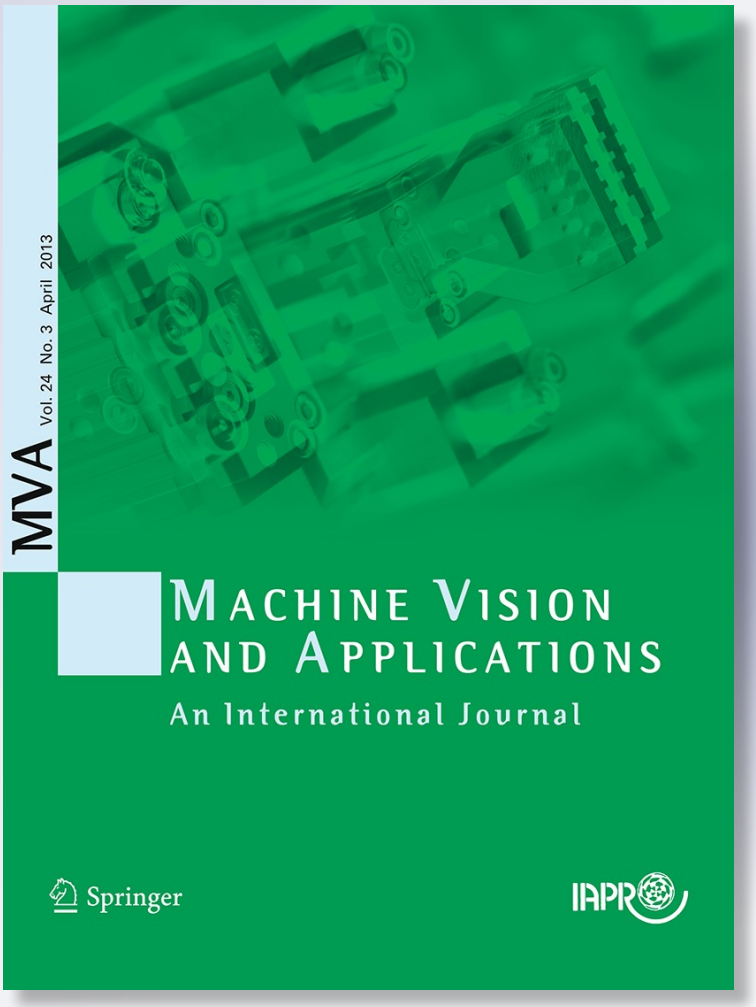
A database for automatic classification of forest species

J. Martins, L. S. Oliveira, S. Nisgoski & R. Sabourin

Machine Vision and Applications

ISSN 0932-8092
Volume 24
Number 3

Machine Vision and Applications (2013)
24:567-578
DOI 10.1007/s00138-012-0417-5



Springer

Machine Vision and Applications (2013) 24:567–578
DOI 10.1007/s00138-012-0417-5

ORIGINAL PAPER

A database for automatic classification of forest species

J. Martins · L. S. Oliveira · S. Nisgoski · R. Sabourin

Received: 7 June 2011 / Revised: 14 November 2011 / Accepted: 13 February 2012 / Published online: 20 March 2012

Abstract Forest species can be taxonomically divided into groups, genera, and families. This is very important for an automatic forest species classification system, in order to avoid possible confusion between species belonging to two different groups, genera, or families. A common problem that researchers in this field very often face is the lack of a representative database to perform their experiments. To the best of our knowledge, the experiments reported in the literature consider only small datasets containing few species. To overcome this difficulty, we introduce a new database of forest species in this work, which is composed of 2,240 microscopic images from 112 different species belonging to 2 groups (Hardwoods and Softwoods), 85 genera, and 30 families. To gain better insight into this dataset, we test three different feature sets, along with three different classifiers. Two experiments were performed. In the first, the classifiers were trained to discriminate between Hardwoods and Softwoods, and in the second, they were trained to discriminate among the 112 species. A comprehensive set of experiments shows that the tuple Support Vector Machine (SVM) and Local Binary Pattern (LBP) achieved the best performance in both cases, with a recognition rate of 98.6 and 86.0% for the first and second experiments, respectively. We believe that researchers will find this database a useful tool in their work on forest species recognition. It will also make future benchmarking and evaluation possible. This database will be available for research purposes upon request to the VRI-UFPR.

Keywords Pattern recognition · Local binary patterns · Texture · Forest species

J. Martins · L. S. Oliveira (✉) · S. Nisgoski
Federal University of Parana (UFPR), R. Rua Cel. Francisco H. dos Santos, 100, Curitiba 81531-990, PR, Brazil
e-mail: lesoliveira@inf.ufpr.br

R. Sabourin
Ecole de Technologie Superieure,
1100 rue Notre Dame Ouest,
Montreal, QC, Canada

1 Introduction

In recent years, with the advent of globalization, the safe trading of logs and timber has become an important issue. Buyers must certify that they are buying the correct material, and supervisory agencies have to certify that the wood has been not extracted illegally from forests. Millions of dollars are spent with the aim of preventing fraud, on the part of wood traders who might mix a noble species with cheaper ones, for example, or even try to export the wood of an endangered species.

Computer vision systems could be a very useful tool in this effort. However, in the past decade, most of the applications of computer vision in the wood industry have been related to quality control, grading, and defect detection [1–3]. Only recently have some authors begun to use computer vision to classify forest species. Tou et al. [4–6] have reported two forest species classification experiments in which texture features are used to train a neural network classifier. They report recognition rates ranging from 60 to 72% for five different forest species.

Khalid et al. [7] have proposed a system to recognize 20 different Malaysian forest species. Image acquisition is performed with a high performance industrial camera and LED array lighting. Like Tou et al., the recognition process is based on a neural network trained with textural features. The database used in their experiments contains 1,753 images for training, and only 196 for testing. They report a recognition rate of 95%.

Paula et al. [8] have investigated the use of GLCM and color-based features to recognize 22 different species of Brazilian flora. They propose a segmentation strategy to deal with large intra-class variability. Experimental results show that when color and textural features are used together, the results can be improved considerably.

Identifying a log or a piece of timber outside its natural environment (the forest) is not an easy task, since there are no flowers, fruits, or leaves to provide clues. Therefore, this task must be performed by well-trained specialists. However, good classification accuracy is difficult to achieve because there are not enough of these specialists to meet industry demands, in part because it takes so long to train them. Another factor to be taken into account is that the process of manual identification is a time consuming, repetitive process, which might result in specialists losing their focus and becoming prone to error. This can make the task impractical when cargo is being checked for export.

One way to simplify the task of forest species classification is to identify its taxonomy, i.e. to determine the group, genus, and family to which a given forest species belongs. Of the two groups, Hardwood species (Angiosperms), which include flowering ornamentals and all vegetables and edible fruits in addition to Hardwood trees, are the most sophisticated, and have adapted to survive in a wide range of climates and environments. Hardwood tree species are a valuable source of lumber for furniture and construction. Softwood species (Gymnosperms) consist of all the conifers: cedar, redwood, juniper, cypress, fir, and pine, including the giant sequoias. Pine and fir are used for lumber, and to make paper and plywood. They also constitute the raw materials used to make substances such turpentine, rosin, and pitch.

A major challenge to pursuing research involving forest species classification is the lack of a consistent and reliable database. To the best of our knowledge, the databases reported in the literature contain few classes, and information about their taxonomy is not readily available. To overcome this difficulty, we introduce a database in this work composed of 112 species from all over the world. As well as labeling the species, we also present their taxonomy in terms of groups, genera, and families. This database has been built in collaboration with the Laboratory of Wood Anatomy at the Federal University of Parana (UFPR) in Curitiba, Brazil, and it is available upon request for research purposes.[1]

In order to make it easier to understand the structure of our database, we have assessed various feature sets and classifiers in two different contexts. In the first experiment, the classifiers were trained to discriminate between Hardwoods and Softwoods, and in the second, they were trained to discriminate among the 112 different species. A comprehensive set of experiments shows that the tuple SVM (Support Vector Machine) and LBP (Linear Binary Pattern) achieved the best performance in both cases, with a recognition rate of 98.6 and 86.0% for the first and second experiments, respectively. The database introduced in this work makes future benchmark and evaluation possible.

Fig. 1 Microscope used to acquire the images

This paper is structured as follows: Sect. 2 introduces the proposed database. Section 3 describes the feature sets we have used to train the classifiers. Section 4 reports our experiments and discusses our results. Finally, Sect. 5 concludes the work.

2 Database

The database introduced in this work contains 112 different forest species which were catalogued by the Laboratory of Wood Anatomy at the Federal University of Parana in Curitiba, Brazil. The protocol adopted to acquire the images comprises five steps. In the first step, the wood is boiled to make it softer. Then, the wood sample is cut with a sliding microtome to a thickness of about 25 μ ($1\mu = 1 \times 10^{-6}$ m). In the third step, the veneer is colored using the triple staining technique, which uses acridine red, chrysoidine, and astra blue. In the fourth step, the sample is dehydrated in an ascending alcohol series. Finally, the images are acquired from the sheets of wood using an Olympus Cx40 microscope with a 100× zoom (Fig. 1). The resulting images are then saved in

[1] http://web.inf.ufpr.br/vri/forest-species-database.

Table 1 Softwood species (Gymnosperms)

ID	Family	Gender	Species
1	Ginkgoaceae	Ginkgo	biloba
2	Araucariaceae	Agathis	becarii
3	Araucariaceae	Araucaria	angustifolia
4	Cephalotaxaceae	Cephalotaxus	drupacea
5	Cephalotaxaceae	Cephalotaxus	harringtonia
6	Cephalotaxaceae	Torreya	nucifera
7	Cupressaceae	Calocedrus	decurrens
8	Cupressaceae	Chamaecyparis	formosensis
9	Cupressaceae	Chamaecyparis	pisifera
10	Cupressaceae	Cupressus	arizonica
11	Cupressaceae	Cupressus	lindleyi
12	Cupressaceae	Fitzroya	cupressoides
13	Cupressaceae	Larix	lariciana
14	Cupressaceae	Larix	leptolepis
15	Cupressaceae	Larix	sp
16	Cupressaceae	Tetraclinis	articulata
17	Cupressaceae	Widdringtonia	cupressoides
18	Pinaceae	Abies	religiosa
19	Pinaceae	Abies	vejari
20	Pinaceae	Cedrus	atlantica
21	Pinaceae	Cedrus	libani
22	Pinaceae	Cedrus	sp
23	Pinaceae	Keteleeria	fortunei
24	Pinaceae	Picea	abies
25	Pinaceae	Pinus	arizonica
26	Pinaceae	Pinus	caribaea
27	Pinaceae	Pinus	elliottii
28	Pinaceae	Pinus	gregii
29	Pinaceae	Pinus	maximinoi
30	Pinaceae	Pinus	taeda
31	Pinaceae	Pseudotsuga	macrolepsis
32	Pinaceae	Tsuga	canadensis
33	Pinaceae	Tsuga	sp
34	Podocarpaceae	Podocarpus	lambertii
35	Taxaceae	Taxus	baccata
36	Taxodiaceae	Sequoia	sempervirens
37	Taxodiaceae	Taxodium	distichum

PNG (Portable Network Graphics) format with no compression and a resolution of 1024 × 768 pixels.

To date, 2,240 microscopic images (20 images per species) have been acquired and carefully labeled by experts in wood anatomy. Of the 112 available species, 37 are Softwoods and 75 are Hardwoods. Table 1 describes the 37 species of Softwood species in the database, which can be divided into 23 genera and 8 families. The 75 species of Hardwood species are reported in Table 2, and can be classified into 62 genera and 22 families. The two groups of species, Hardwoods and

Table 2 Hardwood species (Angiosperms)

ID	Family	Gender	Species
38	Ephedraceae	Ephedra	californica
39	Lecythidaceae	Cariniana	estrellensis
40	Lecythidaceae	Couratari	sp
41	Lecythidaceae	Eschweilera	matamata
42	Lecythidaceae	Eschweleira	chartaceae
43	Sapotaceae	Chrysophyllum	sp
44	Sapotaceae	Micropholis	guianensis
45	Sapotaceae	Pouteria	pachycarpa
46	Fabaceae-Cae.	Copaifera	trapezifolia
47	Fabaceae-Cae.	Eperua	falcata
48	Fabaceae-Cae.	Hymenaea	courbaril
49	Fabaceae-Cae.	Hymenaea	sp
50	Fabaceae-Cae.	Schizolobium	parahyba
51	Fabaceae-Fab.	Pterocarpus	violaceus
52	Fabaceae-Mim.	Acacia	tucunamensis
53	Fabaceae-Mim.	Anadenanthera	colubrina
54	Fabaceae-Mim.	Anadenanthera	peregrina
55	Fabaceae-Fab.	Dalbergia	jacaranda
56	Fabaceae-Fab.	Dalbergia	spruceana
57	Fabaceae-Fab.	Dalbergia	variabilis
58	Fabaceae-Mim.	Dinizia	excelsa
59	Fabaceae-Mim.	Enterolobium	schomburgkii
60	Fabaceae-Mim.	Inga	sessilis
61	Fabaceae-Mim.	Leucaena	leucocephala
62	Fabaceae-Fab.	Lonchocarpus	subglaucescens
63	Fabaceae-Mim.	Mimosa	bimucronata
64	Fabaceae-Mim.	Mimosa	scabrella
65	Fabaceae-Fab.	Ormosia	excelsa
66	Fabaceae-Mim.	Parapiptadenia	rigida
67	Fabaceae-Mim.	Parkia	multijuga
68	Fabaceae-Mim.	Piptadenia	excelsa
69	Fabaceae-Mim.	Pithecellobium	jupunba
70	Rubiaceae	Psychotria	carthagenensis
71	Rubiaceae	Psychotria	longipes
72	Bignoniaceae	Tabebuia	rosea alba
73	Bignoniaceae	Tabebuia	sp
74	Oleaceae	Ligustrum	lucidum
75	Lauraceae	Nectandra	rigida
76	Lauraceae	Nectandra	sp
77	Lauraceae	Ocotea	porosa
78	Lauraceae	Persea	racemosa
79	Annonaceae	Porcelia	macrocarpa
80	Magnoliaceae	Magnolia	grandiflora
81	Magnoliaceae	Talauma	ovata
82	Melastomataceae	Tibouchiana	sellowiana
83	Myristicaceae	Virola	oleifera
84	Myrtaceae	Campomanesia	xanthocarpa
85	Myrtaceae	Eucalyptus	globulus

Table 2 Continued

ID	Family	Gender	Species
86	Myrtaceae	Eucalyptus	grandis
87	Myrtaceae	Eucalyptus	saligna
88	Myrtaceae	Myrcia	racemulosa
89	Vochysiaceae	Erisma	uncinatum
90	Vochysiaceae	Qualea	sp
91	Vochysiaceae	Vochysia	laurifolia
92	Proteaceae	Grevillea	robusta
93	Proteaceae	Grevillea	sp
94	Proteaceae	Roupala	sp
95	Moraceae	Bagassa	guianensis
96	Moraceae	Brosimum	alicastrum
97	Moraceae	Ficus	gomelleira
98	Rhamnaceae	Hovenia	dulcis
99	Rhamnaceae	Rhamnus	frangula
100	Rosaceae	Prunus	sellowii
101	Rosaceae	Prunus	serotina
102	Rubiaceae	Faramea	occidentalis
103	Meliaceae	Cabralea	canjerana
104	Meliaceae	Carapa	guianensis
105	Meliaceae	Cedrela	fissilis
106	Meliaceae	Khaya	ivorensis
107	Meliaceae	Melia	azedarach
108	Meliaceae	Swietenia	macrophylla
109	Rutaceae	Balfourodendron	riedelianum
110	Rutaceae	Citrus	aurantium
111	Rutaceae	Fagara	rhoifolia
112	Simaroubaceae	Simarouba	amara

Softwoods, in the 112 species database belong to 85 genera and 30 families.

The proposed database is presented in such a way as to allow work to be performed on different problems with different numbers of classes. For the experimental protocol, we suggest the following: 40% (eight images per species) for training, 20% (four images per species) for validation, and 40% (eight images per species) for testing: for example, images 001 through 008 for training, images 009 through 012 for validation, and images 013 through 020 for testing.

Figure 2 shows four different species in the database. It is worth noting that color cannot be used as an identifying feature in this database, since its hue depends on the current used to produce contrast in the microscopic images. All the images were therefore converted to gray scale (256 levels) in our experiments.

Looking at these samples, we can see that they have different structures. Hardwoods usually present some large holes, known as vessels (Fig. 2c, d), whereas Softwoods have a more homogeneous texture (Fig. 2b) or present smaller holes, known as resiniferous channels (Fig. 2a).

Another visual characteristic of a of the Softwood species is the growth ring, which is defined as the difference in the thickness of the cell walls resulting from the annual development of the tree. We can see this feature in Fig. 2b. The coarse cells at the bottom and top of the image indicate more intense physiological activity during spring and summer. The smaller cells in the middle of the image (highlighted in light red) represent the minimum physiological activity that occurs during autumn and winter.

3 Features

In this section, we describe the three feature sets we used to train the classifiers. The first was designed to explore the structure of the wood. As mentioned above, if we take a closer look at the images in the database, we can see that the two groups differ in their structure. The other two feature sets we used, GLCM and LBP, are frequently applied to solve texture classification problems. We describe these feature sets briefly below.

3.1 Structural features

One feature that is very often found in Hardwood species is the vessel (Fig. 2c, d), which constitutes the major part of the water transport system in the plant. Softwoods do not have such vessels, but resiniferous channels (Fig. 2a), which are very similar to vessels, are quite common in these species. Therefore, this feature alone is not sufficient to discriminate between the two groups. Another aspect of the Hardwood species is that they have a denser structure (smaller cells) than Softwoods. What we have observed in some experiments is that, after segmentation, the binary images of the Hardwood species contain large connected components. By contrast, Softwoods have a large number of small connected components. Therefore, the rationale behind the creation of this feature set is to detect those connected components and describe their structure by using simple statistical measures. For this reason, we call it a structural feature set.

The first step in extracting the proposed feature set is to binarize the image. To do this, we used an adaptive thresholding technique with a non overlapping window. The thresholding algorithm applied is the well-known Otsu method [9]. Three different window sizes (small [8 × 8], medium [50 × 50], and large [100 × 100]) were tried, and all were found to have little impact on performance. With the adaptive threshold, it is possible to tone down the darker corners produced during image acquisition, as depicted in Fig. 3a.

Next, the binary image is inverted, so that each cell, vessel, or resiniferous channel can form a connected component.

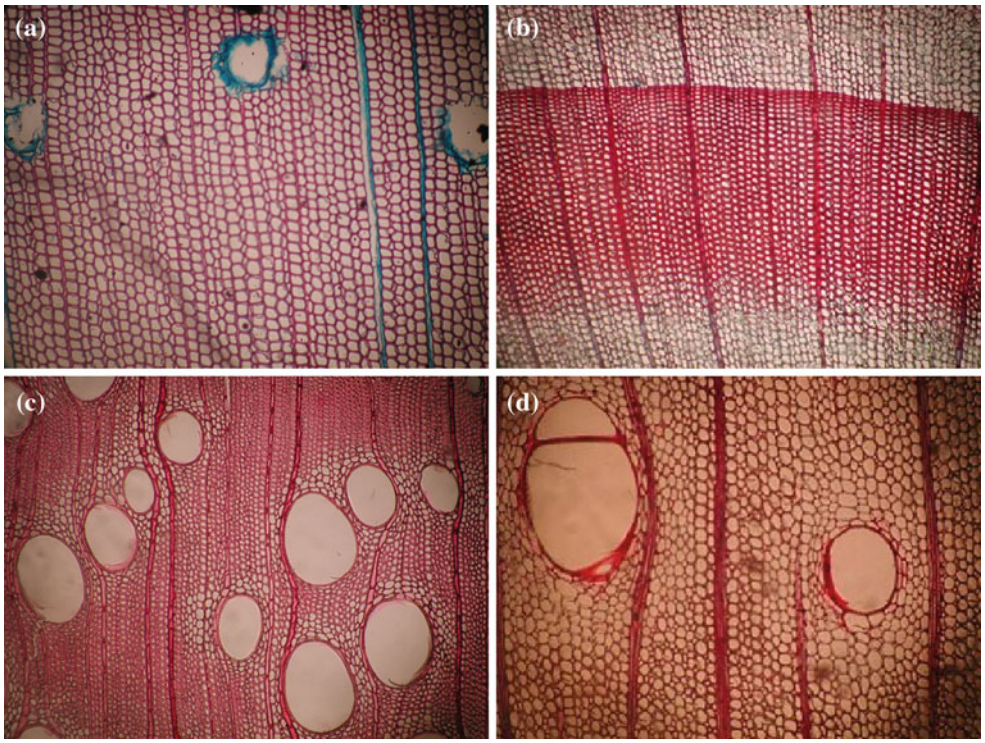


Fig. 2 Species of the database **a** 21, **b** 33, **c** 58, and **d** 95

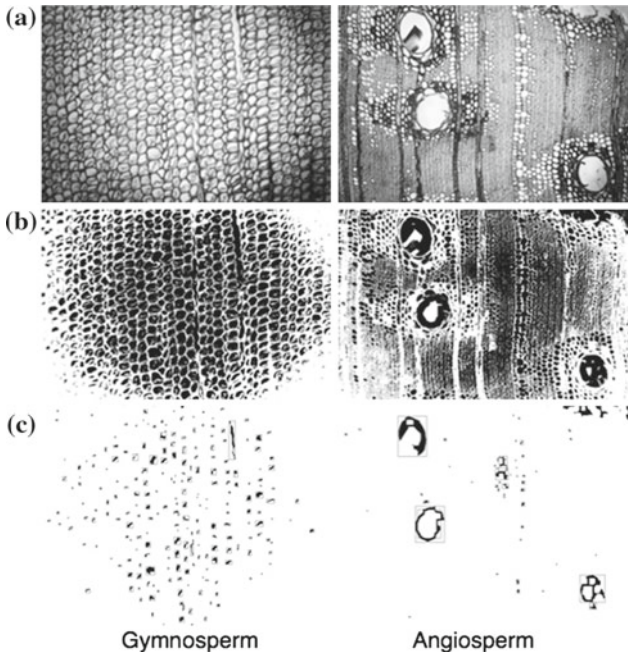


Fig. 3 Feature extraction. **a** Gray-scale images, **b** binary images, and **c** connected components after erosion

In order to keep the main connected components and separate those connected during the thresholding process, we applied an erosion, which is a basic operator in the area of mathematical morphology [10]. The structuring element employed was the square (3×3) and the number of iterations was determined empirically. In Sect. 4, we discuss the impact on the recognition rate of this operation and the number of iterations required.

Figure 3 depicts this process. As we can see in Fig. 3c, a large number of components were removed by the erosion process. We can also see that the Hardwoods lost more components during this process than the Softwoods. The number and size of the connected components allow us to discriminate between the two classes.

The next step is to compute the features for the remaining connected components. The feature vector is composed of the following five elements: number of connected components, average size of the connected components, variance, kurtosis, and obliquity. The features are then normalized using the Min–Max rule.

3.2 Gray level co-occurrence matrix (GLCM)

Among the statistical techniques available for texture recognition, the GLCM has been one of the most widely used and successful. This technique consists of statistical experiments conducted on how a certain level of gray occurs on other levels of gray [11]. It intuitively provides measures of properties such as smoothness, coarseness, and regularity. Haralick [12], who originated this technique, suggested a set of 14 characteristics, but most works in the literature

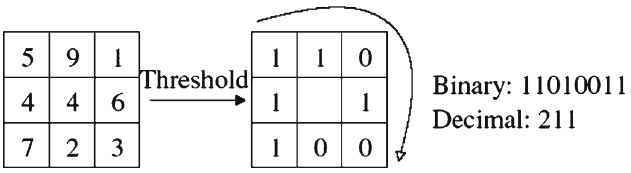


Fig. 4 The original LBP operator

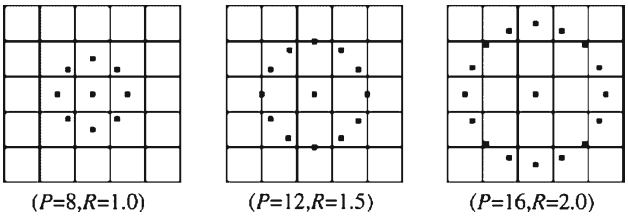


Fig. 5 The extended LBP operator [14]

consider a subset of these descriptors. In our case, we used the following six: Energy, Contrast, Entropy, Homogeneity, Maximum Likelihood, and 3rd Order Moment.

By definition, a GLCM is the probability of the joint occurrence of gray-levels i and j within a defined spatial relation in an image. That spatial relation is defined in terms of a distance d and an angle θ. From this GLCM, some statistical information can be extracted. Assuming that Ng is the gray-level depth, and $p(i, j)$ is the probability of the co-occurrence of gray-level i and gray-level j observing consecutive pixels at distance d and angle θ, we can use a GLCM to describe wood texture.

In our experiments, we tried different values of d, as well as different angles. The best setup we found is $d = 5$ and $\theta = [0, 45, 90, 135]$. Considering the six descriptors mentioned above, we arrive at a feature vector with 24 components.

3.3 Local Binary Patterns (LBP)

The original LBP proposed by Ojala et al. [13] labels the pixels of an image by thresholding a 3×3 neighborhood of each pixel with the center value. Then, considering the results as a binary number and the 256-bin histogram of the LBP labels computed over a region, they used this LBP as a texture descriptor. Figure 4 illustrates this process.

The limitation of the basic LBP operator is its small neighborhood, which cannot absorb the dominant features in large-scale structures. To overcome this problem, the operator was extended to cope with larger neighborhoods. By using circular neighborhoods and bilinearly interpolating the pixel values, any radius and any number of pixels in the neighborhood are allowed. Figure 5 depicts the extended LBP operator, where (P, R) stands for a neighborhood of P equally spaced sampling points on a circle of radius R, which forms a neighbor set that is symmetrical in a circular fashion.

The LPB operator $\mathrm{LBP}_{P,R}$ produces 2^P different binary patterns that can be formed by the P pixels in the neighbor set. However, certain bins contain more information than others, and so, it is possible to use only a subset of the 2^P LBPs. Those fundamental patterns are known as uniform patterns. A LBP is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. For example, 00000000, 001110000 and 11100001 are uniform patterns. It is observed that uniform patterns account for nearly 90% of all patterns in the (8,1) neighborhood and for about 70% of all patterns ib the (16, 2) neighborhood in texture images [13,15].

Accumulating the patterns that have more than two transitions into a single bin yields an LBP operator, denoted $\mathrm{LBP}^{u2}_{P,R}$, with fewer than 2^P bins. For example, the number of labels for a neighborhood of 8 pixels is 256 for the standard LBP but 59 for LBP^{u2}. Then, a histogram of the frequency of the different labels produced by the LBP operator can be built. We have tried out different configurations of LBP operators, but the one that produced the best results was the $\mathrm{LBP}^{u2}_{8,2}$, with a feature vector of 59 components.

4 Experiments and discussion

An important aspect of pattern recognition problems that is very often neglected is class distribution. A tacit assumption in the use of recognition rate as an evaluation metric is that the class distribution among examples is constant and relatively balanced. In the database we propose here, this is not the case. In our context, receiver operating characteristic (ROC) curves are attractive because they are insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change [16]. For this reason, we present the ROC curves and area under the curve (AUC) values for all the experiments. The AUC of a classifier has an important statistical property: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

The overall error rate that has been used for evaluation purposes in this work is given by Eq. 1.

$$\text{Overall error rate} = \frac{\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

where FP, FN, TP, and TN stand for false positive, false negative, true positive, and true negative, respectively. These statistics are defined in the 2×2 confusion matrix depicted in Fig. 6.

Hence, the recognition rate can be calculated using Eq. 2

$$\text{Recognition rate} = 100 - (\text{Overall error rate} \times 100) \quad (2)$$

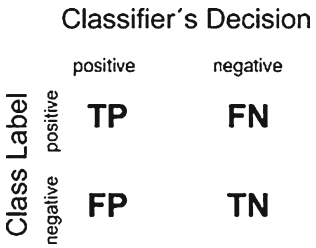


Fig. 6 2 × 2 confusion matrix

As pointed out earlier, three different classifiers were used to assess these feature sets: $k-$NN, LDA, and SVM. For the SVM, different kernels were tried, but the Gaussian kernel produced the best results. The kernel parameters γ and C were defined empirically through a grid search on the validation set. In our experiments, the database was divided into training (40%), validation (20%), and testing (40%), as suggested before, i.e. the validation set is used in a holdout validation scheme. In order to show that the choice of the images used in each subset does not have a significant impact on the recognition rate, each experiment was performed five times with different subsets (randomly selected) for training, validation, and testing. The small standard deviation values show that the choice of the images for each dataset is not an important issue.

The next two subsections report the experiments for the 2-class problem and the multi-class problem, respectively. Then in Sect. 4.3 we show some results on different configurations of the database.

4.1 2-Class problem

The first set of experiments was performed for the 2-class problem using the structural feature set. Our first concern was to determine the impact of the morphological operator on feature extraction, and, consequently, on classifier performance. To do this, we extracted features using different numbers of iterations for the erosion operator. Figure 7 shows the impact in terms of performance on the validation set for the three classifiers. As we can see, all the classifiers behave in a similar way to the different numbers of iterations used with the erosion operator. For a small number of iterations, a large number of connected components is considered for feature extraction. Aggressive erosion, by contrast, removes important components. This, of course, drastically reduces the performance of all the classifiers. We found that the best compromise was achieved using six iterations.

The best results in this experiment were achieved by the SVM and k-NN with 93.1 and 92.6% of the recognition rate on the test set respectively. In the case of the k-NN, the best value for k was 3. The LDA classifier had the worst performance, at 88.9% of the recognition rate. In all cases, we

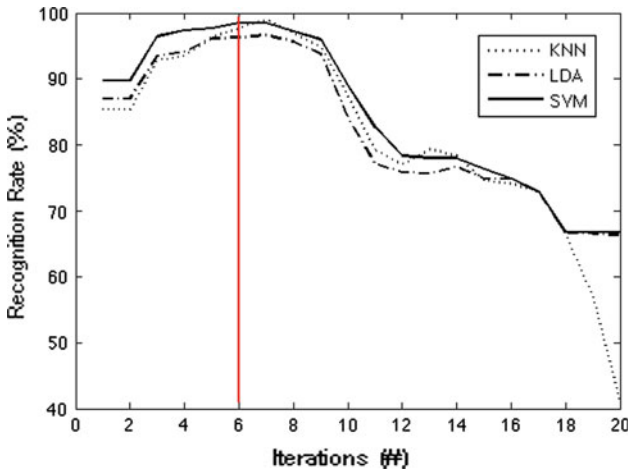


Fig. 7 Impact of the erosion in the performance of the classifiers on the validation set

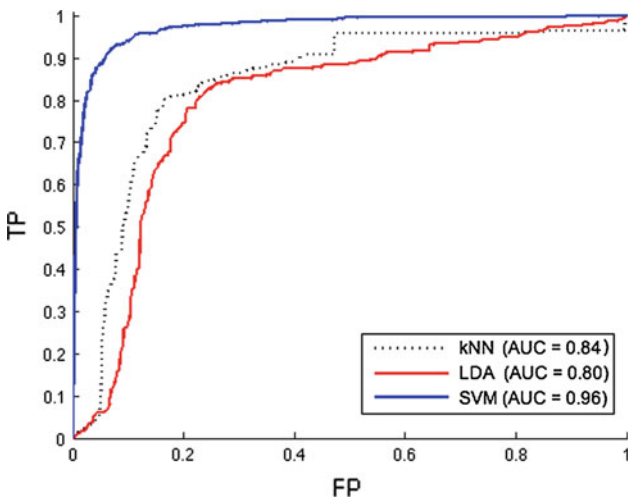


Fig. 8 ROC curves for the classifiers trained with structural features on the testing set

considered the features extracted from the images after six iterations of the erosion operator. The results also show that the size of the window used for the adaptive thresholding has little impact on the final recognition rate. Figure 8 shows the ROC curves and AUC values for all the classifiers. In spite of the similar performance achieved by the SVM and k-NN classifiers in terms of recognition rate, we can see from Fig. 8 that the k-NN produces much higher false positive (FP) rates for any given true positive (TP) rate. Table 3 shows the confusion matrices for the classifiers trained with structural features.

In the second series of experiments, all the classifiers were trained with the GLCM feature set. As in the previous experiments, the SVM classifier outperformed the other classifiers, achieving a recognition rate of 97.4%. The LDA classifier performed much better with this feature set, at 95.0%, however the k-NN classifier seems unsuitable for this

Table 3 Confusion matrices (in %) for the classifiers trained with structural features

	kNN		LDA		SVM	
	Softwood	Hardwood	Softwood	Hardwood	Softwood	Hardwood
Softwood	89.7	10.3	85.3	14.7	88.1	11.9
Hardwood	5.9	94.1	9.3	90.7	4.5	95.5

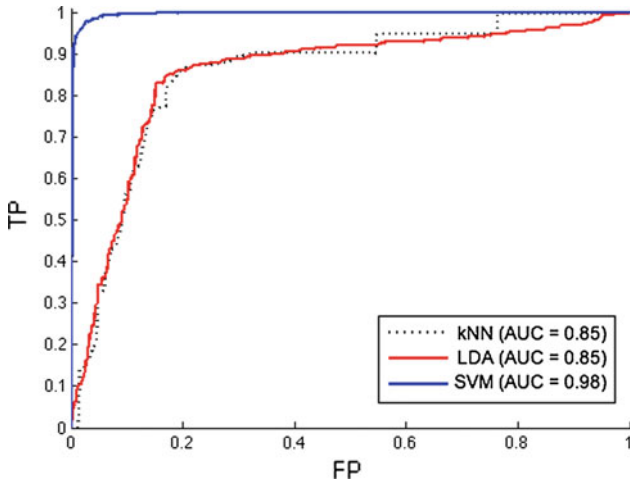


Fig. 9 ROC curves for the classifiers trained with GLCM features

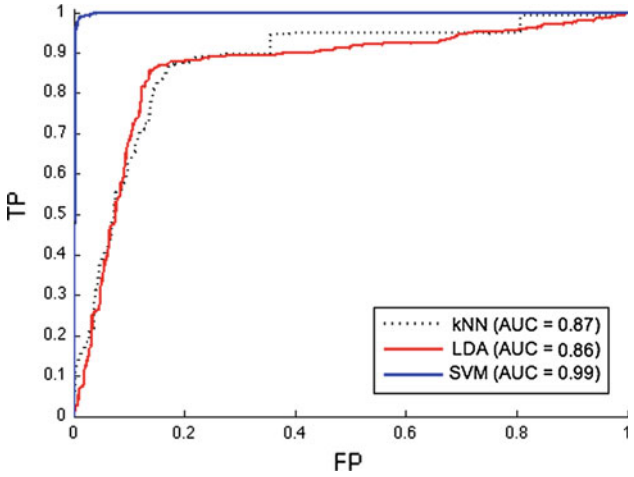


Fig. 10 ROC curves for the classifiers trained with LBP features on the testing set

representation, as it produces no better than an 89% recognition rate. The superior performance of the SVM classifier had an impact on the ROC curve, producing an AUC value of 0.98. The same cannot be said for the LDA, which improved the recognition rate substantially, by about six percentage points, but had little impact on the ROC curve. Figure 9 depicts the ROC curves for the classifiers trained with GLCM. Table 4 shows the confusion matrices for the classifiers trained with GLCM features.

In the last set of experiments, we trained the classifiers using the LPB feature set. As pointed out earlier, different configurations of LBP operators were tried, but the one that produced the best results was $LBP^{u2}_{8,2}$, with a feature vector of 59 components. This feature set generates the biggest feature vector; however, achieving the best performance compensates for this. For example, the SVM classifier achieved a performance of 98.6%, while LDA and k-NN achieved performance values of 95.8 and 92.3%, respectively. In spite of this, we noted no significant improvement in the AUC values for either LDA or k-NN. The AUC value for the SVM, by contrast, is 0.999. Figure 10 presents the ROC curves for the classifiers from this last experiment. Table 5 shows the confusion matrices for the classifiers trained with LBP features.

Table 6 summarizes the results of all the experiments, reporting the recognition rates achieved, as well as the standard deviations, on the test set.

In spite of the good results produced by the SVM trained with LBP, there is still some confusion that must be resolved. Figure 11 presents two misclassified samples. In Fig. 11a, the Gymnosperm *Cedrus sp* was confused with an Angiosperm, mainly because of its structure, which that features long and well-defined veins. These veins are also common in Angiosperms. The reverse occurs as well, as depicted in Fig. 11b. In this case, the vessels of Angiosperms *Tibouchiana sellowiana* are quite small, and similar to the resiniferous channels found in Gymnosperms. One point worth noting, though, is that not all the classifiers make the same mistakes. Therefore, combining different classifiers and feature sets could resolve some of the confusion.

Table 4 Confusion matrices (in %) for the classifiers trained with GLCM features

	kNN		LDA		SVM	
	Softwood	Hardwood	Softwood	Hardwood	Softwood	Hardwood
Softwood	67.8	32.2	89.7	10.3	97.6	2.4
Hardwood	0.5	99.5	2.3	97.7	2.8	97.2

Table 5 Confusion matrices (in %) for the classifiers trained with LBP features

	kNN		LDA		SVM	
	Softwood	Hardwood	Softwood	Hardwood	Softwood	Hardwood
Softwood	79.2	20.8	93.0	7.0	97.2	2.8
Hardwood	1.1	98.8	2.9	97.1	0.7	99.3

Table 6 Summary of all experiments on the testing set

Feature set	Number of features	Rec. rate (%)					
		k-NN	σ	LDA	σ	SVM	σ
Structural	6	92.6	0.8	88.9	0.5	93.1	0.8
GLCM	24	89.0	1.2	95.0	0.1	97.4	0.6
LBP	59	92.3	0.5	95.8	1.1	98.6	0.5

Table 7 Recognition rates for the multi-class problem

Feature set	Number of features	Rec. rate (%)					
		k-NN	σ	LDA	σ	SVM	σ
GLCM	24	46.6	0.6	60.6	1.4	55.3	1.2
LBP	59	70.1	0.5	80.7	0.2	79.3	1.4

Table 8 Different database configurations

Number of pieces	Image size	Number of images
2	512 × 768	4,480
3	341 × 768	6,720
4	512 × 384	8,960
6	341 × 384	13,440
9	341 × 256	20,160

4.2 Multi-class problem

In this experiment, our objective was to analyze the usefulness and reliability of the proposed database. To do that, we expanded the data analysis into a multi-class problem. With this in mind, we used the same protocol as in the previous experiment, but replaced the 2-class classifiers with multi-class classifiers, which were built to discriminate among the 112 classes available. In the case of the SVM classifier, two strategies are commonly used to deal with multi-class problems: pairwise, and one-against-others. In this work, we used the one-against-others strategy, which works by constructing an SVM ω_i for each class q that first separates the class from

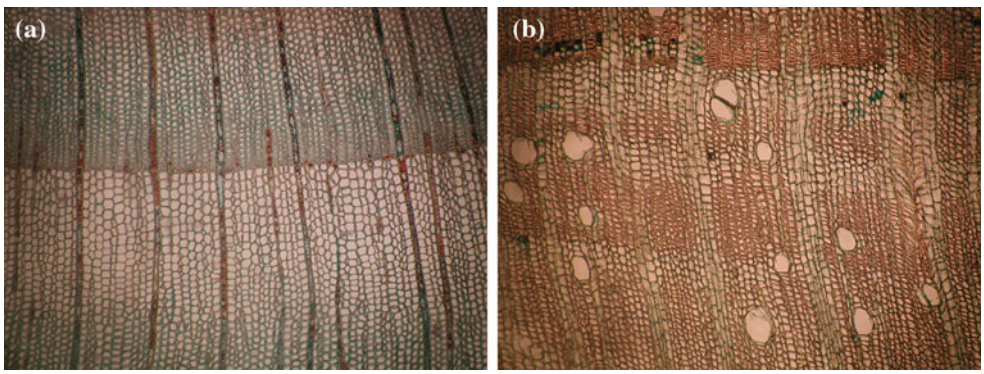


Fig. 11 Some confusions between the two classes

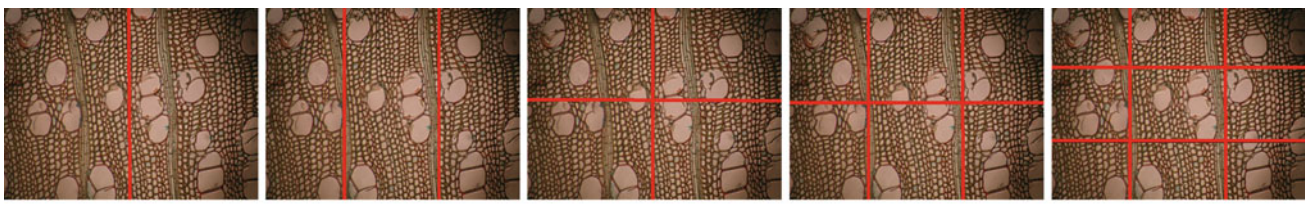


Fig. 12 Dividing the original image into smaller pieces

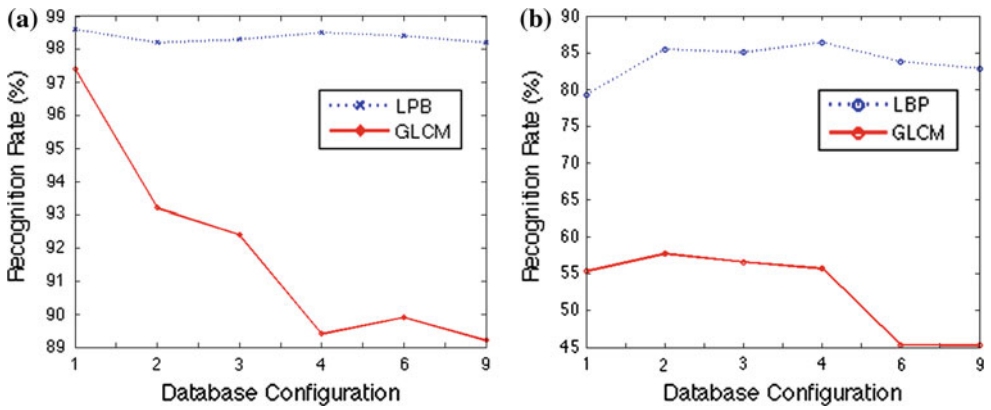


Fig. 13 Performance of the classifiers on the different configurations of the database. **a** 2-class problem and **b** multi-class problem

all the other classes and then uses an expert F to arbitrate between their SVM outputs, in order to produce the final decision. A good reference work listing the multi-class SVM methods is [17].

Table 7 reports the results achieved by the classifiers trained with GLCM and LBP for the multi-class problem. The classifier trained with structural features achieved very poor results, and for this reason it is not reported here.

Like the results of the previous experiments, the results for the classifiers trained with LBP surpassed those for the other classifiers. However, in this experiment, where there is much greater complexity, the LPB proved to be an excellent texture descriptor, superior to the classifiers trained with GLCM by about 20 percentage points. The best results for the 112 classes were achieved by the SVM and LDA classifiers, both of which achieved a recognition rate of about 80%.

4.3 Working with smaller images

As explained earlier, the proposed database contains 20 images of 1,024 × 768 pixels per species, totaling 2,240 images. A larger number of samples can be useful in a number of situations, such as increasing or creating more validation sets that can be used for feature selection, for example. Since acquiring more data is not a trivial task, we divided the original images into several chunks, so that larger databases could be created, as shown in Fig. 12. These databases are reported in Table 8.

To determine the impact of dividing the original image into smaller chunks, we used the best classifier we found in the previous experiments, i.e. the SVM trained with the two feature sets that yielded the best results (LBP and GLCM). Preliminary tests with the structural feature set on smaller images considerably reduced the performance of the classifiers. This is mainly because important structural information can find its way into several of those smaller pieces, which compromises the structural feature vector. This is why we did not pursue the experiments with the structural feature set.

For the experiments with GLCM and LBP on smaller images, we followed the same scheme for dividing the images for training, validation, and testing, i.e. 40, 20, and 40% respectively. In Fig. 13, we can compare the recognition rates on the testing set for both feature sets for all database configurations.

We can see from Fig. 13a, b that the performance of the SVMs trained with LBP is homogeneous across all database configurations, from which we conclude that LBP is able to extract the necessary information for classification from the smaller images. In the case of the multi-class problem, increasing the number of images improves the SVM classifier. The best result in this case was achieved using configuration #4 (Fig. 12c), with a recognition rate of 86%.

The same is not true for the classifiers trained with GLCM. In the case of the 2-class problem, no configuration provided any improvement. In the case of the multi-class problem, we can see a slight improvement for configurations #2 and #3. However, the enormous difference between GLCM and LBP makes this irrelevant.

5 Conclusion

In this paper, we have introduced a new database composed of 2,240 microscopic images of 112 different species divided into 2 groups (Hardwoods and Softwoods), 85 genera, and 30 families. A comprehensive set of experiments using three different feature sets and three different classifiers has shown that an SVM trained with a 59-dimensional LBP feature vector is a good option for both 2-class and multi-class problems. We also proposed another five configurations of the database with a larger number of images, and show that the same tuple SVM-LBP, is able to maintain a homogeneous performance in all five configurations.

The results for the 2-class problem presented in this work are very promising, since they can be used as a preclassification step in any forest species recognition system, and

reduce the possible confusion between species of different genera and families. In future work, we plan to refine and test other feature sets to solve any remaining confusion, and use this dataset to build a system to automatically identify forest species. Our expectation is that this database will contribute to the field of forest species recognition and motivate more researchers to work in this field.

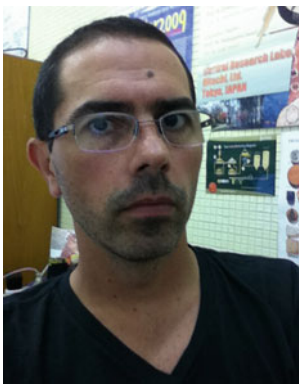
Acknowledgments This work have been supported by The National Council for Scientific and Technological Development (CNPq)—Brazil grant #301653/2011-9 and the Coordination for the Improvement of Higher Level Personnel (CAPES).

References

1. Thomas, L., Milli, L.: A robust gm-estimator for the automated detection of external defects on barked hardwood logs and stems. IEEE Trans. Signal Proc. **55**, 3568–3576 (2007)
2. Huber, H.A.: A computerized economic comparison of a conventional furniture rough mill with a new system of processing. For. Prod. J. **21**(2), 34–39 (1971)
3. Buechler, D.N., Misra, D.K.: Subsurface detection of small voids in low-loss solids. In: First ISA/IEEE Conference Sensor for Industry, pp. 281–284 (2001)
4. Tou, J.Y., Lau, P.Y., Tay Y.H.: Computer vision-based wood recognition system. In : Proceedings of International Workshop on Advanced Image Technology (2007)
5. Tou, J.Y., Tay, Y.H., Lau, P.Y.: One-dimensional grey-level co-occurrence matrices for texture classification. In: International Symposium on Information Technology (ITSim 2008), pp. 1–6 (2008)
6. Tou, J.Y., Tay, Y.H., Lau, P.Y.: A comparative study for texture classification techniques on wood species recognition problem. Int. Confer. Nat. Comput. **5**, 8–12 (2009)
7. Khalid, M., Lee, E.L.Y., Yusof, R., Nadaraj, M.: Design of an intelligent wood species recognition system. IJSSST. **9**(3) (2008)
8. Paula, P.L., Oliveira, L.S., Britto, A.S., Sabourin, R.: Forest species recognition using color-based features. In: International Conference on Pattern Recognition, pp. 4178–4181 (2010)
9. Otsu, N.: A threshold selection method from gray-level histogram. IEEE Trans. Syst. Man Cybern. **8**, 62–66 (1978)
10. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, New York (1984)
11. Tuceryan, M., Jain, A.K.: Texture analysis. The Handbook of Pattern Recognition and Computer Vision. World Scientific, Singapore (1998)
12. Haralick, R.M.: Statistical and structural approaches to texture. Proc. IEEE **67**(5) (1979)
13. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distribution. Pattern Recogn. **29**(1), 51–59 (1996)
14. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**, 803–816 (2009)
15. Ojala, T., Pietikainen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
16. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **227**(8), 861–874 (2006)
17. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multi-class support vector machines. IEEE Trans. Neural Netw. **13**, 415–425 (2002)

Author Biographies

J. Martins received his B.S. degree in Informatics from the Western Paraná State University (UNIOESTE), 2000, Cascavel, PR, Brazil, and the M.E. degree in Production Engineering from the Federal University of Santa Catarina (UFSC), 2002, Florianópolis, SC, Brazil. In 2003, he joined the Technological Federal University of Parana (UTFPR) where he is currently an assistant professor. Since 2010, he is a Ph.D. candidate at Federal University of Parana (UFPR) investigating forest species recognition based on texture features and microscopic images. His research interests include Pattern Recognition and Image Analysis.



L. S. Oliveira received his B.S. degree in Computer Science from UnicenP, Curitiba, PR, Brazil, the M.Sc. degree in electrical engineering and industrial informatics from the Centro Federal de Educação Tecnologica do Paraná (CEFET-PR), Curitiba, PR, Brazil, and Ph.D. degree in Computer Science from Ecole de Technologie Superieure, Universite du Quebec in 1995, 1998 and 2003, respectively. From 2004 to 2009 he was professor of the Computer Science Department at Pontifical Catholic University of Parana, Curitiba, PR, Brazil. In 2009, he joined the Federal University of Parana, Curitiba, PR, Brazil, where he is professor of the Department of Informatics and head of the Graduate Program in Computer Science. His current interests include Pattern Recognition, Machine Learning, Image Analysis, and Evolutionary Computation.

S. Nisgoski is Forest Engineer, received her M.Sc. and the Ph.D. degrees in Forest Science, on Forest Products Utilization Technology from Federal University of Parana, Brazil, in 1999 and 2005, respectively. Since 2009, she is an adjunct professor on Forest and Engineering Department of same institution. Her research interests include wood and charcoal anatomy and identification and material characterization by non destructive techniques

R. Sabourin received his B.Ing., M.Sc.A., and Ph.D. degrees in Electrical Engineering from the Ecole Polytechnique de Montreal in 1977, 1980 and 1991, respectively. In 1977, he joined the physics department of the Universite de Montreal where he was responsible for the design and development of scientific instrumentation for the Observatoire du Mont Megantic. In 1983, he joined the staff of the Ecole de Technologie Superieure, Universite du Quebec, Montreal, P.Q. Canada, where he is currently a professeur titulaire in the Departement de Genie de la Production Automatisee. In 1995, he joined also the Computer Science Department of the Pontificia Universidade Catolica do Parana (PUC-PR, Curitiba, Brazil) where he was corresponsible since 1998 for the implementation of a Ph.D. program in Applied Informatics. Since 1996, he is a senior member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI). His research interests are in the areas of handwriting recognition and signature verification for banking and postal applications.