Non-parametric classification algorithm with an unknown class

B. Gorte

Dept. of Geoinformatics
International Institute for Aerospace
Survey and Earth Sciences (ITC),
P.O.Box 6, 7500 AA Enschede,
the Netherlands

N. Gorte-Kroupnova

Electrical Engineering Department

University of Twente,
P.O.Box 217, 7500 AE Enschede
the Netherlands

Abstract

In the classification of pixels of a multispectral image by methods of supervised classification, a problem can arise in case when an unknown class is present. In this paper, we suggest a method that gives good results in such a case. The method provides an estimation for a posteriori probability vectors(and consequently, classification), and, besides, estimates the prior probability of classes, including the unknown one, and thus, the areas occupied by every class.

1 Introduction

In order to find objects on an image, one can classify pixels. If information about probability distributions of classes is available, supervised classification methods, such as maximum a posteriori probability (MAP) or nearest neighbor, are used. By MAP classification, the a posteriori probability vector is calculated for every pixel of the image, and each pixel is attributed to the class with maximum a posteriori probability.

To implement such a classifier, prior probabilities and conditional probability densities of all classes must be known or must be estimated from the available training sets. However, in some applications (in fact, in many), not all necessary information is present. There can be an unknown class present in the image, and the information about the number of pixels that belong to it is usually not available. For instance, in remote sensing images used for estimation of different vegetation types, classes of plants can be sampled very accurately. But there can be other classes with unknown probability distribution in the image, e.g. built-up areas, distorting the classification. Another example, also considered in this paper, concerns recognition of different electronic components on printed circuit boards (PCB). Some of the classes, such as integrated circuits (IC) and metal, are known, but others, such as the board itself or capacitors, can be of different colors from one PCB to another.

In these cases, classification with rejection can be used. In [4] two kinds of reject are considered: ambiguity reject and distance reject. The ambiguity reject indicates that there is not enough information in the training set to classify the pixel, or in other words, the pixel belongs to a region between the different classes. In case of nearest neighbor classification, the pixel is not attributed to any class (rejected) if the number of neighbors of each class is less than the qualifying majority level [1] - [3]. This type of rejection is not sufficient when an unknown class is present. In this case a distance reject that indicates that a pixel is located too far from all known classes to be attributed to any of them, must be used. In terms of probabilities, we would like to reject the pixels for which the maximum a posteriori probability belongs to the unknown class. In case of nearest neighbor the search radius [3] or the mean distance to the neighbors [4] can be thresholded. But this takes neither the unconditional probability density, which is different in every point of the feature space, nor the priors into account. This leads to classification errors. When a MAP classifier is used, it would be an advantage to be able to threshold the a posteriori probabilities in order to find the unknown class, for which, however, the prior probabilities and the unconditional probability density must be known. So Instead, Mahalanobis distance is, therefore, thresholded [5]. As a result, also in this case the unconditional probability density and the priors are not considered here. Another problem is that in both cases the threshold is not known and have to be found interactively, or have to be estimated from the training set.

The method suggested in this paper enables, with some assumptions, estimation of the a posteriori probabilities of every class, including the unknown class, and provides good classification results without the necessity of thresholding. It also provides the estimation of a priori probabilities and thus the areas occupied by

0-8186-7190-4/95 $4.00 © 1995 IEEE

every class.

2 Method of classification with unknown class

Our aim is to classify pixels of a multispectral image. Every pixel is characterized by a feature vector $\mathbf{x}$. Pixels must be attributed to one of the classes C_i, where i is a number of the class, including the unknown class. We assume to have complete information about known classes. This means that conditional probability densities of these classes $\mathrm{P}(\mathbf{x}|\mathrm{C_i})$ are known. We also assume that there are at least several "pure" pixels in every class which have an a posteriori probability near 1. These assumptions are reasonable in the mentioned applications, where a representative sampling of the known classes can be obtained. The unconditional probability density $\mathrm{P}(\mathbf{x})$ and the prior probabilities of classes $\mathrm{P}(\mathrm{C_i})$ are considered unknown.

We shall try to estimate a posteriori probabilities $\mathrm{P}(C_i|\mathbf{x})$ of M known classes $C_1 \ldots C_N$ as faithfully as possible, applying Bayes' formula:

$$\mathrm{P}(C_i|\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|C_i)\mathrm{P}(C_i)}{\mathrm{P}(\mathbf{x})} \tag{1}$$

Then we shall add those up and subtract the sum from 1, in order to obtain the (a posteriori) probability for the "unknown" class.

Of course, most maximum a posteriori probability (MAP) classifiers are based on Bayes' formula, but often they substitute

$$\mathrm{P}(\mathbf{x}) = \sum_{j=1}^{M} \mathrm{P}(\mathbf{x}|C_j)\mathrm{P}(C_j) \tag{2}$$

for the denominator of the right hand side of (1), which means that they only calculate the numerators of the right hand side and normalize the results to obtain the a posteriori probabilities. This has a few pleasant consequences:

- There is no need for $\mathrm{P}(\mathbf{x})$, which cannot be estimated easily. For example, while the (conditional) class distributions are often assumed to be Normal, this assumption can certainly not be made for the unconditional distribution of feature vectors. To simply take this distribution from the frequency of occurrence in the image (the multidimensional histogram) introduces too much noise, since the number of different feature vectors, compared to the number of pixels in the image, is quite large and, therefore, many of them occur only a few times. (Nevertheless, in the proposed method, we will be looking for unconditional feature densities.)
- The a posteriori probabilities look sensible: they are well in between 0 and 1 and their sum is, of course, exactly 1, even if the estimates of the conditional probability densities and the prior probabilities are not accurate.
- To shift the decision boundaries to the optimal positions (in terms of expected classification accuracy, or in terms of minimum cost of misclassification), it helps to exactly know the a priori probabilities. Even without this exact knowledge, sensible results can still be obtained. Note that in many applications exact prior probablities cannot be given beforehand. They are proportional to the areas covered by the various classes, and to find these is precisely the reason for making the classification. (How much tropical rain forest is there in the area?)

Unfortunately, the normalisation method leaves no room for an "unknown" class, since the sum of the a posteriori probabilities of the "known" ones already equals 1. The attempt to add an extra term to the normalisation sum fails immediately: to estimate the prior probability for "unknown" is tricky, whereas the (conditional) probability density for "unknown" is unknown almost by definition.

2.1 Estimation of conditional and unconditional densities

We will use an extended k-nearest neighbor (kNN) algorithm to estimate the conditional feature probability density $\mathrm{P}(\mathbf{x}|C_i)$ and the unconditional density $\mathrm{P}(\mathbf{x})$, at the same time.

The way to obtain the first is described, among others, in [6] and [7] . While looking in the feature space for k neighboring training samples, meanwhile counting k_i per class, we also keep track of the volume of the "ball" (in 3 dimensions; generally the "hypersphere") that we need to traverse. At the centre of the ball is $\mathbf{x}$; its radius increases until k neighbors are found. The volume $v_\mathbf{x}$ of the ball is the discrete number of feature space cells inside it; the ball itself is a set $B_\mathbf{x}$ of feature space cells.

If we use the symbol N_i for the total number of training samples of class C_i, we estimate the probability for such a training sample to be inside $B_\mathbf{x}$ as

$$\widehat{\mathrm{P}}(\mathbf{x} \in B_\mathbf{x}|C_i) = \frac{k_i}{N_i} \tag{3}$$

Now. assuming that the conditional feature density is constant inside the ball, it can be estimated as

$$\widehat{\mathrm{P}}(\mathbf{x}|C_i) = \frac{k_i}{N_i v_\mathbf{x}} \tag{4}$$

The assumption implies that the ball must not be too large, which will be elaborated later.

In addition to the above, we also count the total number of image pixels $T_{\mathbf{x}}$ that have their feature vectors inside the ball $B_{\mathbf{x}}$. Knowing how many pixels out of the total number T of pixels in the entire image are similar to $\mathbf{x}$, we estimate the probability that this happens to a "random" image pixel as

$$\widehat{\mathrm{P}}(\mathbf{x} \in B_{\mathbf{x}}) = \frac{T_{\mathbf{x}}}{T} \tag{5}$$

Assuming also the unconditional density to be constant inside the ball, this becomes

$$\widehat{\mathrm{P}}(\mathbf{x}) = \frac{T_{\mathbf{x}}}{T v_{\mathbf{x}}} \tag{6}$$

When using both the unconditional and the conditional density in Bayes' formula, the second is divided by the first and $v_{\mathbf{x}}$ disappears from the calculation. This can be considered advantageous: all the used quantities (k_i, $T_{\mathbf{x}}$ and $v_{\mathbf{x}}$) are of a stochastic nature; the possibility to eliminate one of them reduces the statistical "noise".

If we use $Q_i(\mathbf{x})$ to denote

$$Q_i(\mathbf{x}) = \frac{\mathrm{P}(\mathbf{x}|C_i)}{\mathrm{P}(\mathbf{x})}, \tag{7}$$

and rewrite (1) as

$$\mathrm{P}(C_i|\mathbf{x}) = Q_i(\mathbf{x})\mathrm{P}(C_i). \tag{8}$$

We estimate $Q_i(\mathbf{x})$ as $\widehat{Q}_i(\mathbf{x})$ using (4) and (6) as

$$\widehat{Q}_i(\mathbf{x}) = \frac{\widehat{\mathrm{P}}(\mathbf{x}|C_i)}{\widehat{\mathrm{P}}(\mathbf{x})} = \frac{k_i T}{N_i T_{\mathbf{x}}} \tag{9}$$

and use it like in (8) to obtain

$$\widehat{\mathrm{P}}(C_i|\mathbf{x}) = \widehat{Q}_i(\mathbf{x})\widehat{\mathrm{P}}(C_i). \tag{10}$$

2.2 Estimation of a priori and a posteriori probabilities

In a "pure" pixel of class C_i, having no spectral overlap with other classes, the a posteriori probability $\mathrm{P}(C_i|\mathbf{x}) = 1$.

In such a pixel, $Q_i(\mathbf{x})$ is at its maximum value Q_i^M, so there we can derive the prior probability $\mathrm{P}(C_i)$ of class C_i from (8)

$$1 = Q_i^M \mathrm{P}(C_i) \tag{11}$$

Since $\mathrm{P}(C_i)$ does not depend on $\mathbf{x}$ (it is valid for the entire image), we can now also substitute it in (8) for "unpure" pixels

$$\mathrm{P}(C_i|\mathbf{x}) = Q_i(\mathbf{x})\mathrm{P}(C_i) = \frac{Q_i(\mathbf{x})}{Q_i^M} \tag{12}$$

The key statement in this paper is that the same calculations can also be applied using our estimates $\widehat{Q}_i$. After calculating

$$\widehat{Q}_i(\mathbf{x}) = \frac{k_i T}{N_i T_{\mathbf{x}}} \tag{13}$$

for class C_i, and in every pixel which produces an image, we extract $\widehat{Q}_i^M$ from its histogram, estimate the prior probability, in accordance with (11) using

$$\widehat{\mathrm{P}}(C_i) = \frac{1}{\widehat{Q}_i^M} \tag{14}$$

and finally calculate an "image" of estimated a posteriori probabilities for class C_i with

$$\widehat{\mathrm{P}}(C_i|\mathbf{x}) = \frac{\widehat{Q}_i(\mathbf{x})}{\widehat{Q}_i^M} \tag{15}$$

Because $\widehat{Q}_i(\mathbf{x})$ is stochastic, some care must be taken during the calculation of $\widehat{Q}_i^M$. We cannot just take the global maximum; this is very likely to be some "outlier". Instead, during the experiments we used the following strategy. First we took only those pixels into account that are "pure", in the sense that all k neighbors belong to the class under consideration. From these pixels we calculated the histogram of $\widehat{Q}_i(\mathbf{x})$ and took the value where the cumulative histogram passes 0.95 as the estimate of the maximum.

It is interesting to observe that the stochastic nature of $\widehat{Q}_i(\mathbf{x})$ is not the only reason for its variability. Also among the pixels that belong to the "unknown" class, there are some that are much more similar to class C_i than to any other, and before knowing the unknown class they can only be considered "pure" C_i pixels. They, however, will get a low $\widehat{Q}_i$ value and, accordingly, $\widehat{\mathrm{P}}(C_i|\mathbf{x})$, as can be seen on a histogram (Figure 1, calculated for one of the known classes from the image on Figure 9).

2.3 Classification

After the a posteriori probabilities of all known classes are estimated, the a posteriori probability of the unknown class, as was already mentioned, can be easily calculated by subtracting the sum of the obtained probabilities from 1. Pixels are then classified in accordance to MAP algorithm that is attributed to the class with maximum aposteriori probability.

3 Experimental results

The suggested method was tested in two different applications. The first one is of a color image of a PCB, segmented in order to find electronic components on it. If we must look at many different PCB's, there are some classes which stay the same from one

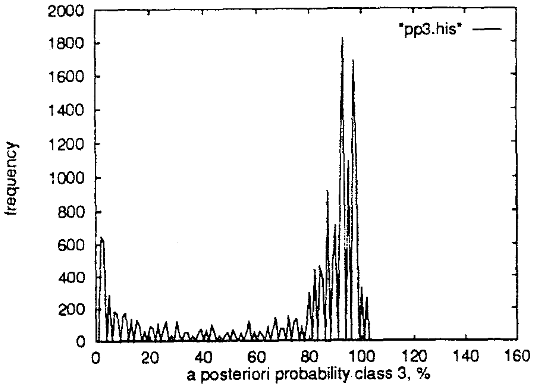

Figure 1: Histogram of $\widehat{\mathrm{P}}(C_i|\mathbf{x})$

PCB to another, such as pixels attributed to IC's and metal. These classes were thoroughly sampled and are assumed known. 1000 samples per class were used. Other classes can change from one board to another (background, other components). All these classes are to be classified as an unknown class. Every pixel is characterized by its color (intensity of red, green and blue in an RGB image), so we have 3D feature vectors. Such an RGB image of a PCB is shown (in grey-scale) in Figure 2. The a posteriori probabilities for every pixel were estimated with the algorithm described above, with the number of neighbors set to 50 and the search radius restricted to 5 in order to support the assumption about constant probability densities inside the neighborhood. The resulting a posteriori probabilities for all classes, including unknown, can be found in Figures 3 - 5, and the resulting MAP classification on Figure 6. For comparison, in Figure 7 the results of 50-NN classification with all classes known are presented. As can be seen, the results are quite similar.

Another kind of rejection - ambiguity rejection - was also applied to this image. The results are presented in Figure 8. Pixels, for which the difference between the highest a posteriori probability and the next one is smaller than a threshold, were rejected. Threshold 0.2 was used in this case. This is, in general, dependent on the number of classes. As expected, this works as a kind of edge detector: as a result of smoothing by point spread function of the camera, pixels on edges between objects have feature vectors situated somewhere between the classes.

The second experiment concerns a satellite image of an agricultural area in the Netherlands. The image was recorded by the "Thematic Mapper" sensor of a LANDSAT satellite, which measures reflected sunlight in six spectral bands (visible and infrared) with a spatial resolution of 30 m. One spectral band is presented in Figure 9. In the experiment three bands were used. The purpose was to make a map of agricultural crops in the area. Seven crops are predominant: grass, potatoes, wheat, sugar beets, peas, beans and onions, defining the seven classes of the classification. Agricultural survey data were available, from which 200 samples were taken for each class. Not part of the survey, although present in the area are a village, a few canals, some forested areas, roads, farmhouses and at least one orchard. Together these constitute our "unknown" class.

During the experiment, the prior probabilities were estimated as 0.02, 0.21, 0.17, 0.19, 0.03, 0.07 and 0.08, respectively, for the seven crops, leaving 0.23 for the "unknown" class. This result is reasonably in accordance with the survey data; an exact comparison is not possible since the survey is incomplete. Next, a posteriori probability images were calculated for each class. The histogram of one of these (wheat) is shown in Figure 1. Subtracting their sum from 1 gives an image of a posteriori probabilities, which is shown in Figure 10. This can be compared with the topographic map (which includes all our "unknown" subclasses); see Figure 11.

4 Conclusions and further work

The algorithm described in this paper allows classification of patterns in case an unknown class is present, and prior probabilities of classes are not known either. The method looks mathematically solid, without too many euristical assumptions. The classification is implemented in accordance to MAP criterium. The conditional and unconditional probability densities are estimated by a non-parametric method (modified kNN), then a posteriori probability densities of all classes, including unknown, are estimated, and patterns are honestly classified in accordance with maximum a posteriori probability. The method proved to work in two very different applications.

It must be noticed, that the method is based on the assumption of a representative sampling of the known classes. An investigation of what happens if the assumption is not true can be a future research topic.

One of the other questions to be considered in the future is further use of the obtained unknown class. We do not necessarily have only one unknown class in the scene, and it may be desirable to obtain further classification. One of the possibilities, as suggested in [8] and also in [4], is to use unsupervised classification (clustering) to reveal the structure of the unknown class. Another possibility is to use region-based segmentation techniques, such as region merging, to distinguish areas inside the unknown class.

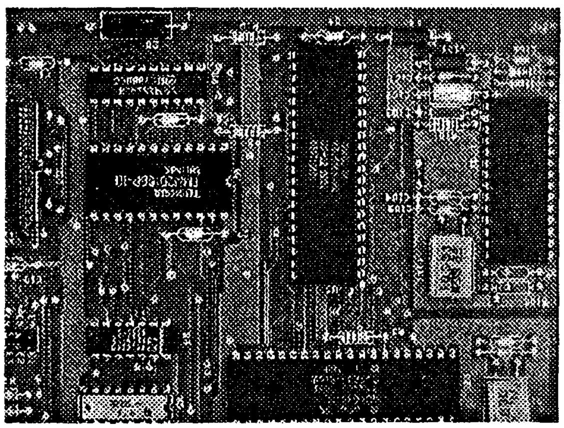

Figure 2: Image of a PCB

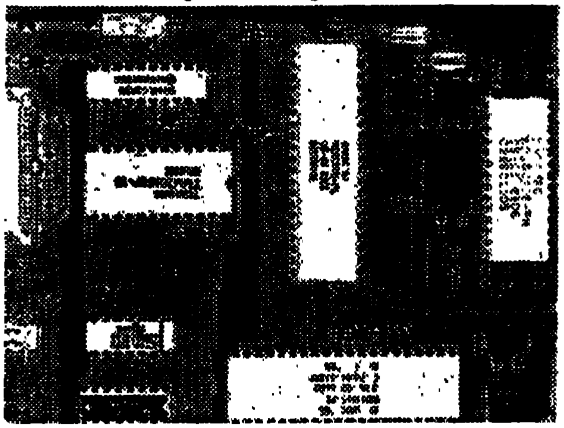

Figure 3: A posteriori probabilities of IC's

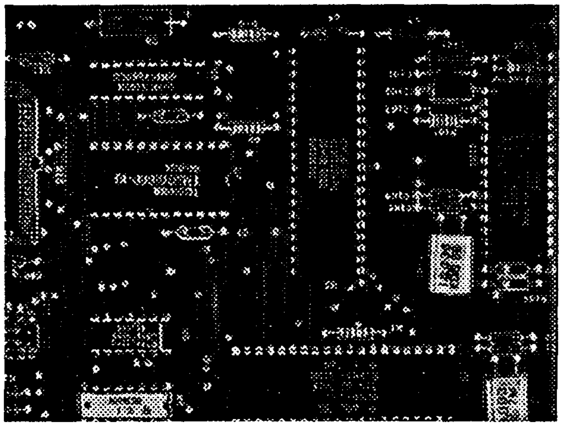

Figure 4: A posteriori probabilities of metal

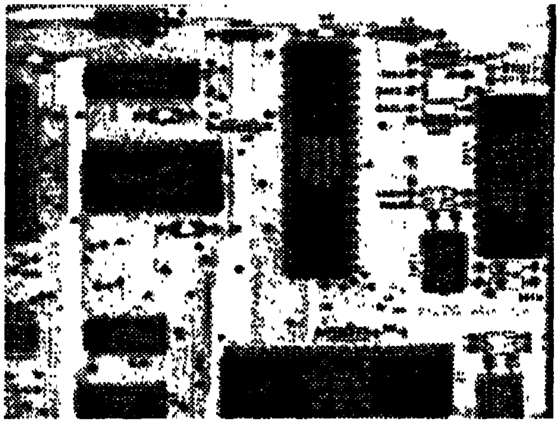

Figure 5: A posteriori probabilities of the unknown class

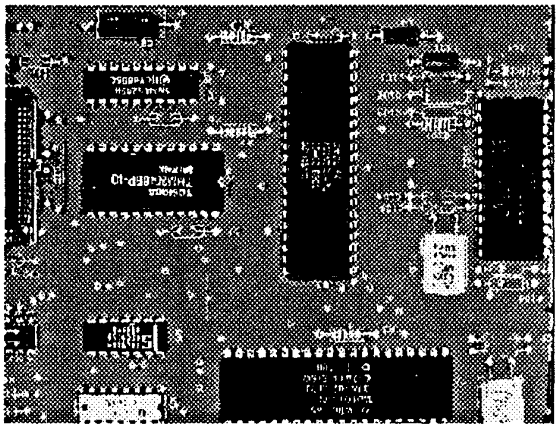

Figure 6: Classification results with unknown class

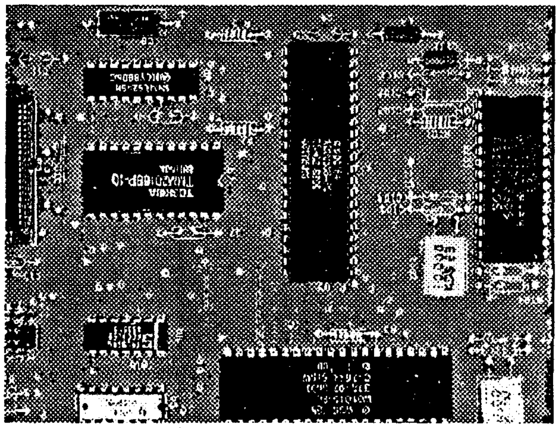

Figure 7: kNN classification with all classes known

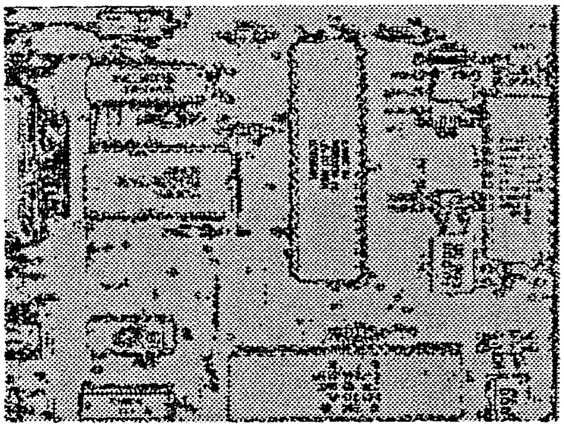

Figure 8: Pixels rejected by ambiguity criterium

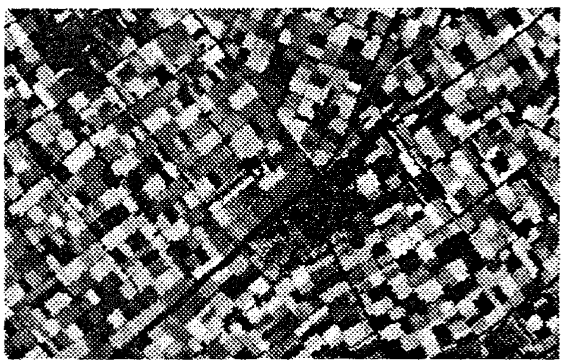

Figure 9: Remote sensing image

Figure 10: A posteriori probabiliti of unknown class from the remote sensing image

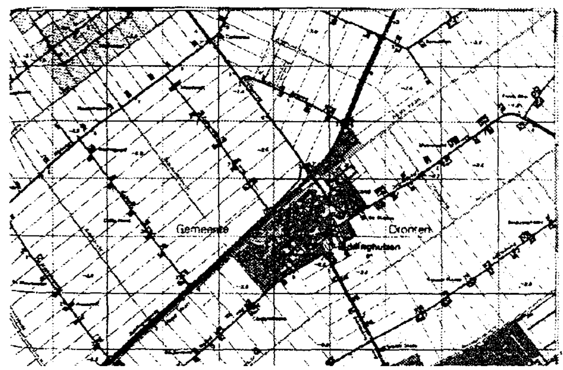

Figure 11: Topographic map

References

[1] M.E.Hellman, "The nearest neighbour classification rule with a reject option," *IEEE Trans. Syst. Man Cybern.*, Vol. 3, pp. 179-185, 1970.

[2] B.V. Dasarathy, "CODE: Cognition of Deliberate Entrants to a Scene" *Proceedings of the SPIE Symposium on Digital Processing of Aerial Images"*, Vol. 186, pp.208-215, 1979.

[3] B.V. Dasarathy, "Noising around the neighbourhood: a new system structure and classification rule for recognition in partially exposed enviroment," *IEEE Trans. PAMI.*, Vol. 2, 1980.

[4] B. Dubuisson, M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognition.*, Vol. 26, pp. 155-165, 1993.

[5] "ERDAS field guide" *Software documentation"*

[6] K.Fukunaga, D.M. Hummels, "Bayes error estimation using Parzen and k-NN procedures", *IEEE Trans. PAMI*, Vol. 9, pp. 634-643, 1987.

[7] C.W. Therrien, *"Decision, estimation and classification"* John Wiley & Sons, 1989.

[8] B.V. Dasarathy, "All you needed to know about the neighbors (but never could figure out how!): recognition under partial exposure and imperfect supervision", *Proceedings of the International Conference on Cybernetics and Society*, pp.218-221, 1979.