. More precisely, being given a direction (e.g. the horizontal direction), for each allowed gray level it is checked how many sets of two consecutive pixels with the same value exist. Then the procedure is repeated for groups of three pixels, four pixels and so on. From the resulted matrix different characteristics can be computed, resulting in a set of 11 features.

3.3. **Haralick features.** The gray level co-occurrence matrix represents a technique of computing statistical measures of the distribution of pixel pairs within an image, by considering the relationships between two pixels, called the *reference* pixel and the *neighbor* pixel. Starting from the top left corner and continuing to the right down corner, each pixel becomes in turn a reference pixel. Given a separation distance, also called *offset*, occurrences of pairs of pixel with a certain gray level are counted and added to the matrix. Out of this matrix Robert M. Haralick has proposed the extraction of different characteristics that describe texture, called Haralick features [5], out of which we mention the energy, entropy, contrast, inverse difference moment, inertia and correlation.

3.4. **Gabor filters.** Gabor Filters are linear filters that model the functions of simple cells from the mammalian visual cortex. Therefore, image processing using Gabor filters is thought to be similar to the perceptions from the human visual system [14]. More precisely, considering the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave [4]. Gabor filters are often used for edge detection, as they detect edges with a given frequency and orientation.

3.5. **Histograms of Oriented Gradients.** The principal idea behind the Histograms of Oriented Gradients (HOG) descriptor is that object appearance and shape within an image can be described by the distribution of intensity gradients. More precisely, HOG is an image descriptor used for object detection which computes the number of gradients orientation in localized portions of an image [3]. The image is divided in small spatial regions, called cells, and for each cell a 1-D histogram of gradient orientation is computed. The first step in computing the histogram is represented by the gamma and color normalization. Then the gradient magnitude is computed for each pixel, by applying certain mask filters. After that follows the computation of the gradient orientation, which is the direction of the fastest gradient change, obtaining a matrix with $n$ x $m$ values, where $n$ and $m$ represent the dimensions of the image in pixels. From this matrix the HOG is computed by counting for each cell the number of pixels for which the gradient orientation falls in the respective cell. Finally, in order to obtain invariance to illumination and shadowing, the histogram can be normalized, common options being the L1 and L2 norms [3].

## 4. GENETIC PROGRAMMING

After extracting relevant features from the images, the next step is building a machine learning system that will use those features as input data. For this purpose, we have decided to use a Genetic Programming (GP) algorithm [13] which, after learning from annotated examples, will be able to generalize and classify newly seen mammograms. GP algorithms were chosen because they are a flexible and powerful evolutionary technique. GP are able to both represent data and to perform computations, and can be used not only for classifying data, but also for data pre-processing and post-processing. Moreover, due to the fact that the discriminant function evolved by the algorithm is similar to the mathematical operations and transformations used in image processing, GP algorithms are considered to be very suitable for image classification tasks [6].

We have chosen to use a tree representation of chromosomes, with nodes consisting of image features and constants, and the functions {+, -, *, /}, where +, - and * represent the usual mathematical operators and / is the safe division (given $a$ and $b$, it returns $a$ / $b$ if $b = 0$ and 1 otherwise). The genetic operators are the usual ones: crossover is performed by switching two randomly selected sub-trees, mutation consists of changing a randomly selected sub-tree with a newly generated tree and the selection method is tournament selection. The population is initialized with the ramped half-and-half method [13].

For GP algorithms, the classical approach in computing the fitness measure in the case of a binary classification problem is choosing one or several threshold points and evaluating, for each of them, the value corresponding to the true positive and false positive rates. In the case of multiple threshold, we will obtain a ROC curve, and the fitness will be obtained by computing the area underneath [7]. However, as this method can lead to increased execution time, we have proposed a new approach that, as far as we know, has not been tried on mammograms, by using the Wilcoxon-Mann-Whitney (WMW) statistic [20] to estimate the area under the ROC curve. WMW does not require the actual building of the ROC curve and therefore is less expensive in concerns of execution time. The basic thought is that pairwise comparisons of the negative class observations and positive class observations are performed, collecting rewards when certain constraints are satisfied. The first constraint is the fact that the observations of the positive class need to be correctly labeled (e.g. bigger than zero) and they also need to be bigger than the observations of the negative class [1]. In this way, WMW is efficient not only in evaluating the accuracy of the classifier on the training data, but also in separating the instances of the two classes.

## 5. Numerical results

5.1. **Data Set.** In order to evaluate the performance of the considered classifier, *The Digital Database for Screening Mammography* (DDSM) [11] has been chosen due to the high number of annotated mammograms with biopsy proven diagnostic. The data set was originally build from two types of film mammograms, mediolateral oblique and craniocaudal, which were then scanned to obtain digital images in LJPEG format. In order to be able to process those images, we converted them in PNG format, using the freely available program

DDSM Software [17]. The resulted images have a size ranging from about 1,600 x 3,700 pixels to about 3,800 x 6,800 pixels.

Out of the 2620 available cases, we have chosen a sample of the mammograms from the normal volumes 1-6 and all the available images from the cancer volumes (volumes 1-15), resulting in a number of 3555 images from four different scanners. About two thirds of the images were used for training the classifier, and the rest for validating the system.

5.2. **Experimental framework.** Experiments were performed for each of the considered descriptors taken individually and for seven descriptor combinations, resulting in 12 scenarios. For each scenario 200 experiments were ran and the performance of the best solution, of the worst solution and the mean performance of all of the solutions were evaluated. The classification performance was measured by computing the accuracy, the precision and the recall and, for each measure, confidence intervals of 95% were provided.

For each experiment we considered a population of 100 individuals that evolved during 500 generations or until no changes were recorded in the population for at least 100 generations. The mutation and the crossover probabilities were the ones recommended by J. R. Koza, namely 10% and, respectively, 90% [13]. The maximum depth of each chromosome was computed by the formula from Eq. 1 [13] where *terminalsNumber* represents the number of terminals.

$$\frac{\log(terminalsNumber)}{\log(2)}, \tag{1}$$

5.3. **Individual descriptors results.** Out of the five considered descriptors, the best classification performance was recorded by statistical moments, which obtained the highest scores for maximum accuracy, maximum precision and for all of the performance measures for the mean and for the worst solution. The maximum accuracy was equal to 0.83 and the maximum recall was 0.86, both registered by the moments descriptor. The maximum precision, of 0.83, was scored by Haralick features. Figure 1 shows complete results, with confidence intervals, for accuracy, precision and recall.

5.4. **Descriptors fusion results.** Out of the five considered descriptors, we chose for the fusion the four descriptors that scored the best performances when evaluated individually: statistical Moments, Haralick features, Grey Level Run Length and Histogram of Oriented Gradients. In general, the maximum classification performances were better when using descriptors fusion: out of the seven tried fusions, five scored a maximum accuracy that was better than the maximum accuracy obtained by individual descriptors. Furthermore, one other fusion (HOG + GLRL) had a slightly better accuracy than the ones

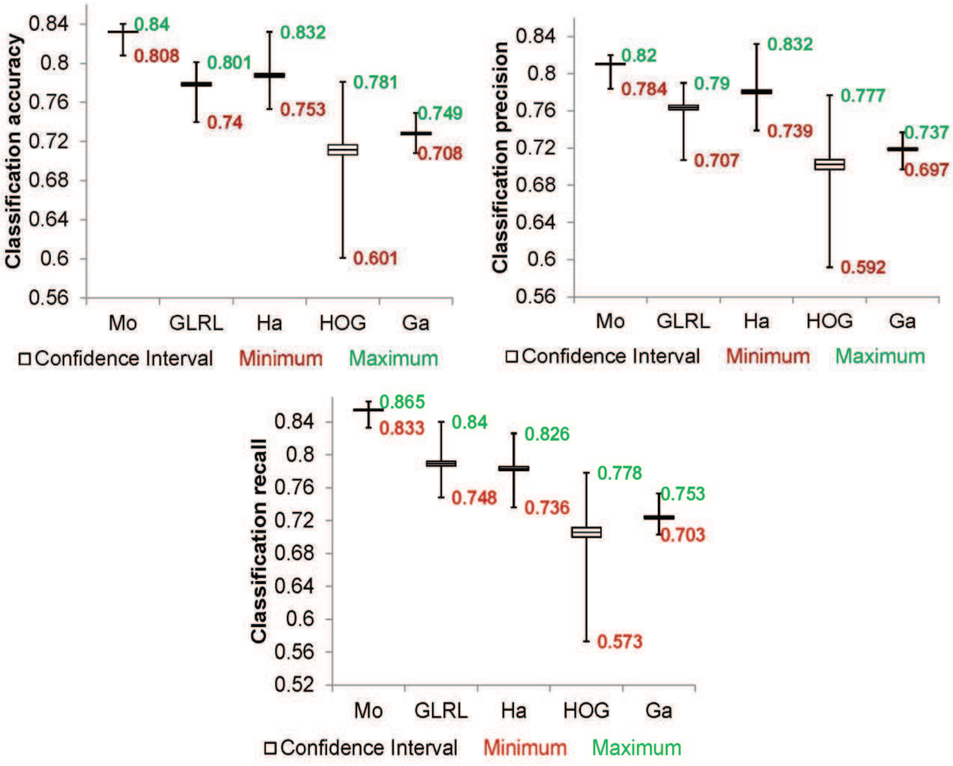

FIGURE 1. Classification results of individual descriptors. Mo: statistical moments; Ha: Haralick features; Ga: Gabor filters.

obtained by using the two descriptors individually (0.8042 for the fusion, compared with 0.8008 for GLRL and 0.7814 for HOG). The best performance of the system resulted from using the combination of Moments, GLRL and Haralick features. For this fusion, the maximum accuracy was of 0.87, the maximum recall was of 0.85 and the maximum precision was equal to 0.88. Figures 2 shows the results obtained by using descriptors fusion.

5.5. **Comparison to related work.** In order to assess the performance of the GP classifier, we analyzed the results of individual descriptors with the ones reported in the study [15], where Support Vector Machines, Random Forests, Logistic Model Trees, K Nearest Neighbors and Naive Bayes classifiers were used. However, only one set of results was presented, and the authors do not mention to which learning algorithm it belongs.

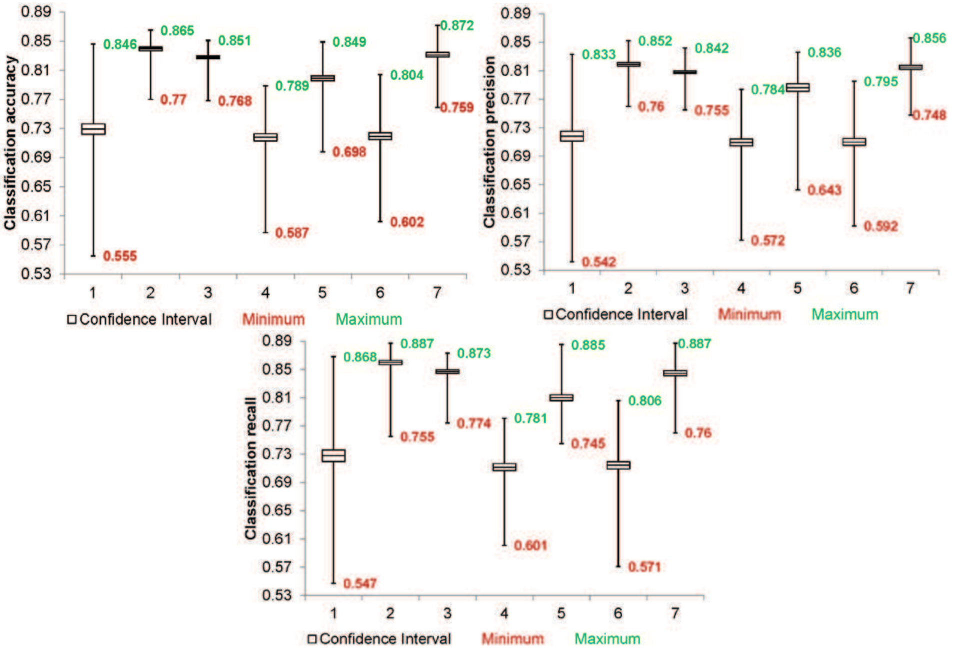

Figure 2. Classification results of descriptors fusion. 1: Mo + HOG; 2: Mo + Ha; 3: Mo + GLRL; 4: Ha + HOG; 5: Ha + GLRL; 6: GLRL + HOG; 7: Mo + Ha + GLRL.

For a more accurate comparison, we only considered the results obtained using an experimental framework similar to ours, where images were sampled from the DDSM data set and no clinical data was used. Table 1 compares the mean accuracies obtained by the GP classifier with the accuracies presented in [15], showing that GP performed better for all of the considered image descriptors. However, since these results were obtained using different images, in different numbers, the comparison does not have a statistical characteristic and can only serve as a guideline.

Table 1. Mean accuracies of the GP compared with the ones reported in [15].

| | GP | [15] |
|---|---|---|
| Moments | 0.831 | 0.707 |
| GLRL | 0.778 | 0.733 |
| Haralick | 0.787 | 0.718 |
| HOG | 0.711 | 0.707 |
| Gabor | 0.723 | 0.711 |

## 6. Conclusions

Breast cancer diagnosis can be a difficult process that involves high workload and relies mostly on the expertise of the radiologist. In order to reduce the chance of human error while decreasing the costs implied by double-reading, automated diagnosis systems could aid the radiologist in taking a decision and determining a diagnostic. This paper investigates a new approach to such a system, by using a Genetic Programming algorithm. Therefore we show that GP is suitable for the task of mammogram classification, the system obtaining a maximum accuracy of 0.87. Furthermore, we propose the use of descriptors fusion, and we show that in most of the cases this approach outperforms results obtained by individual descriptors.

Future work includes testing all other possible fusions of the considered descriptors in order to verify whether a less efficient descriptor could lead to an increase in the overall performance of the combination. Another possible direction is evaluating different machine learning classifiers (e.g. Support Vector Machines) on the given data sets, in order to be able to perform a statistical comparison with the performance of the GP classifier. Also, we intend to validate the obtained results on other data sets such as BCDR and MIAS.

## References

[1] U. Bhowan, M. Zhang, and M. Johnston. Improving gp with new fitness functions. In *Genetic Programming 13th European Conference Proceedings*, pages 3–6, 2010.

[2] I. Christoyianni, E. Dermatas, and G. Kokkinakis. Fast detection of masses in computer-aided mammography. *IEEE Signal Proc Mag*, 17(1):54–64, 2000.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages I: 886–893, 2005.

[4] J. G. Daugman. Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of Optical Society of America*, 2(7):1160–1169, 1985.

[5] I. Dinstein, K. Shanmugam, and R. M. Haralick. Textural features for image classification. In *CMetImAly77*, pages 141–152, 1977.

[6] Pedro G. Espejo, Sebastian Ventura, and Francisco Herrera. A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(2):121–144, March 2010.

[7] Tom Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, Hewlett Packard Laboratories, June 2 2003.

[8] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F.Bray. Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase. Technical Report 11, International Agency for Research on Cancer, Lyon, France, 2013.

[9] M. M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics Image Processing*, 4(2):172–179, June 1975.

[10] P. Gotzsche and M. Nielsen. Screening for breast cancer with mammography. *The Cochrane Library*, 2011.

[11] M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. In *Proceedings of the Fifth International Workshop on Digital Mammography*, pages 212–218, 2001.

[12] J. K. Kim and H. W. Park. Statistical textural features for detection of microcalcifications in digitized mammograms. *IEEE Trans. Medical Imaging*, 18(3):231–238, March 1999.

[13] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection.* MIT Press, 1992.

[14] S. Marcelja. Mathematical description of the responses of simple cortical cells. *Journal of Optical Society of America*, 70:1297–1300, 1980.

[15] Daniel C. Moura and Miguel Ángel Guevara-López. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. *Int. J. Computer Assisted Radiology and Surgery*, 8(4):561–574, 2013.

[16] H. D. Nelson, K. Tyne, A. Naik, C. Bougatsos, B. K. Chan, and L. Humphrey. Screening for breast cancer: systematic evidence review update for the us preventive services task force. *Ann Intern Med*, 151(10):727, 2009.

[17] D. C. Rose. Ddsm software. ttp://microserf.org.uk/academic/Software.html. Accessed: 2014-06-03.

[18] L. Tabar, B. Vitak, T. Chen, A. Yen, A. Cohen, T. Tot, S. Chiu, S. Chen, J. Fann, J. Rosell, H. Fohlin, R. Smith, and S. Duffy. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*, 260(3):658–663, 2011.

[19] Defeng Wang, Lin Shi, and Pheng-Ann Heng. Automatic detection of breast cancers in mammograms using structured support vector machines. *Neurocomputing*, 72(13-15):3296–3302, 2009.

[20] Lian Yan, Robert H. Dodier, M. Mozer, and Richard H. Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. 2003.

[21] S. Yu and L. Guan. A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. *IEEE Trans. Medical Imaging*, 19(2):115–126, February 2000.

Computer Science Department, Babeş Bolyai University, Cluj Napoca, Romania

*E-mail address*: fratean.stefana@gmail.com
*E-mail address*: lauras@cs.ubbcluj.ro