Bayesian Classification of Cork Stoppers Using Class-Conditional Independent Component Analysis

Jordi Vitrià, Marco Bressan, and Petia Radeva, *Member, IEEE*

***Abstract*—In this paper, a real-time application for visual inspection and classification of cork stoppers is presented. The process of cork inspection and quality grading is based on analyzing a large set of characteristics corresponding to visual features that are related to cork porosity. We have applied a set of nonparametric and parametric classification methods for comparing and evaluating their performance in this real problem. The best results have been achieved using Bayesian classification through probabilistic modeling in a high-dimensional space. In this context, it is well known that high dimensionality represents a serious problem for density estimation. We propose a class-conditional independent component analysis representation of the data that allows an accurate estimation of the data probability density function by factorizing it. The method has achieved a success of 98% of correct classification.**

***Index Terms*—Independent component analysis, machine vision, object recognition, visual inspection.**

I. Introduction

Automatic inspection of cork stoppers is a complex problem influenced by many objective and subjective factors, and as a consequence, it represents the least automated task in the production cycle of the cork stopper in the cork industry. Due to the inspection difficulty and the high production rates in the manufacture even the most experienced quality inspection operators frequently make mistakes. In addition, it is increasingly difficult to find labor willing and able to do a job that is at the same time both skilled and highly repetitive. Moreover, human inspection leads to a lack of objectivity and uniformity in the rules applied by different people at different times. As a result, there is a urgent need to modernize the cork industry in this direction. In this paper, we consider a real machine vision application for classification of natural cork stoppers by using a novel statistical classification method based on independent component analysis (ICA) [1].

Cork stoppers are routinely classified in different quality classes (see Fig. 1) and inspected for different defects like small holes due to insect attacks, stopper breaking, or cracks (see Fig. 2). Although it does not seem hard for human beings to visually analyze the cork stopper, it turns out difficult to precisely formulate the characteristics of the cork surface that are relevant to these tasks due to the porosity of the natural material. It is difficult even for the cork quality experts to exactly define all cork features that they take into account in the process of quality grading or define the feature values and ranges in order to take a final decision about the presence of a cork defect.

Manuscript received February 1, 2004; revised August 4, 2004. This work was supported in part by MCYT Grant TIC2003-00654, Spain.

The authors are with the Computer Vision Center, Departament d'Informàtica, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain.

Digital Object Identifier 10.1109/TSMCC.2006.876043

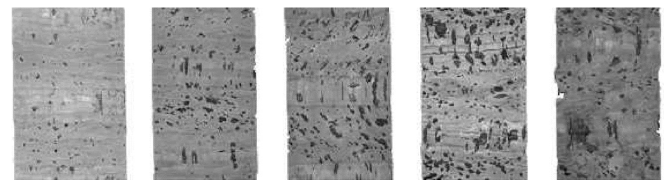

Fig. 1. Surfaces of cork stoppers of five quality groups ordered in best to worst quality (from left to right).

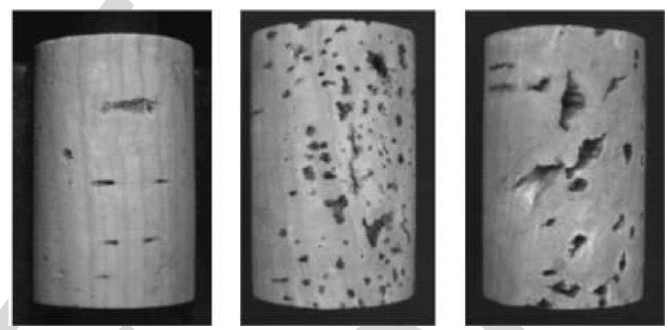

Fig. 2. Cork stopper without defects, woody stopper, and a stopper with a crack.

In order to cope with the problem of subjectivity in the process of cork quality classification, we study the applicability of different techniques from the fields of computer vision and pattern recognition to the cork analysis. We state that the problem of cork analysis is based on considering a large set of visual features (in our case more than 40 blob features, e.g., number of cork holes, average stopper gray level, average holes gray level, etc.). Thus, the problem of cork inspection is defined as a problem of learning and classification in a high-dimensional feature space.

Learning and classification in a high-dimensional space is a difficult task due to the problem known as the *curse of dimensionality*. The main inconvenient faced by statistical classification methods is proper density estimation due to dimensionality concerns. Additionaly, data can be contaminated by noise, and not necessarily all measurements contribute to classification. The classical approach to this problem is based on the application of feature extraction or selection techniques for modeling low-dimensional pattern structures or selecting the most relevant dimensions. Transformed data are low dimensional and a whole set of techniques can then be directly applied.

The most common approach for transforming data can be defined so as to maximally reduce dimensionality while preserving relevant information. This approach can be seen as a particular case of noise reduction, where anything not contributing to classification is considered noise. Discriminant

IEEE Proof

1094-6977/$20.00 © 2006 IEEE

analysis or principal component techniques usually take this approach. But a more direct method can result by considering linear transformations exclusively from the simplification they might provide on density estimation. ICA can provide this simplicity in terms of density marginalization: Density estimation in feature space is reduced to a number of unidimensional estimations. Additionally, higher order dependencies between the features are removed allowing single feature interpretations.

After extensive analysis of different existing techniques for the classification of high-dimensional data, we show that ICA represents a technique with optimal performance for our problem. In order to improve the density estimation of our data we define a class-conditional independent component analysis (CC-ICA) representation of the data that assures fast and robust approach for automatic inspection of cork porosity.

This paper is organized as follows: The problem of high dimensional data classification is presented in Section II, where our approach is exposed. Section III describes the visual features used in the characterization of cork stoppers. Results and comparisons to other methods are presented in Section IV. Finally, in Section V we expose our conclusions.

II. Classification of High-Dimensional Data

Trying to describe and classify cork stoppers from their visual appearance leads to addressing a classification problem in a high-dimensional space. High-dimensional data appear in many pattern recognition problems, such as remote sensing, appearance-based object recognition, text categorization, etc. A stochastic approach for the classification of high-dimensional data is always a delicate issue. For linear or quadratic classifiers, the number of training samples depends linearly or quadratically on the data dimensionality subset of features, and for selecting the best features, an exhaustive sequential feature selection procedure is required so that the size of the problem grows combinatorially on the dimension. Furthermore, the training sample size needs to increase exponentially in order to effectively estimate multivariate densities needed to perform nonparametric classification. To avoid the problem of dimensionality, the most common approach is the implementation of feature extraction or dimensionality reduction algorithms. Principal component analysis (PCA) [3] is widely used due to its noise-reduction properties. PCA treats the data as if they belong to a single distribution and so it has nothing to do with discriminative features optimal for classification. Linear discriminant analysis (LDA) [2] can be used to derive a discriminative transformation that is optimal for certain cases. LDA makes use of only second-order statistical information; hence, if the difference in the class mean vectors is small, the features chosen will not be reliable. Moreover, by definition LDA cannot produce more features than the number of classes involved. Feature subset selection [4] is yet another perspective on feature extraction. This problem considers a subset of all linear combinations of the original feature set (in our case, more than 40 features!), according to a certain criterion. In order to produce an optimal subset of features, an exhaustive feature selection procedure is required and so the size of the problem grows combinatorially on the dimension.

Since there is no golden rule for choosing a dimensionality reduction approach given an implementation of a parametric or nonparametric technique for density estimation, we propose to focus on the higher level statistical properties of the data. These data will be transformed in such a way that density estimation in the transformed space is simplified and more accurate. For this purpose, we consider an ICA [5] representation for each class. This representation projects our data into a space where the components have maximized statistical independence, and in many real problems, sparse distributions. As a result, the independence turns an M-dimensional density estimation into M one-dimensional (1-D) estimations of our data. This allows simple density estimation without relying on dimensionality reduction techniques.

A. Classification of High-Dimensional Data by CC-ICA

The classification problem of cork stoppers is stated as follows: Given a set of quality classes $C_1, C_2, \ldots, C_K$ (in our case, five) and a set of labeled cork stoppers, we want to learn a data representation space where the data density function can be estimated with higher reliability than in the original space and to assign a sample object $\boldsymbol{h}_{\mathbf{test}}$ to a particular class using the probability of misclassification as an error measure. It is well known that the solution to this problem is to assign $\boldsymbol{h}_{\mathbf{test}}$ to the class that maximizes the *posterior probability*. This is called the maximum a posteriori or MAP solution. Using the Bayes rule we can formulate the posterior probability in terms of quantities that are easier to estimate, and the MAP solution takes the form

$$C_{\mathrm{MAP}} = \arg_{k=1,\ldots,K} \max\{P(\boldsymbol{h}_{\mathbf{test}} \,|\, C_k) P(C_k)\}$$

where $P(C_k)$ is usually called the *prior* probability, $P(\boldsymbol{h}_{\mathrm{test}} \,|\, C_k)$ is referred to as the *class-conditional probability* or, when seen as a function of the parameters, the *likelihood*. In practice, the class-conditional probabilities can be modeled parametrically or nonparametrically from our training set. The priors, as their name indicates, are estimated from our prior knowledge of the problem, and if such knowledge is not available, equiprobable priors are usually assumed.

In our problem, the test object $\boldsymbol{h}_{\mathrm{test}}$ is represented by a high-dimensional feature vector $h_1, \ldots, h_D$. If we assume conditional independence in the occurrence of a particular value for the feature vector, the MAP solution takes a new form commonly known as the naive Bayes rule

$$C_{\text{Naive}} = \arg_{k=1,\ldots,K} \max \prod_{l=1}^{D} P(h_l \,|\, C_k) P(C_k).$$

We choose to represent the data from each class using the transform provided by ICA. This linear transform represents our data in a space where the statistical dependence between the components is minimized [5], [6]. After introducing the ICA model and detailing this statistical technique, we show how the assumption of independence greatly simplifies density estimation and analyze the consequences in Bayesian decision.

The ICA of a D-dimensional random vector is a linear transform that minimizes the statistical dependence between its components. This representation in terms of independence proves

IEEE Proof

useful in an important number of applications such as data analysis and compression, blind source separation, blind deconvolution, denoising, etc. Assuming the random vector we wish to represent through ICA has no noise and is zero centered, the ICA model can be expressed as

Q1

$$\mathbf{h} = \mathbf{A}\mathbf{s}$$

where $\mathbf{h} = (h_1, h_2, \ldots, h_D)$ is the random vector representing our data, $\mathbf{s} = (s_1, s_2, \ldots, s_M)$ is the random vector of independent components with dimension $M \leq D$, and $\mathbf{A}$ is a $D \times M$ matrix called the mixing matrix. To avoid ambiguities, the mixing matrix is chosen such that the independent components have unit variance (they already are zero centered). The pseudoinverse of $\mathbf{A}$, which we will represent as $\mathbf{W}$, is called the filter, projection, or demixing matrix, and it provides an alternative representation of the ICA model

$$\mathbf{W}\mathbf{h} = \mathbf{s}. \tag{1}$$

Under certain assumptions stated in [5], $\mathbf{W}$ is completely determined and, as its name indicates, projects our data into the ICA space. Several objective functions have been proposed for the estimation of the projection matrix. Basically, the likelihood, network entropy, mutual information, and approximations of these [7]–[9]. So far, the most satisfying and intuitive contrast function is mutual information, a natural measure for the independence of random variables. It is defined as

$$\mathcal{I}(s_1, s_2, \ldots, s_M) = \sum_M \mathcal{H}(s_m) - \mathcal{H}(\mathbf{s}) \tag{2}$$

where

$$\mathcal{H}(\mathbf{s}) = -\int_{-\infty}^{\infty} f(\mathbf{s}) \, \log(f(\mathbf{s})) \, d\mathbf{s} \tag{3}$$

is the differential entropy of random vector $\mathbf{s}$ with density $f(\mathbf{s})$. The mutual information is always nonnegative and zero if and only if the variables are statistically independent. The main drawback of mutual information is its estimation since the definition of entropy requires the estimation of the density of the independent components. Some authors have used approximations of the mutual information based on polynomial density expansions [5], [10] such as the Edgeworth expansion [11]. Approaches using approximations of the differential entropy were also suggested [12]. These approximations can be used for the minimization of mutual information, estimating the individual independent components as projection pursuit directions. In [13], together with the theory, Hyvärinen and Oja present a fast and reliable implementation of this scheme through the FastICA algorithm. This algorithm proves efficient and robust regardless of the dimension and density of the sources and is the algorithm of choice for the estimation of the ICA model in this study.

When the conditions for performing ICA on a certain dataset hold, density estimation of the projected data is strongly simplified. From a statistical perspective, the product of the marginal densities of the projected data (independent components) best fits the global distribution for the observations. But straightforward application of ICA for classification, i.e., unsupervisedly learning ICA from the dataset and working with the marginal densities of the independent components are incorrect. The Bayesian classification scheme makes use of the class-conditional densities and, as we will see, global independence does not imply class-conditional independence. This is why a representation that takes into account class belonging is required. This is why this representation is known as CA-ICA.

Let x and y be random variables taking values in $\Re$. And let $p(x, y)$, $p(x)$, $p(y)$, and $p(x \mid y)$ be, respectively, the joint density of (x, y), the marginal densities of x and y, and the conditional density of x given the value of y. We say that x and y are independent if any of the following two equivalent definitions hold

$$p(x, y) = p(x)p(y) \tag{4}$$

$$p(x \mid y) = p(x). \tag{5}$$

It proves useful to understand independence from the following statement derived from (5): Two variables are independent when the value one variable takes gives us no knowledge on the value of the other variable. For the multivariate case $\mathbf{x} = (x_1, \ldots, x_N)$, independence can be defined by extending (4) as $p(\mathbf{x}) = p(x_1) \ldots p(x_N)$. Conditional independence is defined as a natural extension of (4) and (5) through the incorporation of the conditional operator: $p(x, y \mid z) = p(x \mid z)p(y \mid z)$ and, equivalently, $p(x \mid y, z) = p(x \mid z)$. A frequent mistake is to think that global independence implies conditional independence, being Simpson's paradox [14] probably the most well-known counterexample.

Given K classes $C_1, \ldots, C_K$ in a D-dimensional space, the ICA model is estimated from the training set for each class. If $\mathbf{W_k}$ and $\mathbf{s_k}$ are the projection matrix and the independent components for class C_k with dimensions $M_k \times D$ and M_k, respectively, then

$$\mathbf{s}^k = \mathbf{W}^k(\mathbf{h} - \overline{\mathbf{h}^k}) \tag{6}$$

where $\mathbf{h}$ belongs to C_k and $\overline{\mathbf{h}^k}$ is the mean of the class, estimated from the training set. We use the following notation for the density distribution of the independent components $p^k(\mathbf{s}) \stackrel{\text{def}}{=} p(\mathbf{s}^k)$. The density distribution of the projected data can be rewritten using the independence assumption for the independent components

$$p^k(\mathbf{s}) = \prod_{m=1}^{M_k} p^k(s_m). \tag{7}$$

When using a Bayesian scheme for comparing different models, special care should be taken on the normalization of the class-conditional probabilities. If Ω is the set of all possible projected feature vectors, we have that for all k

$$\int_\Omega p^k(\mathbf{s}) \, d\mathbf{s} = 1.$$

For a training set $T_k \in \Omega_k$, it is recommendable to introduce a normalization constant ν_k on each class-conditional probability

IEEE Proof

defined as

$$\nu_k = \left(\frac{1}{\#T} \sum_{s \in T} p^k(\mathbf{s}) \right)^{-1}$$

and to use the normalized likelihood $\nu_k p^k(\mathbf{s})$.

If $\mathbf{h}_l$ is a representative feature vector for a test object, we can project it into the ICA model learnt for class C_k using (6) and obtain the independent components $\mathbf{s}_l^k$. The likelihood of the representative feature vector is obtained from the easier to calculate likelihood of the transformed feature vector. Using the log-likelihood to turn the products into sums and incorporating (7), the Naive Bayes rule can be rewritten as

$$C_{\text{Naive}} = \arg \max_{k=1 \ldots K} \sum_{l=1}^{L} \left(\sum_{m=1}^{M_k} \log p^k(s_{lm}) \right) + L\nu_k. \quad (8)$$

In practice, classification is performed as follows: Representative feature vectors are extracted from the objects belonging to class C_k, conforming training set T_k. T_k is then used to estimate the ICA model and projected into this model. From the projected feature vectors, the M_k 1-D densities are estimated, together with the normalization constants. If we have no prior information on these marginal distributions, nonparametric or semiparametric methods can be used in the 1-D estimation.[1] Given a test object, its representative feature vectors are projected on each class and the class-conditional likelihoods calculated. The test object is assigned to the class with the highest probability.

III. Cork Stopper Image Analysis

Given the problem of automatic classification of cork stoppers, we need to extract a set of visual features that describes the cork porosity and allow to classify the stoppers in five different quality groups. It has been clearly stated by visual inspection that this problem is more related to analyze the distribution of blob characteristics than to texture analysis. In fact, human inspectors describe cork quality using a large set of blob features. Moreover, production requirements stated that cork stopper analysis system should allow the inspection of four stoppers per second.

A. Feature Extraction and Analysis

In order to construct robust algorithms to classify cork stoppers, original images should be preprocessed and segmented. The lateral part image of cork stoppers is acquired by a high-resolution linear camera [see Fig. 3 (left)] during the continuous movement of the cork stopper by the machine conveyor. The image is geometrically corrected by applying linear regression to determine the image distortion that is introduced by stopper linear movement with respect to the camera. Using simple correlation based image processing techniques the overlapping parts of the stopper image are detected and removed [see Fig. 3 (right)]. Afterwards, standard image enhancing techniques are applied and an adaptive threshold is used to binarize the blobs of the cork stoppers.

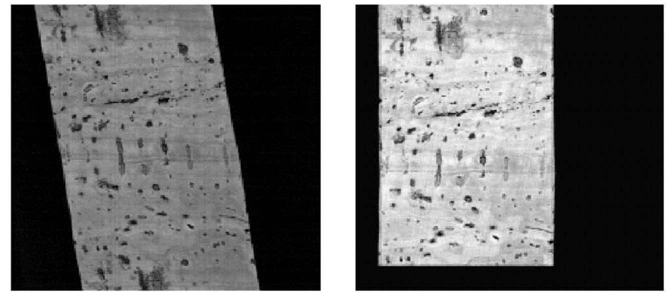

Fig. 3. Original cork stopper image (left) and its geometrical and photometric correction (right).

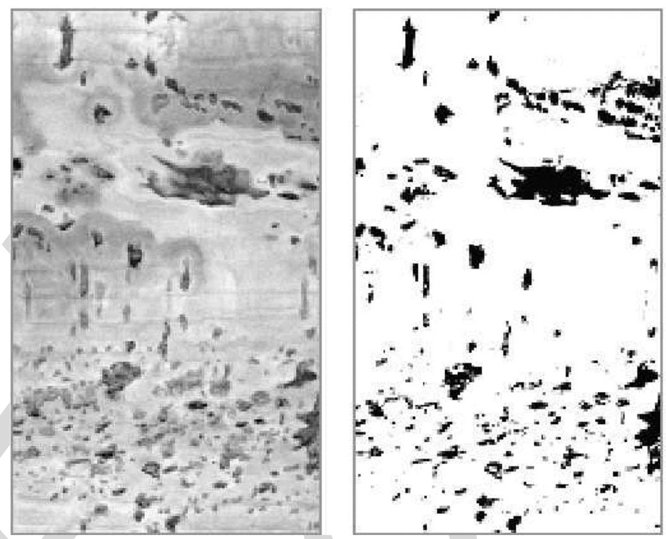

Fig. 4. Blob extraction of cork stoppers images.

Quality of cork stoppers directly depends on the presence and appearance of blobs on the stopper surface. Human beings take into account global as local stopper features when classifying cork stoppers. Global features can be expressed as overall gray number of blobs, overall distribution of blobs, overall gray-level appearance of stopper blobs, overall area of blobs, etc. Local stopper features usually refer to the first- and second-largest blobs of the cork stoppers and particularly their appearance: area, length, perimeter, convex perimeter, compactness, edge smoothness, etc. Following this strategy suggested by experts of cork classification, we defined 43 cork stopper features. The blob analysis (see Fig. 4) involves blob features as follows: Stopper area, number of blobs, average blob area, average blob elongation, average blob gray level, average compactness, average roughness, features of the first two blobs with the largest area (area, length, perimeter, convex perimeter, compactness, roughness, elongation, length, width, average blob gray level, position with respect to the center of the stopper), and the corresponding features of the two longest blobs.

[1]In the appendix, we show how the ICA model gives us *a priori* information that can be used in the estimation of the marginal densities of the independent components.

IEEE Proof

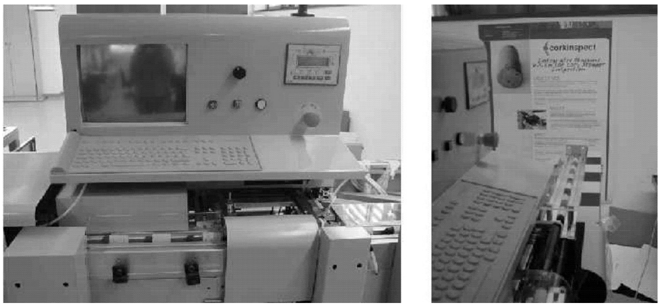

Fig. 5. External aspect of the artificial vision machine for automatic cork stopper inspection.

IV. Results of Classification of Cork Stoppers by CC-ICA

In order to compare our approach and to assess its performance for the problem of cork stopper classification, we used 5000 cork stoppers (1000 stopper per class) that were selected by experts in cork stopper quality of a company for cork stopper production. The images have been acquired by a specialized machine (see Fig. 5) and blob features have been extracted by the procedure described in Section III-A. We implemented and tested the following methods for the problem of cork stopper quality classification.

- The most simple classification technique, the nearest neighbor (NN) classifier [15], which classifies each cork representation - on the original space - to the class of the nearest representation of a stopper from the learning set.
- PCA of the data for dimensionality reduction, followed by NN classification [15].
- LDA, as described in [2], followed by NN classification.
- Nonparametric discriminant analysis (NDA), as described in [16].
- Maximum likelihood (ML) classification using a single unrestricted Gaussian distribution per class [15].
- Global ICA, as described in [13] without taking into account the class labels of the data. Prior to ICA, data was analyzed using PCA in order to keep only components with a significant associated eigenvalue (we kept the eigenvectors corresponding to 99.5% of data variation).
- CC-ICA, as described in Section III. Prior to ICA, data was analyzed using PCA in order to keep only components with a significant associated eigenvalue (we kept the eigenvectors corresponding to 99.5% of data variation).

Each experiment was validated using a variant of tenfold cross validations: We randomly divided the data into a test set (30% of the data) and a learning set (70% of the data) ten different times.

Fig. 6 shows the performance of these methods, measured from test data using cross-validation. It can be observed that for this problem, parametric methods are clearly superior to the nonparametric ones. CCA-ICA gets the maximum score.

In Fig. 7, the results corresponding to various implementations of the nearest neighbor approach are shown: a) 1-NN classification on original data; b) 1-NN on linearly normalized data; and c) 1-NN on data normalized on variance. The graphics

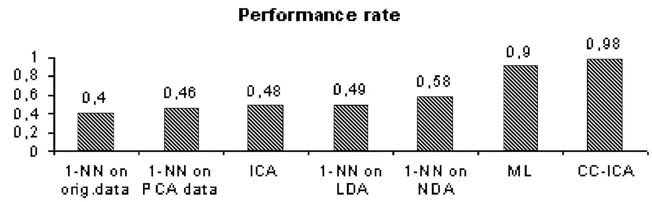

Fig. 6. Classifier performance.

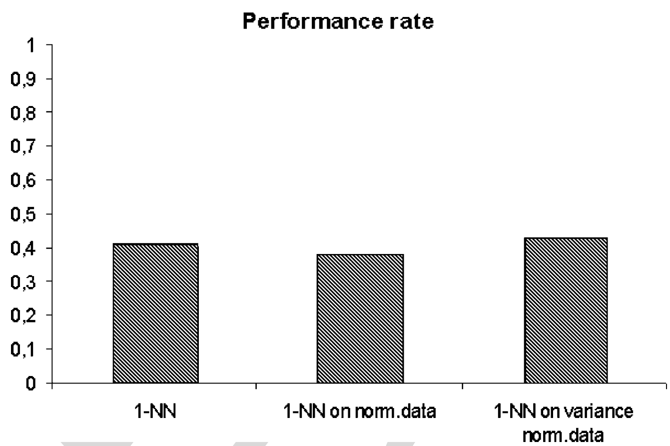

Fig. 7. Dependence of k-NN classifier on representation of data.

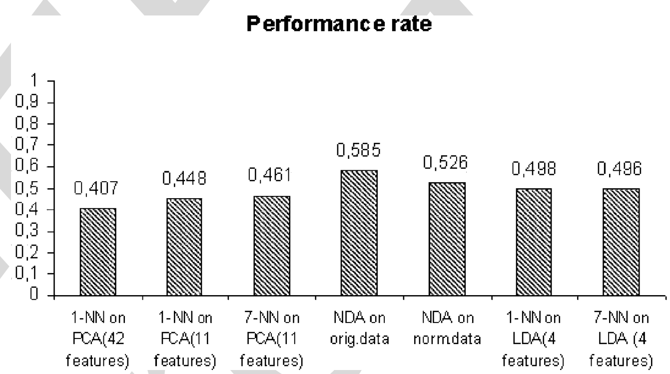

Fig. 8. Dependence on data representation.

shows that the performance of the method is not significantly affected by this aspect.

Fig. 8 shows the dependence of nonparametric methods on the data representations. Different tests have been run on original data, normalized data, and representation on reduced feature spaces by PCA and LDA. Different classifiers have been tested on original and normalized data. The figure shows the final results of applying NDA that gave a success rate of 52% and 58%. After reducing the feature space by PCA to different dimensions, the final results were quite similar, e.g., classifying a reduced space of 42 and 11 dimensions the success rate was 40% and 46%, respectively. When LDA was applied to obtain an optimal subspace for class discrimination (the space has been reduced to R^4) applying two different training sets, the results achieved 50% of success. Summarizing, the results show that the classification did not depend significantly on the data representation.

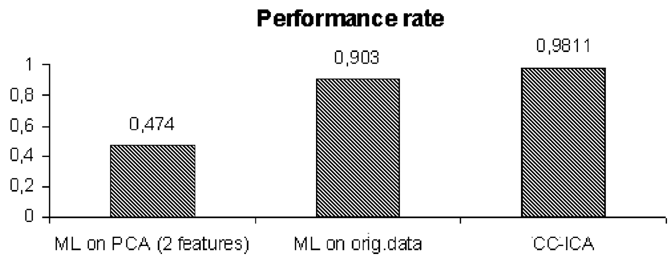

Fig. 9. Results of parametric classification.

The results differed meaningfully when parametric classifiers are applied (see Fig. 9). If the mean result of classification success was about 45% with nonparametric classifiers, parametric classifiers doubled the success rate achieving performance rate up to 98%. Furthermore, the results on the graphics show that keeping the full dimensionality of data is important for the classification performance. Maximum likelihood classification of original data got 90% of success, while Naive Bayes classification on CC-ICA got 98%. The results from CC-ICA can be explained by a better estimate of the probability density function of the different classes of cork stoppers thanks to complexity reduction when we transform the estimate of an N-dimensional density function to N estimates of 1-D density functions.

V. Conclusion

We tested the prototype by using cross-validation in more than 5000 cork stoppers applying a set of parametric and nonparametric classifiers. Special attention was paid to data representation. As a result, we obtained that data representation of feature space of cork stoppers which is important to the purposes of the correct classification. In particular, standard techniques, like PCA and LDA, for reducing feature space (in terms of blob characteristics of cork stoppers) hide the danger of loosing important information for the following process of classification. Our conclusion is that a sufficient representation of the probability density function of full-dimensional cork data is the best approach. When faced with this problem, the CC-ICA method proved its advantages as a robust estimate of this function. After testing an extensive set of classifier methods, we found that nonparametric group of classifiers showed a low-performance rate (mean of 45% of success). In contrast, Bayesian classification achieved high-performance rate (mean 94%) in different tests. A strong advantage of the proposed approach is that although the computational complexity of the selected method of CC-ICA during the learning phase can be high, the process of classification of new examples can be implemented in real time achieving the best performance rate.

Appendix

ICA Model and Marginal Density Estimation

Sparse coding is a coding of the data such that for any given input vector only a few components of the code will be significantly active (nonzero). A close relationship between sparse coding and ICA has been pointed out [17]–[19], [22]. In our particular problem, as in many others, a very high sparsity is observed in the independent components.

Sparsity of data is reflected on a distribution that presents a strong peak at zero and heavy tails. Kurtosis is a natural measure of how "peaky" a distribution can be. Kurtosis or the fourth-order cumulant is defined as $k(s) = E(s^4) - 3$ for a zero mean, unit variance variable (true for the independent components).

Depending on the sparsity (which we can measure using kurtosis as statistic), we can use different density models [20]. For moderately sparse variables, a good approach is the Laplace or double-exponential density

$$p(s) = \frac{1}{\sqrt{2}\alpha} \exp\left[-\left(\sqrt{2}|s|/\alpha\right)\right].$$

The main disadvantage of the Laplace density is that it only has one parameter α and so it is not suited for adapting the density to the data.

Another approach is to work with a mixture of Laplacians. A mixture of two zero-mean Laplace distributions proves easy to estimate and highly adaptive to strong variations in the level of sparsity. The general form for a mixture of L zero-mean Laplacians is

$$p(s) = \sum_{l=1}^{L} \frac{\beta_l}{\sqrt{2}\alpha_l} \exp\left[-\left(\sqrt{2}|s|/\alpha_l\right)\right]$$

where α_l is related with the variance of each component of the mixture and the β_l are normalizing constants. These parameters can be estimated, for instance, using the expectation-maximization algorithm [21].

A quite robust parametrization that provides an accurate approximation of very sparse data was introduced by Hyvärinen [13]

$$p(s) = \frac{1}{2} \frac{(\alpha+2)[\alpha(\alpha+1)/2]^{(\alpha/2+1)}}{\sqrt{\alpha(\alpha+1)/2} + |s|^{(\alpha+3)}}.$$

As $\alpha \leftarrow \infty$, this approaches the Laplace density. The parameters are estimated as follows:

$$\alpha = \frac{2 - k + \sqrt{k(k+4)}}{2k-1}$$

where

$$k = p(0)^2$$

and $p(0)$ is estimated with a suitable kernel.

Hyvärinen's density and the mixture of Laplacians were the most frequently used densities in the experiments.

References

[1] M. Bressan, D. Guillamet, and J. Vitrià, "Using an ICA representation of local color histograms for object recognition," *Pattern Recognit.*, vol. 36, no. 3, pp. 691–701, Mar. 2003.
[2] P. Belhumer, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[3] I. T. Jolliffe, *Principal Component Analysis*. Cambridge, MA: MIT Press, 1986.
[4] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. Intell. J.*, vol. 7, no. 1, pp. 273–324, 1995.

IEEE Proof

[5] P. Comon, "Independent component analysis— A new concept?," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[6] C. Jutten and J. Herrault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, pp. 1–10, 1991.

[7] A. Bell and T. Sejnowski, "An information-maximization approach for blind signal separation," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[8] J. F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Lett. Signal Process.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[9] T. W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and supergaussian sources," *Neural Comput.*, vol. 11, pp. 609–633, 1998.

[10] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind source separation," *Adv. Neural Infor. Process.*, vol. 8, pp. 757–763, 1996.

[11] M. Kendall and A. Stuart, *The Advanced Theory of Statistics.* London, U.K.: Charles Griffin, 1958.

[12] A. Hyvärinen, "New approximatins of differential entropy for independent component analysis and projection pursuit," *Adv. Neural Process. Syst.*, vol. 10, pp. 273–279, 1998.

[13] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, pp. 1483–1492, 1999.

[14] E. H. Simpson, "The interpretation of interaction in contingency tables," *J. Roy. Statis. Soc., Ser. B*, vol. 13, pp. 238–241, 1951.

[15] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[16] K. Fukunaga and J. M. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, Nov. 1983.

[17] D. J. Field, "What is the goal of sensory coding?," *Neural Comput.*, vol. 6, pp. 559–601, 1994.

[18] A. Hyvärinen, "Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation," *Neural Comput.*, vol. 11, pp. 1739–1768, 1999.

[19] R. Vigario, J. Särelä, V. Jousmäki, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, 2000.

[20] P. O. Hoyer, "Independent Component Analysis in Image Denoising" Ph.D. dissertation, Helsinki Univ. Tech., 1999.

[21] A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood for incomplete data via the em algorithm," *J. Roy. Statis. Soc.*, vol. 39, pp. 1–38, 1977.

[22] A. Bell and T. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Neural Comput.*, vol. 11, pp. 1739–1768, 1999.

Jordi Vitrià received the Ph.D. degree in mathematical morphology from the Universitat AutAutònoma de Barcelona (UAB), Barcelona, Spain, in 1990.

In 1991, he joined the Computer Science Department of the UAB, where he became an Associate Professor. He is the author of more than 60 scientific publications and 1 book. His research interests include machine learning, pattern recognition, and visual object recognition.

Marco Bressan received the B.A. degree in applied mathematics from the University of Buenos Aires, Buenos Aires, Argentina, the M.Sc. degree in computer vision from Computer Vision Centre, Barcelona, Spain, and the Ph.D. degree in computer science and artificial intelligence from the Universitat Autònoma de Barcelona (UAB), Barcelona.

He is currently with the UAB. He is a permanent researcher in XRCE's image processing research group (IPC). He has authored more than 15 refereed publications. His main research interests are statistical learning and classification; image and video semantic scene understanding; and object detection, recognition, and enhancement, particularly when dealing with uncontrolled environments.

Petia Radeva (M'01) received the Ph.D. degree from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 1998.

Currenty, she is an Associate Professor at the Computer Science Department, UAB. She has more than 90 international publications in international journals and proceedings. Her present research interests are development of physics-based approaches (in particular statistics methods and deformable models) for medical image analysis, industrial vision, and remote sensing.

IEEE Proof

QUERIES

Q1. Author: Note that vectors are set in bold italics and matrics are set in bold roman. Kindly mark for the same in the proof.

IEEE Proof