Combining Texture and Shape Descriptors for Bioimages Classification: A Case of Study in ImageCLEF Dataset*

Anderson Brilhador, Thiago P. Colonhezi,
Pedro H. Bugatti, and Fabrício M. Lopes

Federal University of Technology - Paraná, Brazil
{pbugatti,fabricio}@utfpr.edu.br

Abstract. Nowadays a huge volume of data (e.g. images and videos) are daily generated in several areas. The importance of this subject has led to a new paradigm known as eScience. In this scenario, the biological image domain emerges as an important research area given the great impact that it can leads in real solutions and people's lives. On the other hand, to cope with this massive data it is necessary to integrate into the same environment not only several techniques involving image processing, description and classification, but also feature selection methods. Hence, in the present paper we propose a new framework capable to join these techniques in a single and efficient pipeline, in order to characterize biological images. Experiments, performed with the ImageCLEF dataset, have shown that the proposed framework presented notable results, reaching up to 87.5% of accuracy regarding the plant species classification, which is highly relevant and a non-trivial task.

Keywords: image descriptors, feature selection, classification, pattern recognition.

1 Introduction

Currently the digitization of information is becoming more common, generating a massive volume of data, leading to a new paradigm of data analysis known as eScience [1]. In 2011 it was estimated that the amount of information in the digital universe exceeded 1.8 zettabytes [2]. Among these data are largely digital content such as images and videos. However, these contents can only become useful when accessed efficiently, meaning not only fast, but also accurate. Therefore, it is needed new computational tools to retrieve and index these great volume of generated data.

The recovery of data can occur in textual form, through the inclusion of identifiers known as *tags*. Although, the *tags* inclusion have the advantage of indexing and retrieving multimedia content quickly, it is required to be given the textual information for each content. This task may lead to inconsistencies because it depends on the human perception. If the *tag* inclusion is incorrect, incomplete or not done, the content is not indexed properly, and consequently it

* This work has been supported by Fundação Araucária and CNPq.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 431–438, 2013.

will not be recovered correctly. On the other hand, the retrieval of images based on its content eliminates the human interaction and the allocation of *tags* is done automatically. This process is based on extracting measures or features of the content, which is used to perform the indexing and retrieval of a particular object. A major challenge in this process is to generate features that really represent the data and consequently establish a classifier in order to correctly identify the data under analyses.

In this scenario, emerges an important research area known as Bioimage informatics [3], which focuses attention on developing new techniques for image processing, data mining, database and visualization in order to extract, compare, search and manage the biological knowledge in data-intensive problems. Regarding bioimages, there is an important challenge called the CLEF Cross Language Image Retrieval Track (ImageCLEF) [4]. This challenge illustrates the importance of the image retrieval in the actual data-intensive scenario. In particular, the biological diversity is very significant both in relation to the genetic potential as compared to the number of species and ecosystems. Considering plant biodiversity, the Amazon Rainforest holds the largest reserves of medicinal plants in the world. Then, there is a great necessity to recognize the flora through fast and efficient computational methods in order to deal with big data scenarios.

This paper presents an efficient framework for bioimage processing, feature extraction and classification, based on its texture and shape descriptors, which are combined in order to classify the input images. In this way, the performance of the proposed methodology was evaluated based on ImageCLEF [4] database by using several image features and classification techniques for this task, which are presented in the following sections.

2 Background

2.1 Feature Extraction

The feature extraction is defined as the entire set of operations for image processing and analysis performed in order to obtain numerical values that characterize the images or parts of them. It can also be defined as the capture of the most relevant information from a data given as input. The features extracted from the images can be based on three main classes: color, texture and shape.

The shape descriptors are measures of the boundaries, such as chain code, circularity, width, perimeter and area. The Fourier descriptor is widely used as a shape descriptor through its coefficients. The Fourier coefficients represents a global information of the curvature extracted from the image, which can be used to compare objects, because these coefficients are invariant to rotation, translation and scale. This invariance is achieved by applying simple transformations to the Fourier coefficients [5]. In this way, the Fourier coefficients from object boundary [6] was adopted in this work.

The color descriptors are based on the spectral radiation emitted or reflected by the objects, quantified by the intensity of the pixels in different spectral bands.

In this work, it was not adopted color descriptors by the nature of application in images of leaves, which have little variation in this feature.

The texture is an important descriptor used to identify objects in a digital image. The Haralick descriptors [7] use the distribution of gray levels and co-occurrence matrices to evaluate the different textures, which can be defined as: thin, thick, smooth, wavy, irregular or linear. Another method for texture analysis was proposed by Chao-Bing Lin and Quan [8], called Quantized Compound Change Histogram (QCCH). In this method, given a particular pixel the main idea is to check all gray level variations from its neighbors in the four directions. The differences of intensities in each direction are used for the construction of a histogram. By considering the variation of the intensities, this approach is free from variation between rotation and translation of the image. The Haralick and QCCH descriptors were adopted in the present work.

2.2 Image Classification

The classification is a way to analyse the data set and extracting models that lead to a category (class). The classification process can be defined in two paradigms: supervised learning, in which is known the classes for each available sample and unsupervised learning, in which the samples has no indication about its class.

The supervised learning is commonly divided into two tasks. The first task is called training, in which the classifier is constructed to determine the classes of the input objects from their attributes [9]. The second task is the classification, in which the model created in the first task is applied in order to define the classes for the input samples. There are a wide variety of supervised learning methods. In order to explore some important methods available in the literature the following classifiers: K-NN (K-nearest neighbor) [10], NB (Naive Bayes) [11], MLP (Multilayer Perceptron) [12], RF (Random Forest) [13], J4.8 [14] and SVM (Support Vector Machines) [15] were considered in this work.

In addition, the classifiers can be combined with the adaptive boosting strategy, which can be defined as a machine learning algorithm used to improve the performance of other learning algorithms [16]. In order to evaluate the performance of such technique, it was also considered in this work.

2.3 Feature Selection

The feature selection approach has been investigated, mainly in pattern recognition area, since the 70s [17]. By considering the big data scenario, the feature selection techniques has become essential in many knowledge areas [18–21].

Regarding pattern recognition, the feature selection aims to reduce the volume of features, i.e. the feature space, keeping the maximum of the source information as possible, in order to reduce the computational cost and to increase the accuracy of the classifier. Other aspects may be useful such as to increase the comprehensibility of the classification model and to increase the robustness of learning.

An important consideration in feature selection methods [22] is that much of the search assumes the monotonicity principle, i.e. increasing the number of attributes improves the performance of the classifier. However, adding more features the estimation error also increases, because the number of samples needed for constructing a suitable model.

A well known feature selection technique is the so-called correlation feature selection (CFS) [23]. This technique evaluates the subset of features by considering the consistency measure, seeking for combinations of attributes whose values split the data into subsets associated with a majority class. For a feature fit in this condition, the technique seeks for features that have a high correlation with the observed class and features not correlated among themselves, considering not only the feature individually, but also the relation among them.

3 Proposed Approach

After the image dataset definition it was performed the segmentation of the desired object from the image. Once the segmentation process is not the main focus of this paper, it was applied the baseline thresholding method proposed by Otsu [24]. Its basic principle is to select a threshold that maximizes the variance between classes (foreground and background).

The next step was the extraction of the image features as related in Sec. 2.1. It was developed in Java technology a framework to perform the extraction of the adopted image features from an input image, which is freely available[1].

The image features are extracted from each sample and was built a feature vector with 218 positions, as follows: [1 - 52] Haralick, [53 - 92] QCCH and [93 - 218] Fourier coefficients. Each image has an unique feature vector, which will be considered for its classification (see Sec. 5).

The final step is the classification of the image data set by applying the methods described in Sec. 2.2. In face of the number of the extracted image features, a feature selection was adopted as a filter step before the classification (Sec. 2.3). Besides, the adaptive boosting strategy was also performed with the classifiers.

4 Measuring Effectiveness

In order to evaluate the performance, it is necessary to perform the classification and to compare the results with the correct class for each sample. An approach commonly used in this task is the cross-validation or k-fold cross validation, in which the image data set is splitted in k *folds*, $D_1, D_1, \ldots, D_k$ of equal size. Then, the training and test set is performed k times in order to evaluate the performance of the classifier. More specifically, one fold D_i is used for test and the remaining folds are used for training. The overall accuracy is calculated by averaging the results obtained at each step, thus achieving an estimation of the quality of knowledge generated by the classification model and allowing statistical analyses.

After performing the test, it is also possible to obtain statistical values for measure the performance of the classifier such as *Precision*, *recall* and receiver operating characteristic (ROC) curve [25].

[1] http://code.google.com/p/jimagefeature/

4.1 Image Dataset Description

The ImageCLEF 2012 [26] image dataset was adopted in this work in order to evaluate the proposed methodology by considering the plant identification species from its leaves. The image dataset includes $n = 126$ different species of trees located in the French Mediterranean area, the total number of samples available is 11,527, which are subdivided into three categories of images: scan (57%), scan-like (24%) and free natural photos(19%).

The scan category was adopted in this work, which contains 4870 images divided unevenly among the 126 species (classes). In order to normalize the distribution among classes, it was applied the following procedure: (1) it was calculated the average x of images for each species e_i, $i = 1, \ldots, n$; (2) by considering this average value, it was observed the quantity of images in each specie (q_i) was greater than x. The species with lower quantity of samples, i.e. $q_i < x$, were excluded for not having enough samples. This pre-processing led to a balanced number of samples for each species. As a result, the pre-processed image dataset was reduced to 3,582 samples distributed in 54 different species.

5 Results

The first round of experiments was performed in order to evaluate the contribution of each class of the adopted descriptors (shape and texture). The correlation-based feature approach was applied to the complete feature vector (see Sec. 3) before the classification methods. As a result, the features were selected in order to build a new feature vector as shown in Table 1.

Table 1. Feature vector composition by considering all features and the feature selection results

Descriptors	All Features	Selected Features
Haralick	52 (24%)	11 (20%)
QCCH	40 (18%)	20 (37%)
Fourier	126 (58%)	23 (43%)
Total	**218**	**54**

These results points out that the texture features are slightly more relevant than shape descriptors with respect to the number of selected features. However, the feature selection indicates that both were important. The texture features have 58% and the shape descriptors have the 42% of the selected features.

The second round of experiments was performed in order to investigate the behavior of the adopted classifiers by considering all features, the filtered feature vector and the the Adaptive Boosting technique. Table 2 presents the average results by adopting the 10-fold cross validation approach for each configuration. It is important to notice that the SVM classifier presented the best performance over all configuration, achieving 87,5% of precision when all features were applied. The MLP and Random Forest classifiers showed slightly lower results. However, these results were achieved only after performing the feature selection and adaBoost approaches, respectively, indicating the robustness of the SVM classifier. The J4.8, Random Forest methods had a significant improvement when using the feature selection and adaBoost approaches. Surprisingly, the MLP classifier showed no improvement when combined with the AdaBoost approach. The K-NN and Naive Bayes showed similar results for all adopted variations.

Table 2. Performance comparison among the adopted classifiers by considering all features, the selected features and the selected features with adaBoost technique in terms of the precision measure

Classifier	all features	selected features	adaBoost
K-NN	78.9%	82.6%	82.6%
Naive Bayes	76.8%	78.9%	78.9%
J4.8	67.1%	68.3%	84.1%
Random Forest	77.0%	81.1%	**87.1%**
MLP	50.9%	**83.9%**	83.3%
SVM	**87.5%**	83.2%	86.5%

In order to evaluate the performance of the classifiers regarding the better and the worst classified species, the five species with better and worst results were selected in Tables 3 and 4 respectively. Figure 1 shows an example for each of the species listed in Tables 3 and 4.

Table 3. The five better classified species

Id	**Class**	**K-NN**		**NB**		**J48**		**RF**		**MLP**		**SVM**	
		Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC
(a)	Daphne-cneorum	1	0,992	1	0,985	0,988	0,999	0,988	1	0,975	1	1	1
(b)	Buxus-sempervirens	0,978	0,999	0,983	0,998	0,972	1	0,978	1	0,962	1	0,972	1
(c)	Juniperus-oxycedrus	0,988	1	0,987	0,997	0,963	1	0,988	1	0,898	1	0,975	1
(d)	Albizia-julibrissin	1	1	0,886	0,999	1	1	0,951	1	0,95	0,999	1	0,993
(e)	Nerium-oleander	0,978	0,993	0,965	0,988	0,966	0,996	0,966	0,998	0,925	0,999	0,977	0,997

Table 4. The five worst classified species

Id	**Class**	**K-NN**		**NB**		**J48**		**RF**		**MLP**		**SVM**	
		Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC	Precision	ROC
(f)	Ginkgo-biloba	0,784	0,788	0,515	0,911	0,7	0,985	0,871	0,983	0,583	0,959	0,795	0,974
(g)	Acer-campestre	0,795	0,857	0,515	0,906	0,789	0,956	0,696	0,986	0,707	0,954	0,688	0,979
(h)	Arbutus-unedo	0,562	0,866	0,791	0,782	0,662	0,963	0,593	0,975	0,667	0,953	0,788	0,969
(i)	Laurus-nobilis	0,661	0,926	0,508	0,882	0,681	0,993	0,765	0,99	0,694	0,961	0,706	0,988
(j)	Fraxinus-angustifolia	0,464	0,704	0,712	0,795	0,676	0,99	0,803	0,988	0,615	0,982	0,662	0,932

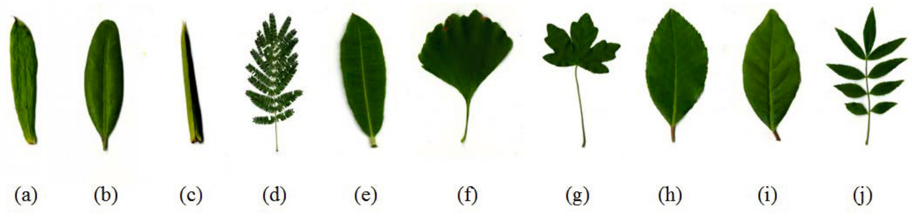

Fig. 1. Samples of the better and worst classified species accordingly Tables 3 and 4

By considering the five better classified species (a–e), it was observed that they have similar features of texture and shape. For instance, the *Daphne cneorum* (a) and *Juniperus oxycedrus* (c) species present similar shape features, but have different texture features being the first roughened and the second smooth. Only the *Albizia-julibrissin* has different texture and shape from other species.

The five worst classified species (f–j) show similar texture and shape features as in the case of the species *Arbutus unedo* (h) and *Laurus nobilis* (i) increasing the difficulty in classification between them. Another case is the *Fraxinus angustifolia* (j), because it varies greatly in shape making very difficult to establish a pattern for this class, the same variation occurs with the species *Ginkgo biloba* (f) and *Acer campestre* (g).

6 Conclusion

This paper presents a novel and flexible framework for Bioimage processing, feature extraction and classification. The proposed framework combines texture-based and shape-based features improving in a great extent the classification accuracy of biological images. Furthermore, it not only allows an easy addition of new methods for processing, description and classification of images, but also provides the evaluation of such methods under the same conditions. It is important to highlight that the great majority of works in the literature neglects this issue. Another point addressed by the proposed approach is related to the high dimensionality of the feature vectors. In order to mitigate this problem we embedded a feature selection method into it.

As shown in the experiments section, the proposed approach presented notable results by considering the plant species classification, reaching up to 87.5% of accuracy in the overall case. Considering each one the species it reached, in many cases, up to 100% of accuracy. Moreover, the dimensionality of the feature vectors was reduced about 4 times less dimensions, diminishing the classification computational cost. Hence, this testifies the usefulness of the proposed approach in real biological applications.

Future work includes to apply the proposed framework to other biological image datasets, to include color-based features and join new steps in the proposed framework, such as unsupervised classification. It is also planned as a future work to apply non-parametric tests for statistical comparisons of classifiers as described in [27].

References

1. Gray, J.: Jim gray on escience: a transformed scientific method. The Fourth Paradigm: Data-intensive Scientific Discovery (2009)
2. Gantz, J., Reinsel, D.: Extracting value from chaos. IDC iView, 1–12 (2011)
3. Peng, H.: Bioimage informatics: a new area of engineering biology. Bioinformatics 24(17), 1827–1836 (2008)
4. Müller, H., Clough, P., Deselaers, T., Caputo, B.: ImageCLEF: Experimental Evaluation in Visual Information Retrieval, vol. 32. Springer (2010)

5. Bartolini, I., Ciaccia, P., Patella, M.: Warp: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(1), 142–147 (2005)
6. da Fontoura Costa, L., Cesar Jr., R.M.: Shape analysis and classification: theory and practice, 2nd edn. CRC Press (2010)
7. Attig, A., Perner, P.: A comparison between haralick's texture descriptor and the texture descriptor based on random sets for biological images. In: Perner, P. (ed.) MLDM 2011. LNCS, vol. 6871, pp. 524–538. Springer, Heidelberg (2011)
8. Huang, C.B., Liu, Q.: An orientation independent texture descriptor for image retrieval. In: Int. Conf. on Communic., Circ. and Systems, pp. 772–776. IEEE (2007)
9. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann (2006)
10. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning 6(1), 37–66 (1991)
11. Lewis, D.D.: Naive (bayes) at forty: The independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 4–15. Springer, Heidelberg (1998)
12. Gardner, M., Dorling, S.: Artificial neural networks–a review of applications in the atmospheric sciences. Atmospheric Environment 32(14-15), 2627–2636 (1998)
13. Statistics, L.B., Breiman, L.: Random forests. Machine Learning, 5–32 (2001)
14. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
15. Abe, S.: Support vector machines for pattern classification. Springer (2010)
16. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P.M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
17. Mucciardi, A.N., Gose, E.E.: A comparison of seven techniques for choosing subsets of pattern recognition properties. IEEE Trans. on Comp. 100(9), 1023–1031 (1971)
18. Lopes, F.M., Martins Jr., D.C., Cesar Jr., R.M.: Feature selection environment for genomic applications. BMC Bioinformatics 9(1), 451 (2008)
19. Lopes, F.M., de Oliveira, E.A., Cesar Jr., R.M.: Analysis of the GRNs inference by using Tsallis entropy and a feature selection approach. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 473–480. Springer, Heidelberg (2009)
20. Lopes, F.M., Martins Jr., D.C., Barrera, J., Cesar Jr., R.M.: SFFS-MR: A floating search strategy for GRNs inference. In: Dijkstra, T.M.H., Tsivtsivadze, E., Marchiori, E., Heskes, T. (eds.) PRIB 2010. LNCS, vol. 6282, pp. 407–418. Springer, Heidelberg (2010)
21. Pinto, S.C.D., Mena-Chalco, J.P., Lopes, F.M., Velho, L., Cesar Jr., R.M.: 3D facial expression analysis by using 2D and 3D wavelet transforms. In: ICIP, pp. 1281–1284 (2011)
22. John, G.H., Kohavi, R., Pfleger, K., et al.: Irrelevant features and the subset selection problem. In: 11th Int. Conf. on Machine Learning, pp. 121–129 (1994)
23. Hall, M.A.: Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato (1999)
24. Sahoo, P.K., Soltani, S., Wong, A.: A survey of thresholding techniques. Computer Vision, Graphics, and Image Processing 41(2), 233–260 (1988)
25. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)
26. Goëau, H., Bonnet, P., Joly, A., Yahiaoui, I., Barthélémy, D., Boujemaa, N., Molino, J.: The ImageCLEF 2012 Plant Identification Task (2012)
27. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)