Feature Selection in AVHRR Ocean Satellite Images by Means of Filter Methods

Jose A. Piedra-Fernández, Manuel Cantón-Garbín, and James Z. Wang, *Senior Member, IEEE*

***Abstract*—Automatic retrieval and interpretation of satellite images is critical for managing the enormous volume of environmental remote sensing data available today. It is particularly useful in oceanography and climate studies for examination of the spatio-temporal evolution of mesoscalar ocean structures appearing in the satellite images taken by visible, infrared, and radar sensors. This is because they change so quickly and several images of the same place can be acquired at different times within the same day. This paper describes the use of filter measures and the Bayesian networks to reduce the number of irrelevant features necessary for ocean structure recognition in satellite images, thereby improving the overall interpretation system performance and reducing the computational time. We present our results for the National Oceanographic and Atmospheric Administration satellite Advanced Very High Resolution Radiometer (AVHRR) images. We have automatically detected and located mesoscale ocean phenomena of interest in our study area (North–East Atlantic and the Mediterranean), such as upwellings, eddies, and island wakes, using an automatic selection methodology which reduces the features used for description by about 80%. Finally, Bayesian network classifiers are used to assess classification quality. Knowledge about these structures is represented with numeric and nonnumeric features.**

***Index Terms*—Automatic image interpretation, feature selection, ocean image analysis, pattern classification, sea surface temperature.**

I. Introduction

CONSIDERING the increasing number and size of satellite image databases, automatic processing is becoming necessary for retrieving and coding useful environmental information from these databases. We are particularly interested in the analysis of mesoscale ocean structures in satellite images. New methodologies have been proposed for the automatic estimation of ocean surface currents using AVHRR and MODIS imagery in [1] and [2]. In these cases, the main goal is to obtain a coarse structure segmentation and achieve the maximum detail of the structure. The problem is knowledge extraction. One of our goals is to be able to select the fewest segmented ocean region features in satellite images which still retain most of the knowledge necessary to compare the relevance of selected feature data sets [3]–[5]. In general, support vector machines (SVMs) do not require a dimensionality reduction for remote sensing data classification. However, in [6], a study with two hyperspectral sensor data sets shows that the classification accuracy using an SVM with a small training data set depends mainly of the addition of new features. In fact, the accuracy of classification by an SVM depends on the dimensionality of the data. For this reason, it is important to include a feature selection.

The most common framework for feature selection is to define the criteria for measuring the goodness of a set of features [7] and then use a search algorithm to find the optimal or suboptimal feature set [8]. Other authors use an evolutionary algorithm for the classification of hyperspectral images that estimates the set of Pareto-optimal solutions [9]. In [10], a sparse conditional random field model is used to select relevant features. Focusing on spectral sensors with overlapping and noisy bands, a canonical correlation-based feature selection (CFS) algorithm was used in [11].

In this paper, we have used filter measures and Bayesian classifiers to select features for ocean structure recognition in satellite images [12], [13].

In the rest of this section, we describe the mesoscale ocean phenomena of interest for classification that appear in infrared satellite images. We will also introduce the symbolic and numeric features used for classification.

A. Ocean Phenomena

The AVHRR sensor has been a powerful tool in environmental, climate, and geophysical and geographical research tasks for more than three decades. This sensor, onboard the Tiros and National Oceanographic and Atmospheric Administration (NOAA) satellite series, covers five channels (second version: AVHRR-2) in the infrared and visible spectra. Infrared information, in particular, has been used in ocean structure identification [14]–[18].

This study was carried out in the Canary Island, Iberian Atlantic, and Mediterranean Seacoast regions. A detailed oceanographic description of this area can be found in [19]–[23]. In this area, there are several different significant mesoscale ocean structures: upwellings, cold eddies, warm eddies, and wakes (Fig. 1).

Manuscript received August 7, 2008; revised January 26, 2009 and March 1, 2010. This work was supported by the Spanish Ministry of Science and Technology through the Project TIN2008-06622-C03-03. The work of J. Z. Wang was supported by the National Science Foundation under Grants 0347148 and 0219272.

J. A. Piedra-Fernández was with the Languages and Computation Department, University of Almería, 04120 Almería, Spain. He is now with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: jpiedra@ual.es).

M. Cantón-Garbín is with the Languages and Computation Department, University of Almería, 04120 Almería, Spain (e-mail: mcanton@ual.es).

J. Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: jwang@ist.psu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2010.2050067

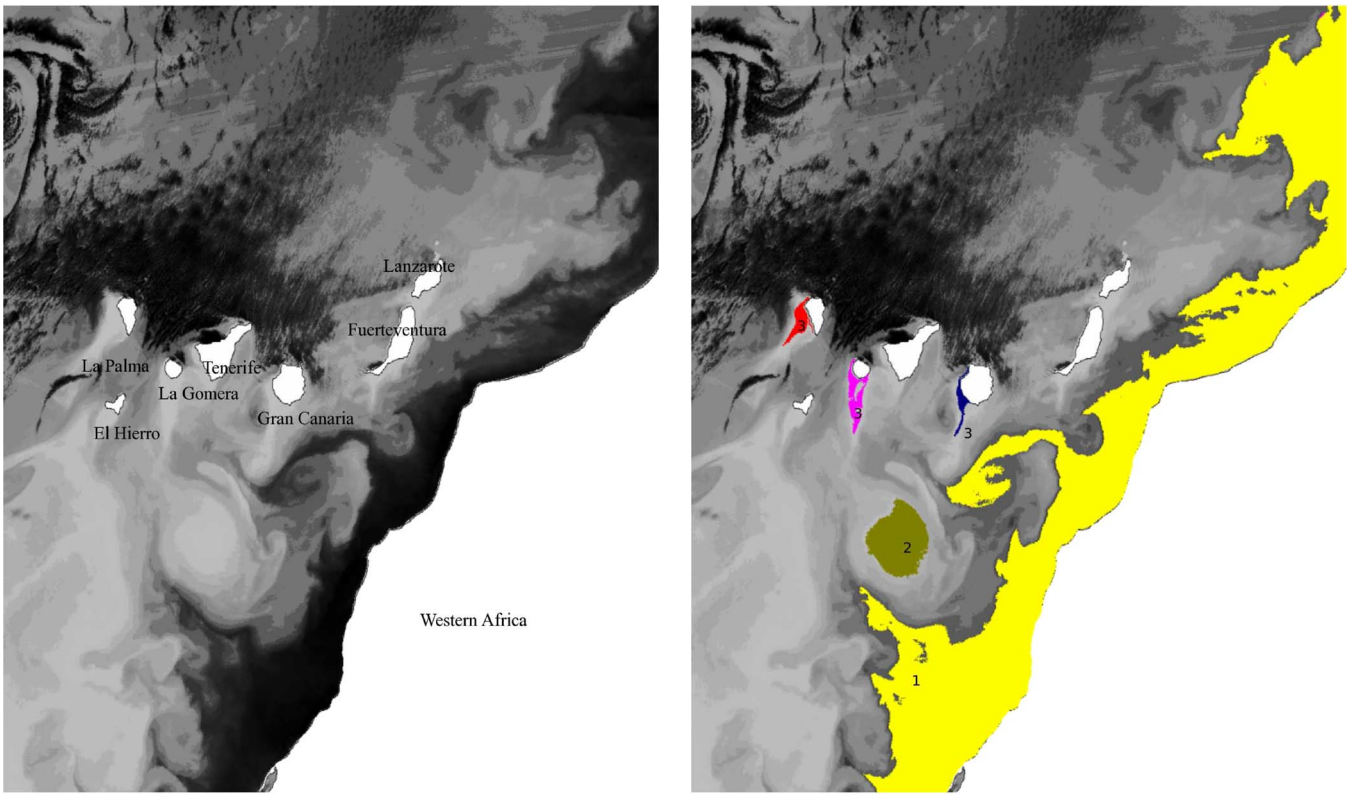


Fig. 1. AVHRR scene (08/10/1993). (a) Original in grayscale. (b) Ocean structure map. Yellow (1): Upwelling. Dark green (2): Warm eddy. Other colors (3): Wakes.

Upwelling is the term used by oceanographers to describe cool nutrient-rich water from the lower layers of the ocean which rises to the surface. Brought into the light, such water is very fertile and a rich feeding ground for fish. For an upwelling to occur, there must be a divergence of the surface currents, and this usually happens as a result of the wind field causing surface-water drift in the presence of topographic constraints. Alongshore winds blowing in the Northern Hemisphere with a coastline to the left of the direction of travel set up a geostrophic flow away from the coast and produce an upwelling. This upwelling is a regular occurrence off the North–West African coast (Fig. 1) and others, like the Peruvian coast, where wind conditions are suitable. Nonetheless, upwellings may well be intermittent, depending on both the weather [20] and, to a certain extent, ocean-wide thermohaline circulation. Because of their importance for commercial fisheries, much oceanographic research has been done to understand and predict upwellings. Remote sensing from satellites has been a powerful tool for such studies.

Eddies are highly morphologically and contextually variable structures difficult to define. Eddy water is different from the surrounding water in temperature and salinity. In addition, an eddy can travel long distances for long periods of time without mixing with the surrounding water [21], [22]. On the other hand, the movement and shape of cold and warm eddies in our region of study are controlled by the trade winds. The cool eddies are more intense under calm conditions, whereas the warm eddies are intensified by strong winds (Fig. 1). In cool eddies, the vertical movement is ascending; cold water, rich in nutrients, rises to the surface [23]. However, warm eddies accumulate and sink warm water, carrying organic matter downward toward the ocean depths.

Wakes are warm oceanic structures associated with islands [20], [21]. Wakes have been observed leeward of the Canary Islands, generated by the obstacle that the islands present to the predominant North–East trade winds in this region. This implies that the intensity of winds is lowered southwest of the islands and the sea surface is warmed in these zones (Fig. 1). These wakes are very thin and warmer than surrounding water.

These ocean structures are shown in Fig. 1(a) (an equalized AVHRR scene) and classified and labeled in Fig. 1(b).

B. Data Set

The original data set used in this study was a symbolic and numeric feature database computed from previously segmented mesoscale ocean regions in AVHRR images. AVHRR data used for calculating these features come from Channel 4 (far infrared). The symbolic database was built from the knowledge of human experts and the competitive high-level knowledge processor networks used in our previous works [14]. Ordinary symbolic logic handles only features which are present (true) or absent (false). The symbolic feature set used in this work can be divided into two categories: morphological features described in Table I and contextual features described in Table II (as explained in [14]).

TABLE I
MORPHOLOGICAL FEATURES

Feature Description
Size in pixels above or below threshold
Above-average regional temperature
Below-average regional temperature
Regional variability above 4
Regional variability above 6
Variability below 4
Presence of a kernel
Hat or V-shape (cold or warm eddies)
Rounded
Not on the boundary
Subregion centroids are aligned
Kernel colder than the rest of the region
Kernel warmer than the rest of the region
Oblique Gaussian shape (upwelling)
Oblique Gaussian shape
Not in any defined region

Table I shows the pixel value and measurements relating a pixel to its neighbors, for example, above/below average regional temperature. One of the more relevant features is the variability measurement, because it is used to detect clouds in AVHRR images [24].

Table II gives the geographic position of segmented regions and features related to the region position in the image [16].

Furthermore, a numeric feature database (Table III) was extracted from the same set of satellite images. The main features used were the invariant moments [25]–[28], which are basically statistical data carrying information about the frequency components of the image in the moment space (in this case, image intensities) and are an effective and compact image feature set that can be used to represent the ocean features in the feature space. Tables I and II show discrete features, and Table III shows continuous features.

C. Main Contributions of the Work

In this paper, we illustrate a novel and effective way of extracting relevant features in the ocean satellite image recognition using filter methods and Bayesian networks. Specifically, by successfully integrating different distances and techniques, our method achieves the following.

1) Remove the irrelevant features: This method gets a ranking of features and tries to find the optimum subset of features. The idea is to choose a ranking feature subset by different thresholds and to evaluate the accuracy by means of Bayesian networks.
2) Identify the relationships between features: Bayesian networks define the conditional dependence of features.
3) Reduce the computational cost.
   a) Filter methods are simplest and fastest to get a ranking of features.
   b) Bayesian classifiers used are easy to design and fast to evaluate.

   The method was evaluated with different classifiers. Experimental results show that our method is advantageous.

TABLE II
CONTEXTUAL FEATURES USING A GIS DATABASE

Feature Description
Land zone the region pertains to
SE Spain
SW Portugal
Melilla
NW Africa
Sahara
Fuerteventura
Isla de Lobos
Lanzarote
Grand Canary Island
Tenerife
Gomera
La Palma
Hierro
N Cape White
S Cape White
Cantabrian Coast and W of France
Regions position to land
North
South
East
West
Type of land closest to the region
Island
Continental platform
Regions position in the ocean
Coast
Transition Zone
High sea
Hemisphere the region is in
North
South
Sea or ocean the region is in
Mediterranean
Atlantic
Other features
No clouds surrounding the region
The region is closer to the islands or to filamentous structures

TABLE III
NUMERIC FEATURES

Features	**Bounding ellipse**	**Bounding box**	**Grayscale level (0-255)**	**Invariant Moments**
Area	Centroid	First point	Min	Hu's [25] (7 invariants)
Perimeter	Major Axis	Height	Max	Maitra's [26] (6 invariants)
Density	Minor Axis	Width	Mean	Zernike's [27] (12 invariants)
Volume	Orientation	Area	Standard deviation	Cantón's [28] (16 invariants until 5th order)
Volume^2	Eccentricity	Extended	Barycenter	
Equivalent diameter	Irradiance			

D. Outline of This Paper

In Section II, we provide a description of the methodology used for feature selection. Experimental results are presented

and discussed in Section III using several feature sets. We conclude and suggest further improvements in Section IV.

II. Methodology for Feature Selection

Feature selection is determined, in general, by three factors.

1) *Available data set*: In our study, the features are mixed (discrete and continuous).
2) *Algorithms used*: This can be defined in terms of evaluation measures (e.g., distance, information, dependence, and consistency) and search strategies (e.g., exhaustive, complete, heuristic, and random). An optimal feature subset is always relative to a certain evaluation measure, i.e., an optimal subset, chosen using one evaluation measure, may not be the same as that using another. In this study, the distance and dependence measures are used. Moreover, algorithms with a heuristic search strategy are used, because they are very simple to implement and produce very fast results.
3) *Performance required*: In this case, accuracy and optimality were considered.

Methods of feature selection are divided into three categories. Wrapper methods require evaluation of each subset of features by a classifier. These methods are black boxes with some parameters. The problem is NP-complete since the number of all subsets grows exponentially with the number of features. The next category is embedded methods that perform feature selection in the training process. The third category of methods is filter methods based on evaluation of individual features independently from the classifier used.

We believe that the third option is suitable for this task, mainly because these methods are independents, faster than other options while getting similar results.

Filter methods are useful for a high-cardinality data set. We measure the similarity between any feature and class. The similarity is represented by the ranking results. We have chosen six distance measures, which are the classic measures used in several papers. In this case, it does not get the real relationships between any features. For this purpose, we have included correlation measure that allows the comparison of any two features and classes at the same time.

The goal of this work was to reduce the search area of the relevant feature space. This methodology consists of several steps (Fig. 2). The step of feature selection orders the features by information relevance. The whole cycle allows us to find the best feature subset that is close to the number of features and accuracy rate specified.

This methodology tries to combine filter methods and Bayesian classifiers. One advantage of these classifiers is that they allow several kinds of causal reasoning: inductive, deductive–predictive, and intercausal. The different steps and their relationships are described in the following sections.

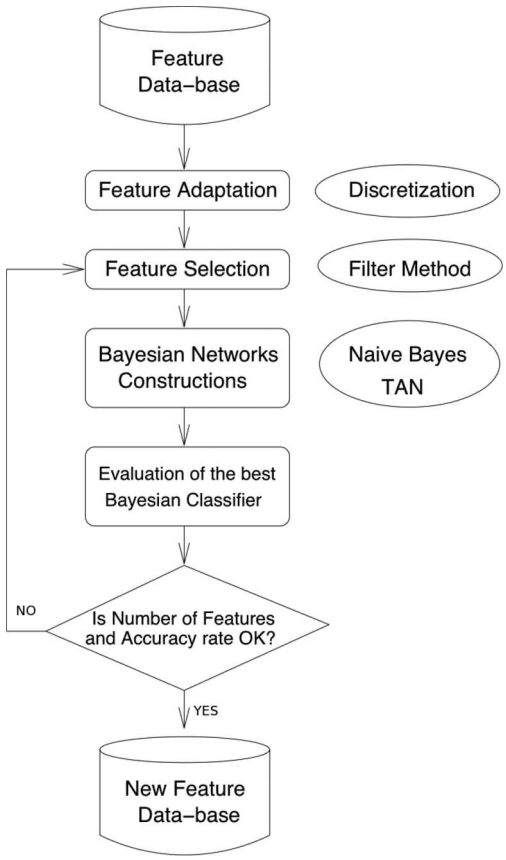


Fig. 2. Methodology of feature selection.

A. Feature Adaptation

Discretization should significantly reduce the number of possible values of continuous features, since a large number of possible feature values contribute to a slow and ineffective process of inductive machine learning. Thus, a supervised discretization algorithm should seek the minimum number of discrete intervals while not weakening the interdependence between the feature values and the class label.

Discretization algorithms can be divided into two categories [29].

1) Unsupervised (class-blind) algorithms discretize attributes without taking class labels into account. The representative algorithm is the equal-frequency (EF) discretization [30].
2) Supervised algorithms discretize attributes by taking into account class-attribute interdependence. The representative algorithm is the K-Means method [31]. This algorithm finds K groups from a data set, where the goal is to minimize the distance between the data in each group.

The main problem in discretization is to choose the optimal number of intervals (EF algorithm) or number of groups (K-Means algorithm). We performed an iterative evaluation of different intervals and groups in which we used two parameters, the number of relevant features and the accuracy rate. The number of features was calculated by means of CFS (Section II-B) because it is actually a filter method that provides good results. On the other hand, the accuracy rate was estimated by the Naive Bayes (NB) network (Section II-C) because it is

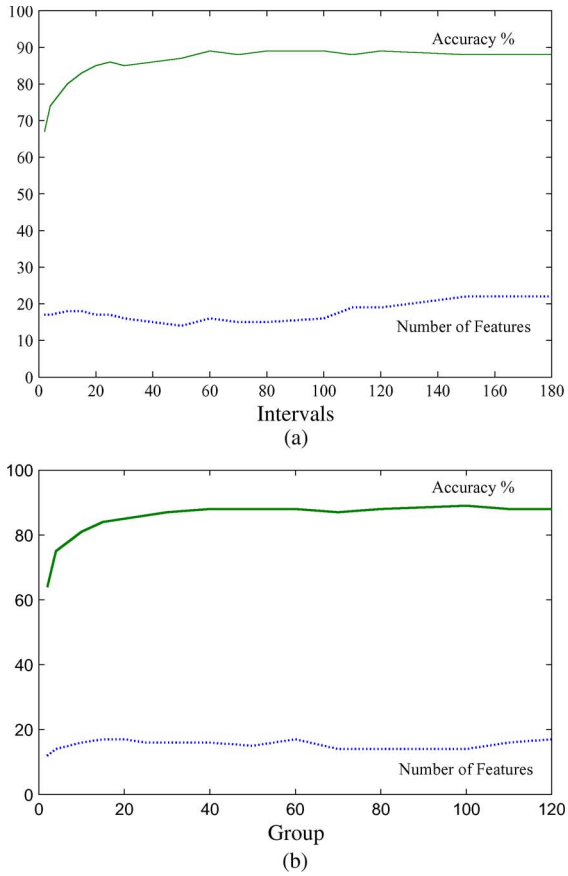


Fig. 3. Discretization evaluation process. (a) EF. (b) K-Means.

the simplest classifier and has produced good results in other studies [32].

Fig. 3 shows the relationship between the accuracy rate and the number of features used to determine the optimal number of intervals (using the EF method) and the optimal number of groups (by the K-Means method) in the discretization process. In Fig. 3(a), there are 100 optimal intervals and 16 features selected with 89% accuracy. Fig. 3(b) shows that 100 groups would be reasonable for discretization by the K-Means method, because only 14 features are necessary for an 89% accuracy rate.

Moreover, we use the expectation maximization (EM) algorithm [33] to automatically find the optimal number of groups. EM is a statistical model that makes use of the finite Gaussian mixture model. It is similar to the K-Means in that a set of parameters is computed until a desired convergence is achieved. We therefore use the optimal number of groups found by the EM algorithm for discretizing the continuous features and the K-Means algorithm to check its efficiency.

B. Feature Selection

Dimension reduction is often used in clustering, classification, and many other machine-learning and data-mining applications. It usually retains the more important dimensions (features), removes noisy dimensions (irrelevant features), and reduces computational cost.

Filtering methods, which are functions returning a relevance index estimating the efficiency of a feature subset for classification, are based on performance evaluation metrics calculated directly from the data. These methods are computationally less costly than wrapped methods [34].

The relevance index assigned to each feature should be positively correlated with the accuracy rate found by the classifier. One of the problems is that this is not always the case, and it is rather difficult to argue the theory that filtering methods are better than a classifier. In this case, Bayesian networks have been chosen as the experimental classifier because of the good results found with them and their simplicity.

In our study, the filtering methods used for experimental testing can be divided into two groups, i.e., those based on distance measures and those based on correlation measures or feature dependence.

Distance Measures are used for a two-class problem [35]. A feature X is preferred to another feature Y if X induces a greater difference than Y between the two-class conditional probabilities. If the difference is zero, then X and Y are indistinguishable. These measures have been adapted for multiple classes.

There are many ways to measure feature and class dependence based on evaluating differences between the probability distributions. Six univariate metrics were chosen in this work. Detailed discussions on six univariate metrics are in [35]. To help the readers understand our work which applies these techniques in image analysis, we include here a brief summary.

The metric Euclidean distance (ED) is based on the classic two-class evaluation formula, modified for multiclass evaluation (1). To work with Bayesian nets by means of probability distribution, the expression used is

$$ED(XC) = \left[\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{k<j} P(c_i)P(c_j)|P(x_i|c_k) - P(x_i|c_j)|^2\right]^{\frac{1}{2}} \quad (1)$$

where m is the maximum number of states X (i.e., discretize variable with one specific number of intervals), n is the maximum number of states C (or classes), $P(X|C)$ is the conditional probability distribution function of X given C, and $P(C)$ is the marginal probability distribution function of C.

The Bhattacharya distance metric (BD) measures the distance between two probability distributions (2) (in this case, it measures the dependence between a feature and a class)

$$BD(XC) = \sum_{j=1}^{n} -\log\left[P(c_i)\sum_{i=1}^{m}\sqrt{P(x_i|c_j)P(x_i)}\right] \quad (2)$$

where $P(X)$ and $P(C)$ are the marginal probability distribution functions of X and C, respectively.

The Jeffries–Matusita distance metric (MD) is similar to the BD, as it measures the distance between two probability distributions (3). The difference is that it measures the dependence

for each feature and the class, considering the mean distance of the conditional distributions

$$MD(XC) = \sum_{i=1}^{M} \sum_{k=1}^{k<i} P(c_i)P(c_k) \left[\sum_{j=1}^{n} \sqrt{P(x_j|c_i)P(x_j|c_k)} \right] \tag{3}$$

where $P(C|X)$ is the conditional probability distribution function of C given X and $P(C)$ is the marginal probability distribution function of C.

The Mutual information (MI) is the measure of the amount of information that one feature contains about another [36]. Formally, the MI of two discrete random variables X and Y can be defined as

$$MI(XC) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(x_i, c_j) \log \frac{P(x_i, c_j)}{P(x_i)P(c_j)} \tag{4}$$

where $P(X, C)$ is the joint probability distribution function of X and C, and $P(X)$ and $P(C)$ are the marginal probability distribution functions of X and C, respectively.

The Kullback–Leibler divergence (KL) measures the distance between two probability distributions, which makes it possible to measure the dependence of each feature and the class [36]. The idea is to select the feature with the highest dependence with respect to the class variable.

P typically represents data or observations on the probability distributions. Q represents a theory, a model, a description, or an approximation of P. For probability distributions P and Q of a discrete random variable, the KL of Q from P is defined by

$$D\left(P(X), Q(X)\right) = \sum_{x_i} P(x_i) \log \frac{P(x_i)}{Q(x_i)}. \tag{5}$$

The probability distributions to be used had to be decided. The prior marginal probability distribution was chosen because it produces good results. The expressions are

$$KL_{ij}(XC) = D(P(X|c_i), P(X)) + D(P(X|c_j), P(X)) \tag{6}$$

$$KL\,D(XC) = \sum_{i=1}^{m} \sum_{j=1}^{j<i} P(c_i)P(c_j)KL_{ij}(XC). \tag{7}$$

The Shannon entropy metric (SE) is the most popular for measuring the relevance of a feature. For the multiclass problem, this metric is [36]

$$SE_{ij}(X) = -\sum_{i=1}^{m} P(x_i|c_i) \log_2 P(x_i|c_j) + P(x_i|c_j) \log_2 P(x_i|c_i) \tag{8}$$

$$SE\,D(XC) = \sum_{i=1}^{m} \sum_{j=1}^{j<i} P(c_i)P(c_j)SE_{ij}(XC). \tag{9}$$

Dependence or correlation measures predict the relationship between features. Correlation measures can be used to find the correlation between a feature and a class [37]. If the correlation of feature X with class C is higher than the correlation of feature Y with C, then feature X is preferred to Y because X recognizes C (classes) better than Y. The CFS method is described in [37]. The CFS measures the relevance of a feature when classifying an object class in two ways individually (only taking the feature and the class into account) and by correlating with other features (taking into account a group of features and the class). The main difference between the distance and dependence measures is that the distance measures only consider the relationship between one feature and the class; however, the dependence measures consider the relationship among two or more features and the class. This is very important since it allows one to consider the relationships among features that, in the distance measures, are not considered. The dependence method uses a heuristic search strategy based on the greedy hill-climbing algorithm, which constructs a search tree with all the features and generates all the possible combinations of one feature with the rest. In each step, the algorithm selects the next features having a higher correlation with the feature subset and the class to be included. The dependence expression is

$$H = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \tag{10}$$

where k is the number of features in the subset, $\overline{r_{ci}}$ is the average class correlation, and $\overline{r_{ii}}$ is the average intercorrelation between features in the subset. The numerator represents the classification efficiency, and the denominator measures feature redundancy.

The conditional entropy measures the correlation between features and classes. If X and Y are discrete random variables with ranges R_x and R_y, the next expressions show, respectively, the entropies of Y before and after X is found in (11) and (12).

$$H(Y) = -\sum_{y \in R_y} P(y) \log\left(P(y)\right) \tag{11}$$

$$H(Y|X) = -\sum_{x \in R_x} P(x) \sum_{y \in R_y} P(y|x) \log\left(P(y|x)\right) \tag{12}$$

The measurement of the correlation between Y and X is defined by

$$C(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)}. \tag{13}$$

C. Bayesian Network Constructions

The last step in the proposed methodology is the use of a classifier to test the goodness of each feature previously selected. Bayesian networks [38] have been successfully used as models for representing uncertainty in knowledge databases. The uncertainty is represented in terms of a probability distribution with induced independence relationships encoded by the network structure.

A Bayesian network for a set of variables $X = X_1, \ldots, X_n$ formally consists of a directed acyclic graph where each node is labeled with a variable in X, and a set of probability conditional distributions for each variable X_i given its parents in the graph, which is denoted as $P(X_i|X_{pa_i})$.

A Bayesian network can be used as a classifier. One of its variables represents the class, and the others are the features that describe the object to be classified. Classification is done by instantiating the value of the feature variables and probability propagation [38] over the class variable, which consists of computing the posterior probability of each class given the features observed. Afterward, the class assigned is the one with the highest posterior probability.

We tested two simple methods for training the Bayesian network with a set of numeric data.

1) NB is based on the assumption that all the features are conditionally independent when the class is known. This assumption implies that the network structure is rather simple, since only the arcs in the network link the class variable with each feature, and there are no arcs among feature variables. The advantage of this naive approach is the small number of parameters to be learned from the data, thereby improving the estimation accuracy [39].
2) Tree Augmented Naive Bayes Classifier (TAN): TAN models are a restricted family of Bayesian networks in which the class variable has no parents and the parents of each feature are the class variable and another feature at most [40].

D. Evaluation of the Best Bayesian Classifier

The classification experiments were performed with tenfold cross-validation. This means that the whole data set was partitioned into ten subsets, of which nine were used as training sets and the remaining one was first set aside as a test set, and then used to evaluate second stage results. This was done for the ten subsets.

Several different Bayesian classifiers were constructed for this methodology. Instead of selecting them based only on accuracy, it may often be more effective to select for simplicity. We therefore considered how much they reduced the size (number of relevant features) and feature dependence relationships (in NB, all of the features are conditionally independent of the class, and in TAN, dependence relationships may be established between two features and the class).

III. Experimental Results

To test the methodology performance in our system, we have processed a set of 30 ocean images, where one of the major problems for the automatic interpretation is the high morphological variability in the shapes of the ocean structures. Fig. 4 shows the upwelling structure with different levels of segmentation.

In many cases, the problem with the segmentation is to decide which is the best region, because there is not a well-defined boundary for the ocean structures, even for the oceanographers. For this reason, we have included in the data set all the regions segmented that agree with the size, shape, and position conditions imposed for the expert system that segment the whole image [41] [Fig. 4(b)]. This introduces more information about the shapes improving the classification process.

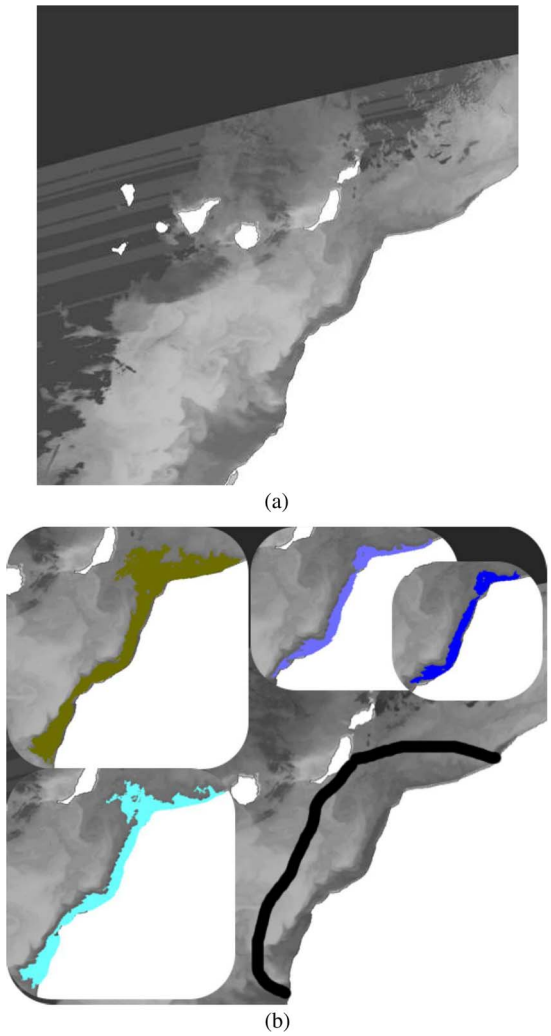


Fig. 4. AVHRR image (02/17/1990). (a) Channel 4. (b) Different levels of upwelling segmentations (black line).

To mask the clouds, we have used a cloud masking developed in previous works in our group. This masking process was built using CH2 and CH4 of AVHRR and two coupled neural nets [24]. We have used this cloudy masking in data set of Tables I and II. However, we have not used cloudy masking in the data set of Table III, but we have included a new class set based on mixed structures. The mixed structures include clouds in their shapes. Some examples are shown in Figs. 5 and 6.

We used about 1000 cases of real ocean structures (472 upwellings, 119 cloudy upwelling, 180 wakes, 10 anticyclonic eddies, 40 cyclonic eddies, and 180 misclassified regions) and 15 classes (3 upwelling classes, 3 cloudy upwelling classes, 2 eddy classes, 6 wake classes, and 1 class of misclassified regions) [Fig. 7(a)]. The upwelling region has been divided into three classes (Fig. 7(a); id class 2: upwellings between Cape Jubi and Cape Bojador, id class 3: upwellings

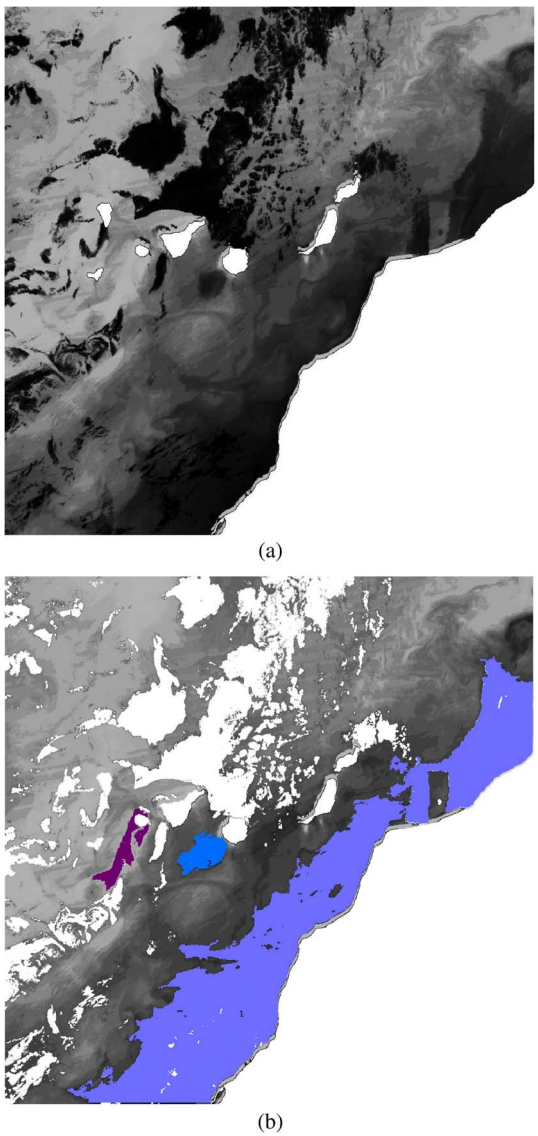


Fig. 5. AVHRR image (07/21/1990) (a) Original. (b) Cloud and land mask (white color), cloudy upwelling (light blue color-1), cyclonic eddy (blue color-2), and wake (pink color-3).

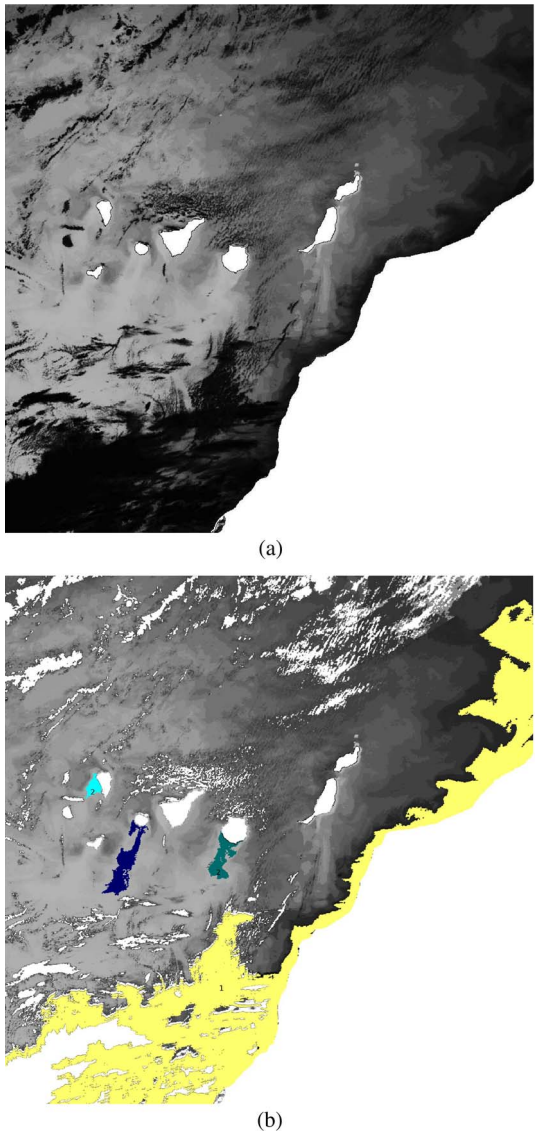


Fig. 6. AVHRR image (05/26/1990) (a) Original. (b) Cloud and land mask (white color), cloudy upwelling (yellow color-1), and wakes (blue and green colors-2).

south of Cape Bojador, and id class 4: upwellings in both coast regions). This solution introduces more knowledge about this big structure within noise or clouds. Classification results of an AVHRR scene are shown in Fig. 7(b).

Bayesian networks were constructed using the Elvira tool [42][1] running in an AMD Athlon (tm) 64 Processor 3200 (2.20 GHz) with 1.5-GB RAM.

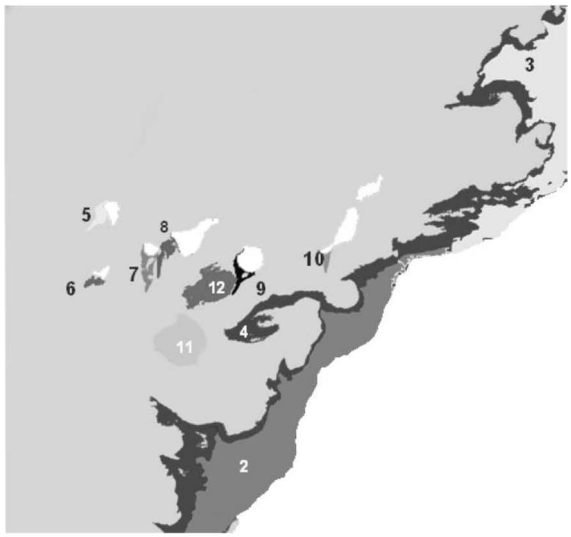
The results of Bayesian classifier evaluation of symbolic features are shown in Table IV, where the average number of relevant features used and the classification accuracy of each filter can be seen. The best accuracy rate was found using SE, although it retains a large number of features. Other options are BD and CFS, which yield approximately the same accuracy rate, but with a smaller number of features than the SE.

The numeric feature accuracy rate is shown in Table V (EF discretization) and in Tables VI and VII (K-Means discretization). Nine optimal groups were found by EM (Table VI). The best discretization was produced by K-Means with $k = 100$. CFS results were the best, because it reduced the size from 80 features to 14 and 16 features with an accuracy rate of around 90%.

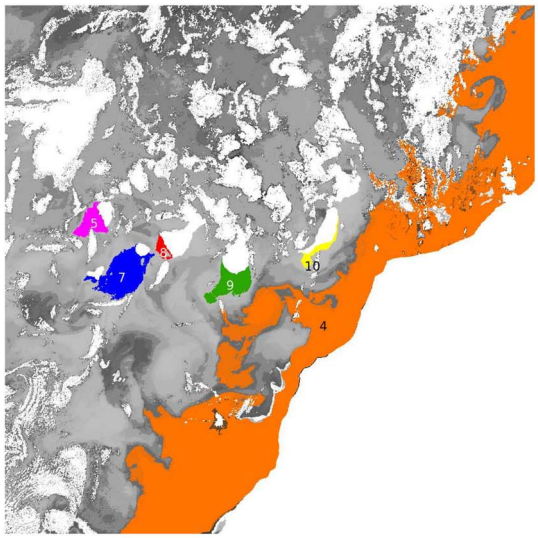
On the other hand, the accuracy rate is better in NB than in TAN. This is because NB assumes fewer errors in feature dependence relationships than TAN. Examples of the networks found in the best experiments in Table VII (Filter: CFS) can be seen in Fig. 8.

After selecting the most relevant features, we used this feature selection methodology with a group of several different classifiers: MLP (multilayer perceptron network), One R (one-rule classification), C4.5 (classification tree), NNGE (nearest neighbor with generalization), NB, TAN, hybrid system (radial basis function network and fuzzy system TSK-Sugeno),

[1] Available at http://leo.ugr.es/~elvira



(a)



(b)

Fig. 7. (a) Ocean structure identifiers (0 no figure; 1 water; 2–4 upwelling; 5–10 wake; 11–12 eddy). (b) AVHRR image (08/02/1993): Classification results are 5, 7, 8, 9, and 10—wakes and 4—cloudy upwelling (cloud and land mask—white color).

TABLE IV
EVALUATION WITH SYMBOLIC FEATURES

Filter	Number of Features	Accuracy % NB	Accuracy % TAN
ED	19	78.84	78.84
BD	13	77.60	77.60
MD	19	78.83	78.77
MI	19	78.84	78.85
KL	19	78.82	78.87
SE	48	79.00	79.10

neuro-fuzzy system (NEFCLASS and NEFPROX), and adaptive neuro-based fuzzy inference system (ANFIS) [43], [44] to find out the efficiency of other classifiers compared with Bayesian networks in evaluating the features extracted by this methodology. These classifiers were used with numeric features only.

TABLE V
EVALUATION WITH NUMERIC FEATURES BY EF (100 INTERVALS)

Filter	Number of Features	Accuracy % NB	Accuracy % TAN
ED	23	89.18	85.88
MD	64	81.88	82.58
MI	64	81.98	82.38
KL	19	88.48	83.18
SE	1	30.63	30.73
CFS	16	89.58	84.58

TABLE VI
EVALUATION WITH NUMERIC FEATURES BY K-MEANS ($K = 9$ GROUPS BY EM)

Filter	Number of Features	Accuracy % NB	Accuracy % TAN
ED	47	81.08	82.08
MD	27	82.28	83.78
MI	36	80.78	82.48
KL	16	81.28	85.68
SE	1	30.67	32.45
CFS	15	82.58	87.08

TABLE VII
EVALUATION WITH NUMERIC FEATURES BY K-MEANS ($K = 100$ GROUPS)

Filter	Number of Features	Accuracy % NB	Accuracy % TAN
ED	24	88.48	85.78
MD	61	85.28	82.48
MI	60	84.78	84.08
KL	19	87.68	84.38
SE	1	30.03	31.33
CFS	14	89.18	87.08

Table VIII shows only the best results, although all discretization algorithms studied (with 100 intervals for EF and 9 and 100 groups for K-Means) were analyzed. The % accuracy column (M) is the accuracy rate with the methodology, while % accuracy (WM) is when it was not used (for 80 features). In almost all cases, the proposed methodology improves the accuracy rate and reduces the number of features necessary to get a good ocean structure classification.

IV. CONCLUSION AND FUTURE WORK

We have explored the use of filter methods as a feature selection mechanism in an automatic ocean AVHRR satellite image recognition system. The use of this methodology has been beneficial for reducing the number of relevant features and in finding the knowledge structure, in terms of the conditional independence relationships of the features. The computational cost has been reduced, because filter methods are the simplest and fastest solution for feature selection. Moreover, the Bayesian classifiers used are very easy to build and train and have produced good results in many real application fields. The methodology was evaluated with different classifiers and was found to be advantageous.

In the future, we plan to improve the systems accuracy rate by including more features. Furthermore, we expect to use models like mixtures of truncated exponentials [45] to avoid the discretization of continuous features during Bayesian network training.

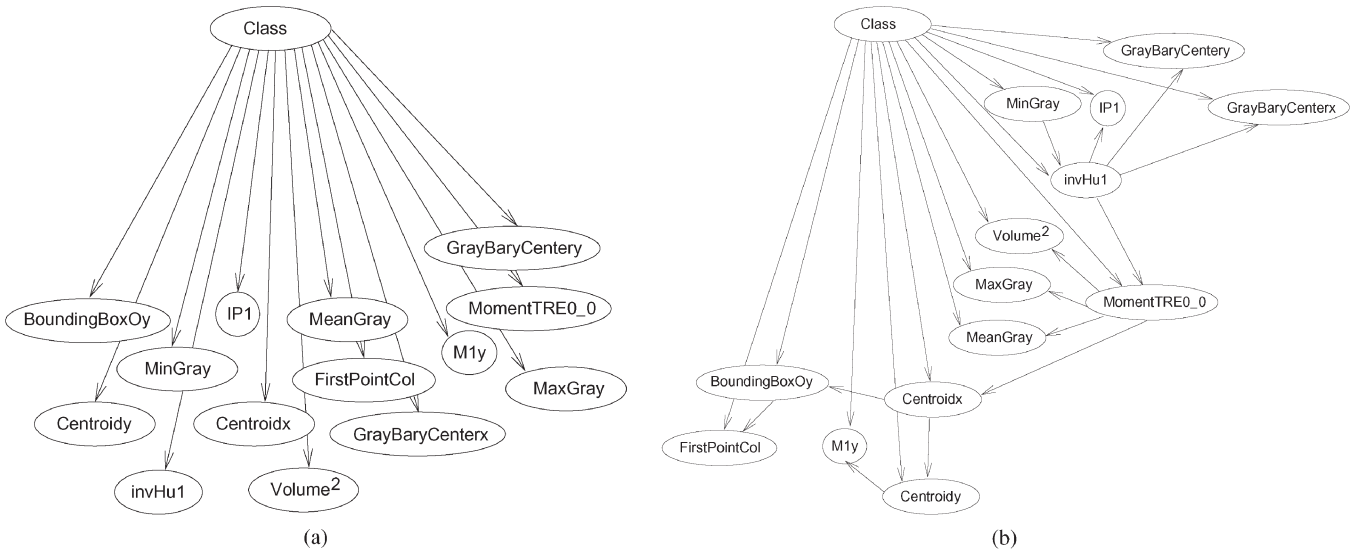


Fig. 8. Examples of Bayesian networks in Table VII—CFS. (a) NB CFS, numeric features. (b) TAN CFS, numeric features.

TABLE VIII
EVALUATION WITH DIFFERENT CLASSIFIER
(CFS 14 NUMERIC FEATURES)

Classifier	Number of Features	Accuracy % M	Accuracy % WM
MLP	14	90.49	90.39
OneR	16	61.96	64.36
C4.5	16	90.09	87.98
NNGE	14	92.49	90.29
Naïve Bayes	16	89.58	85.58
TAN	15	87.08	77.27
RBF-Sugeno	14	70.80	69.13
NEFCLASS	15	77.87	99.89
ANFIS	16	81.83	81.83
NEFPROX	16	89.62	89.62

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments.

REFERENCES

[1] J. Marcello, F. Marques, and F. Eugenio, "Automatic tool for the precise detection of upwelling and filaments in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1605–1616, Jul. 2005.

[2] J. Marcello, F. Eugenio, and F. Marques, "Methodology for the estimation of ocean surface currents using region matching and differential algorithms," in *Proc. IEEE IGARSS*, 2007, pp. 882–885.

[3] I. Inza, P. Larraqaga, R. Etxeberria, and B. Sierra, "Feature subset selection by Bayesian networks based optimization," *Artif. Intell.*, vol. 123, no. 1/2, pp. 157–184, Oct. 2000.

[4] R. S. Lynch and P. K. Willet, "Bayesian classification and the reduction of irrelevant features from training data," in *Proc. 37th IEEE Conf. Decis. Control*, Tampa, FL, 1998, pp. 1591–1592.

[5] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[6] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, May 2010.

[7] Y. L. Murphey and H. Guo, "Automatic feature selection—An hybrid statistical approach," in *Proc. IEEE Int. Conf. Pattern Recog.*, 2000, pp. 382–385.

[8] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Proc. 10th Conf. Uncertainty Artif. Intell.*, Seattle, WA, 1994, pp. 399–406.

[9] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 9, pp. 3180–3191, Sep. 2009.

[10] P. Zhong and R. Wang, "Learning sparse CRFs for feature selection and classification of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4186–4197, Dec. 2008.

[11] B. Paskaleva, M. M. Hayat, Z. Wang, J. S. Tyo, and S. Krishna, "Canonical correlation feature selection for sensors with overlapping bands: Theory and application," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 3346–3358, Oct. 2008.

[12] L. Bruzzone, "Classification of remote sensing images by using the Bayes rule for minimum cost," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 1998, pp. 1778–1780.

[13] Y. Yamagata and H. Oguma, "Bayesian feature selection for classifying multi-temporal SAR and TM data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 1997, pp. 978–980.

[14] J. A. Torres, F. Guindos, M. Lopez, and M. Canton, "Competitive neural-net-based system for the automatic detection of oceanic mesoscalar structures on AVHRR scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 4, pp. 845–852, Apr. 2003.

[15] J. A. Piedra-Fernandez, M. Canton, and F. Guindos, "Application of fuzzy lattice neurocomputing (FLN) in ocean satellite images for pattern recognition," in *Studies in Computational Intelligence (SCI)*. Berlin, Germany: Springer-Verlag, 2007, pp. 215–232.

[16] K. Simhadri, S. Iyengar, R. J. Holyer, M. Lybanon, and J. M. Zachary, "Wavelet-based feature extraction from oceanographic images," *IEEE Trans. Geosci. Remote Sens.*, vol. 36, no. 3, pp. 76–78, May 1998.

[17] M. Lybanon, "Maltese front variability from satellite observations based on automated detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 5, pp. 1159–1165, Sep. 1996.

[18] H. Thonet, B. Lemonnier, and R. Delmas, "Automatic segmentation of oceanic eddies on AVHRR thermal infrared sea surface images," in *Proc. Conf. Challenges Changing Global Environ. OCEANS*, 1995, vol. 2, pp. 1122–1127.

[19] L. Garcia-Weil, M. Pacheco, G. Rodriguez, and T. McClimans, "Topographic effects on the mesoscale ocean circulation near Canarian archipelago examined by means of satellite images and laboratory simulations," in *Proc. IGARSS*, 2000, pp. 1830–1832.

[20] E. D. Barton, J. Armstegui, P. Tett, M. Canton, J. Garcia-Braun, S. Hernandez-Leon, L. Nykjaer, C. Almeida, J. Almunia, S. Ballesteros, G. Basterretxea, J. Escanez, L. Garcia-Weil, A. Hernandez-Guerra, F. Lopez-Laatzen, R. Molina, M. F. Montero, E. Navarro-Pirez, J. M. Rodriguez-Pirez, K. V. Lenning, H. Vilez, and K. Wild, "The transition zone of the Canary current upwelling region," *Prog. Oceanogr.*, vol. 41, no. 4, pp. 455–504, Oct. 1998.

[21] J. Aristegui, P. Sangra, S. Hernandez-Leon, M. Canton, A. Hernandez-Guerra, and J. L. Kerling, "Island-induced eddies in the Canary islands," *Deep-Sea Res.*, vol. 41, no. 10, pp. 1509–1525, Oct. 1994.
[22] P. Sangra, J. L. Pelegri, A. Hernandez Guerra, I. Arregui, J. M. Martin, A. Marrero Diaz, A. Martinez, A. W. Ratsimandresey, and A. Rodriguez Santana, "Life history of an anticyclonic eddy," *J. Geophys. Res.*, vol. 110, no. 19, p. C03 021, Mar. 2005.
[23] A. Tejera, L. Garcia Weil, K. Heywood, and M. Canton, "Observation of oceanic mesoscale features and variability in the Canary islands area from ERS-1 altimeter, satellite infrared imagery and hydrographic measurements," *Int. J. Remote Sens.*, vol. 23, no. 22, pp. 4897–4916, Nov. 2002.
[24] J. A. Torres, F. Guindos, M. Peralta, and M. Canton, "An automatic cloud-masking system using backpro neural nets for AVHRR scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 4, pp. 826–831, Apr. 2003.
[25] M. K. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 2, pp. 179–187, Feb. 1962.
[26] S. Maitra, "Moment invariants," *Proc. IEEE*, vol. 67, no. 4, pp. 697–699, Apr. 1979.
[27] M. Teague, "Image analysis via the general theory of moments," *J. Opt. Soc. Amer.*, vol. 70, no. 8, pp. 920–930, Aug. 1980.
[28] J. M. Galvez and M. Canton, "Normalization and shape recognition of three-dimensional objects by 3D moments," *Pattern Recognit.*, vol. 26, no. 5, pp. 667–680, May 1993.
[29] L. Kurgan and K. J. Cios, "Discretization algorithm that uses class-attribute interdependence maximization," in *Proc. IC-AI*, Las Vegas, NV, 2001, pp. 980–987.
[30] H. Liu and R. Setiono, "Feature selection via discretization," *IEEE Trans. Knowl. Data Eng.*, vol. 9, no. 4, pp. 642–645, Jul. 1997.
[31] M. C. Ludl and G. Widmer, "Relative unsupervised discretization for regression problems," in *Proc. 11th ECML*, 2000, pp. 246–254.
[32] Y. Yang and G. I. Webb, "A comparative study of discretization methods for naive-Bayes classifiers," in *Proc. Pacific Rim Knowl. Acquisition Workshop*, 2002, pp. 159–173.
[33] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1997.
[34] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications*. Berlin, Germany: Springer-Verlag, 2006.
[35] M. Ben-Bassat, "Use of distance measures, information measures and error bounds in feature evaluation," in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1982, pp. 773–791.
[36] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[37] M. A. Hall, "Correlation based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.
[38] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Berlin, Germany: Springer-Verlag, 2001.
[39] R. O. Duda and P. E. Hart, *Pattern Classification*. New York: Wiley, 2001.
[40] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2, pp. 131–164, 1997.
[41] F. Guindos, J. A. Piedra, and M. Canton, "Ocean features recognition in AVHRR images by means of Bayesian net and expert system," in *Proc. 3rd Int. Workshop Pattern Recog. Remote Sens.*, 2004, vol. 11, pp. 1–4.
[42] Elvira Consortium, "Elvira: An environment for creating and using probabilistic graphical models," in *Proc. 1st Eur. Workshop Probab. Graph. Models*, 2002, pp. 222–230.
[43] D. Nauck and R. Kruse, "NEFCLASS—A neuro-fuzzy approach for the classification of data," in *Proc. ACM Symp. Appl. Comput.*, New York, 1995, pp. 461–465.
[44] V. Petridis and V. G. Kaburlasos, "Learning in the framework of fuzzy lattices," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 4, pp. 422–440, Aug. 1999.
[45] S. Moral, R. Rumi, and A. Salmeron, *Mixtures of Truncated Exponentials in Hybrid Bayesian Networks*. Berlin, Germany: Springer-Verlag, 2001, pp. 135–143.

Jose A. Piedra-Fernández was born in Spain in 1978. He received the degree in computer science and the M.S. and Ph.D. degrees from the University of Almería, Almería, Spain, in 2001, 2003, and 2005, respectively.

From 2008 to 2009, he was visiting the College of Information Sciences and Technology, The Pennsylvania State University, University Park. He is currently an Assistant Professor with the University of Almería. His main research interests include image processing, pattern recognition, and image retrieval. He has designed hybrid systems applied to remote sensing problems.

Manuel Cantón-Garbín was born in Adra (Abdera in phenician times) Almería, Spain, in 1955. He received the Physics (Electronics) degree from Granada University, Granada, Spain, in 1979, and the M.S. and Ph.D. degrees from La Laguna University, Canary Islands, Spain, in 1980 and 1982, respectively.

In 1984, he moved to Las Palmas de Gran Canaria University (ULPGC), Canary Islands, and in 1986, he led the first group in Spain working on satellite oceanography. He became a Lecturer in 1986 and a full Professor in applied physics with ULPGC in 1991. In 1994, he moved to the University of Almería, where he is currently a full Professor in computing and artificial intelligence. His research interests include automatic ocean satellite image interpretation and analysis, and computer vision.

James Z. Wang (S'96–M'00–SM'06) received the B.S. degree in mathematics and computer science (*summa cum laude*) from the University of Minnesota, Minneapolis, and the M.S. degree in mathematics, the M.S. degree in computer science, and the Ph.D. degree in medical information sciences from Stanford University, Stanford, CA.

He has been with the faculty at The Pennsylvania State University, University Park, since 2000, where he is currently a Professor with the College of Information Sciences and Technology. He is also currently an Affiliate Professor with the Department of Computer Science and Engineering, and the Integrative Biosciences (IBIOS) Program. From 2007 to 2008, he was a Visiting Professor with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. He has also held visiting positions with SRI International, IBM Almaden Research Center, NEC Computer and Communications Research Lab, and Academia Sinica. His main research interests are automatic image tagging, image retrieval, computational aesthetics, and computerized analysis of paintings.

Dr. Wang was a recipient of a National Science Foundation Career Award and the Endowed PNC Technologies Career Development Professorship.