Comparative study of information fusion methods for sonar images classification

Arnaud Martin
E3I2 EA3876
ENSIETA
2 rue François Verny
29806 Brest Cedex 9, France.
Arnaud.Martin@ensieta.fr

Abstract – *Here, a comparative study of information fusion methods for sonar images classification is proposed. The automatic classification of sonar images is a very difficult problem. Our first task consists in finding a good image representation to classify the sea bottom. Classical approaches are based on texture analysis. Many methods can be considered to deal with this problem, however the best choice of the considered method depends often on the kind of sediment. Once the features extraction method has been considered, many classifiers can be used. In order to extract features, four major texture analysis methods have been considered. The four sets of features are classified and different methods of information fusion, such as the weighted vote approach, or coming from the possibility theory and evidence theory, have been employed.*

Keywords: Weighted vote, possibility theory, evidence theory, classification, sonar images.

1 Introduction

Sonar images are obtained from temporal measurements made by a lateral, or frontal sonar trailed with the back of a boat. Each emitted signal is reflected on the bottom then received on the antenna of the sonar with an adjustable delayed intensity. In order to build images a huge number of physical data (geometry of the device, coordinates of the boat, movements of the sonar,...) are taken into account, but these data are polluted with a large amount of noises due to used instrumentations. In addition, there are some interferences due to the signal travelling on multiple paths (reflection on the bottom or surface), due to speckle, and due to fauna and flora. Therefore, sonar images have a lot of imperfections such as imprecision and uncertainty; thus sonar classification is a difficult problem.

The perfect solution would be the fusion of information coming from several sensors. Nevertheless, this solution is not acceptable because it needs several boats.

The automatic classification approaches are based on texture analysis and a classifier such as a neuronal network. In the literature, techniques for texture analysis can be found and the choice of one or more of them depends on the kind of sonar and the applications [1].

Therefore in order to enhance the classification, we can fuse the data at the level of the texture features or at the level of the decisions of the classifiers (given by symbolic or numeric data). In this article, we compare different fusion techniques such as weighted vote, or methods coming from the possibility theory and evidence theory which work at the level of the decision making for numeric and symbolic data.

The paper is organized as follows. First we present the sonar images database and the problems inherent to such data. Then we briefly describe feature extraction by different kinds of texture analysis and the classifier. In the section 4, three fusion approaches are presented. Finally, some experimental results are discussed.

2 Sonar Images Database

Our database contains 26 sonar images provided by the GESMA (Groupe d'Etudes Sous-Marines de l'Atlantique). Theses images were obtained with a Klein 5400 lateral sonar with a resolution of 20 to 30 *cm* in azimuth and 3 *cm* in range. The sea-bottom deep was between 15 *m* and 40 *m*.

These 26 sonar images have been segmented in small-images with a size of 64x384 pixels (*i.e.* of approximately 1152 *cm* x 1152 *cm*). Figure 1 presents one image obtained by the sonar. On the left the water column is represented, on the right part of image we can not see the kind of sediment because of the noise. Figure 2 shows a sample of these small-images represented in order to obtain a size of 64x64 pixels.

Each small-image is manually characterized either by the type of sediment (rock, cobbles, sand, ripple, silt), or shadow when the information is unknown (see table 1). Moreover the existence of more than one kind of sediment on the small-image is indicated. In this last case, the major type of sediment in the small-image is indicated

From table 1, we notice that the sand sediment is the most represented one. The cobbles sediments are under-represented. A major classification difficulty is due to this difference. There are 39.7% of small-images with more than one kind of sediment (named patch-worked images).

Notice that such a database is quite difficult to realize. Indeed, the expert has a subjective experience, and can he can make mistakes on some small-images. So we only have a subjective perception of reality.

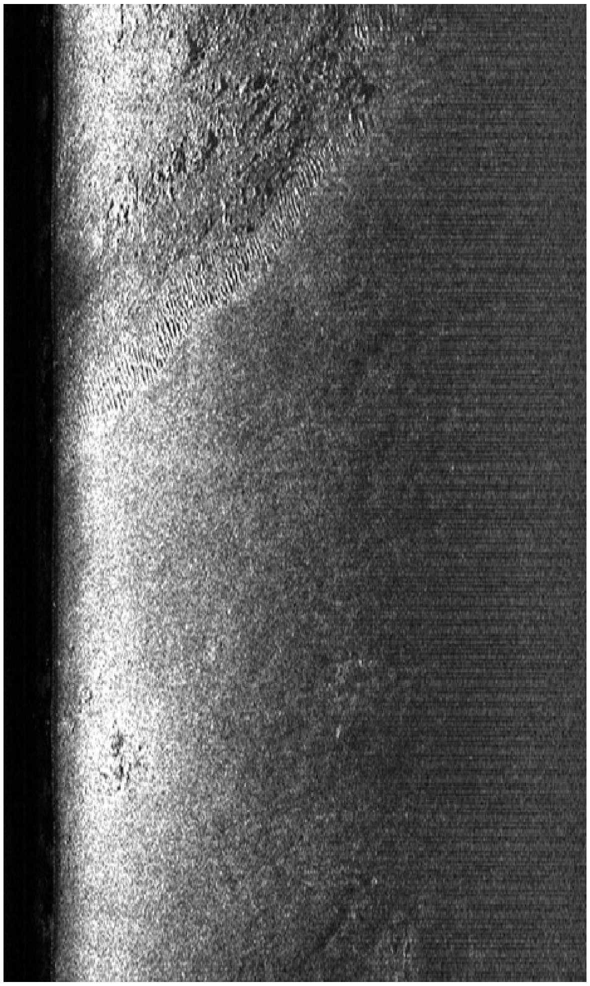

Figure 1. A sample of the sonar images.

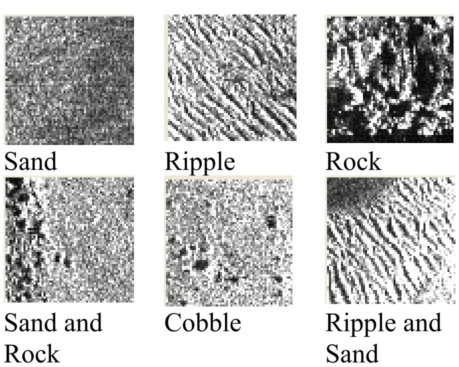

Figure 2. A sample of small-images with different types of sediments.

Sediment	Effective	%
Rock	915	21.35
Cobbles	33	0.77
Sand	2321	54.62
Ripple	374	8.80
Silt	234	5.50
Shadow	102	2.40
Total	4249	100.00

Table 1. Database elements and their effective.

3 Sonar images classification

Automatic classification of sonar images is generally made by automatic characterization of image texture. Many methods for texture analysis can be found in the literature and the selection of a method is not an easy task. This selection depends on the kind of images and on the applications. Once the image texture analysis has been made, any classifier can classify the extracted parameters. In this section, four features extraction methods are briefly discussed. The retained classifier is a multilayer perceptron classifier.

3.1 Features extraction

The features extraction methods presented here are based on four representations of the image: co-occurrence matrices, run-lengths matrix, wavelet transform and Gabor filters [2].

The co-occurrence matrices are calculated by numbering the occurrences of identical grey level of two pixels. Four directions are mainly considered: 0°, 45°, 90° and 135°. Concerning these four directions, six parameters given by Haralick [3] are calculated: homogeneity, contrast estimation, entropy estimation, the correlation, the directivity, and the uniformity. This classical approach yields 24 parameters. The problem for co-occurrence matrices is the non-invariance in translation. Typically, this problem can appear in a ripple texture characterization.

The run-lengths matrix is obtained by counting consecutive pixels with the same grey level in the four previous directions. Hence a matrix $\mathbf{L_d} = (L_d(i,j))$ is obtained, where $L_d(i,j)$ is the number of run lengths j of the pixels with a grey level i in the direction d. So the number of run lengths is given by:

$$N_L = \sum_{i=0}^{n_g-1} \sum_{j=1}^{n_g} L_d(i,j)\,. \tag{1}$$

Then five parameters are extracted for the four directional matrices: the proportion of small run-lengths, the proportion of big run-lengths, the run dispersion between the grey levels, and the run dispersion between the lengths. Hence this approach yields 20 parameters. This method is well suited in the case of optical images for example, where no speckle is present. Anyway, in the case of sonar images, we have to remove first the speckle or adapt the parameter calculation. However, we keep this approach in order to study the effect of a bad extraction of texture features.

The both previous approaches do not consider the translation invariance in the directions. The discrete translation invariant wavelet transform is based on the choice of the optimal translation for each decomposition level. Each decomposition level d gives four new images. We choose here a decomposition level d=3. For each image I_d^i (the i^{th} image of the decomposition d) we calculate three parameters: the energy, the entropy, and the mean. So we obtain 63 wavelet features (3+4x3+16x3).

The last texture analysis is given by the Gabor filter. We consider five different frequencies and six directions, thus 30 filters are designed. Then we calculate four parameters. The first one is the maximum value of the matrix numbers normalized by the mean, which represents the maximum value of the standard deviation with the considered sediment. The mean of all points of the matrix is also calculated. The third parameter represents the mean on the horizontal direction only (pings direction) normalized by the global mean. The last one is the global standard deviation before filtering. This approach takes into account the translation invariant on the directions.

3.2 Multilayer perceptron classifier

These four feature sets are independently considered as the input of a multilayer perceptron (MLP) classifier following the fusion architecture given in figure 3 and in [4].

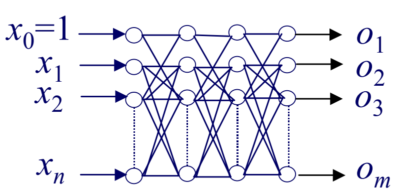

Figure 3. The multilayer perceptron structure.

The learning process of the multilayer perceptron is made with a sigmoid function given by:

$$f(x) = \frac{1}{1+\exp(-x)} \tag{2}$$

So that each neuron k of the output layer gives a value $o_k \in [0,1]$. Figure 4 represents an artificial neuron structure.

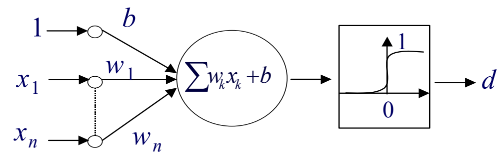

Figure 4. Artificial neuron structure.

These values o_k are considered as numeric outputs of the classifiers. The classifier decisions on the class of each observation x are taken with the maximum of the o_k values:

$$x \in C_k \quad \text{if } o_k(x) = \max_{1 \le j \le m} o_i(x)\,. \tag{3}$$

Hence we consider four MLP classifiers with 24, 20, 63, and 4 unit inputs and 6 unit outputs corresponding to the 6 considered classes.

4 Fusion approaches

We describe hereafter three theory frames of information fusion: weighted vote, possibility theory and evidence theory. The description is made following the four classical stages of fusion: modelization, estimation, combination and decision. Consider m sources S_j with j=1,…,m. Each source S_j gives some information on the observation x on the hypothesis that $x \in C_i$, where C_i is one of the n conceivable classes. Let us define by $M_i^j(x)$ this information.

The studied architecture of fusion is presented on figure 5. We follow the same architecture as the one presented in [4]. We will consider the numeric outputs o_{kj} (vector in $[0,1]^m$) of the classifiers or the classifier decisions C_{kj}.

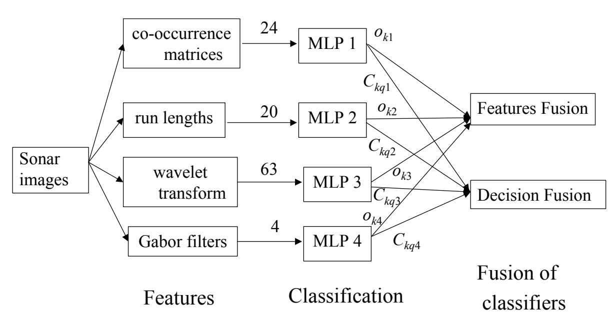

Figure 5. Architecture of fusion.

4.1 Weighted Vote

The weighted vote is the simplest method of information fusion. It is more a combination approach particularly well adapted for decisions fusion. Let us suppose that the n classes are exclusives. Modelization step consists in defining $M_i^j(x)$ by the following function:

$$M_i^j(x)=\begin{cases}1 & \text{if } x\in C_i,\\ 0 & \text{else.}\end{cases} \quad (4)$$

This function needs not to be estimated but it cannot take into account data imperfection. Reliability of sources can be introduced in the combination step [5]:

$$M_k^E(x)=\sum_{j=1}^{m}\alpha_{jk}M_k^j(x), \quad (5)$$

where α_{jk} is the reliability of the source S_j for the decision $x\in C_i$, such as $\sum_{j=1}^{m}\sum_{k=1}^{n}\alpha_{jk}=1$. Notice that this combination is associative and commutative. Adding the weights allows for a reduction of the conflict. The estimation of these weights α_{jk} can be made by the normalized confusion matrices. We can sum up the different decision rules by:

$$E(x)=\begin{cases}k & \text{if } M_k^E(x)=\max_i M_i^E(x)\geq cm+b(x),\\ n+1 & \text{else,}\end{cases} \quad (6)$$

where c is a constant in [0,1] and $b(x)$ is a function of $M_k^E(x)$. *E.g.:* the function b is given by $b(x)=\max_{k',k'\neq k} M_{k'}^E(x)$. Thus the decision $x\in C_k$ is taken if the difference between the number of sources saying $x\in C_k$ and the number of sources returning $x\in C_{k'}$ is large enough, otherwise the decision cannot be taken among the n exclusives classes. If the decision rule is too restrictive, ambiguous observations increase.

We assume that c=0.5, $b(x)$=0 (majority vote case), m odd and that the sources are statistically independent and give the same probability of successful. In this theoretical case [6] demonstrates that weighted vote gives best result in term of probability of successful. This result proves that under some assumptions, a fusion approach allows better classification rates.

4.2 Possibility theory

The possibility theory, proposed by L.A. Zadeh in 70's [7], has been next developed by D. Dubois and H. Prade [8]. This theory can take into account both imprecision and uncertainty data. These imperfections are modelized through a distribution of possibility and two functions based on it used for a event characterization: the possibility and the necessity. The distribution of possibility is defined on the space of discernment D={C$_1$,…, C$_n$} onto [0,1], such that:

$$\pi: D\rightarrow[0,1],\ \sup_{x\in D}\pi(x)=1. \quad (7)$$

Hence the first step of fusion is given by: $M_i^j(x)=\pi_x^j(C_i)$, where M_i^j is seen as a possibility degree for the assumption $x\in C_i$. We should assume that the classes are exclusives. The possibility degree represents a fuzzy number of the possible values of x and is an indication on the imprecision of x. Uncertainty can occur when both the event and its contrary are possible. The possibility of the contrary event is called: necessity. The possibility and necessity function are given for all $A\in 2^D$ (the set of all disjunction of D) by:

$$\prod(A)=\sup_{x\in A}\pi(x), \quad (8)$$

$$N(A)=1-\prod(A^c), \quad (9)$$

where A^c is the contrary event of A. One of the difficulties of such theory consists on the estimation of the distribution possibility. Many functions can be considered. Here, we interpret the output o_i of the classifier as a possibility degree for the class C_i.

Another interest of this theory is the large number of combination operators studied and developed. These operators can be classified into three classes: conjunctive (*e.g.* t-norm), disjunctive (*e.g.* t-conorm), and compromise operators (*e.g.* mean, OWA, Sugeno and Choquet integrals, …). The choice of one operator is not easy to do and depends on the context and application. So we have tested several operators: max, min, mean, median, and Sugeno integral.

The last step of the fusion process is the decision. It is made by the following rule:

$$x\in C_k \quad \text{if } \mu_k(x)=\arg\max_{1\leq i\leq n}\mu_i(x) \quad (10)$$

where $\mu_i(x)$ is the membership coefficient of x of the class C_i given here by the combination of the outputs of the classifiers.

4.3 Evidence theory

The evidence theory allows for a representation of both imprecision and uncertainty through two functions: plausibility and belief [9, 10]. Both functions are derived from a mass function defined by mapping of each subset of the space of discernment D={C$_1$, …, C$_n$} onto [0,1], such that:

$$\sum_{A\subseteq D} m(A)=1 \quad (11)$$

where $m(.)$ represents the mass function.

The first difficulty is the choice of a mass function. We can consider two types of approaches: one based on a probabilistic model [10] and another one based on distance transformation [11]. Appriou in [10] proposes two equivalent models based on three axioms. The first one that we use in this article is given by:

$$\begin{cases} m_j^i\left(\{C_i\}\right)(x) = \alpha_{ij} R_j p(q_j / C_i) / (1 + R_j p(q_j / C_i)) \\ m_j^i\left(\{C_i\}^c\right)(x) = \alpha_{ij} / (1 + R_j p(q_j / C_i)) \\ m_j^i(D)(x) = 1 - \alpha_{ij} \end{cases} \quad (12)$$

where q_j is the j^{th} classifier (supposed cognitively independent), j=1,…,m, α_{ij} are reliability coefficients on each classifier j for each class i=1,…,n (in our application we take α_{ij}=1), and $R_j = (\max_{q_j,i}(p(q_j / C_i)))^{-1}$. Hence a mass function is defined for each classifier j and each class C_i. In this approach, the difficulty is the estimation of the probabilities $p(q_j / C_i)$. In the case of decision level, C_i is the class given by the classifier j. Hence the estimation of these probabilities can be made easily on a learning database using the confusion matrices. In the case of characteristic level, the estimation can be made classically by the frequencies or making assumption on the distribution of theses probabilities. For this level the distance approach is easier.

Indeed, in [11] the mass functions are defined by:

$$\begin{cases} m_j^i\left(\{C_i\} / \mathbf{x}^{(t)}\right)(\mathbf{x}) = \alpha_{ij} \varphi_i(d^{(t)}) \\ m_j^i(D / \mathbf{x}^{(t)})(\mathbf{x}) = 1 - \alpha_{ij} \varphi_i(d^{(t)}) \end{cases} \quad (13)$$

where $\left(\mathbf{x}^{(t)}\right)_t$ is a set of learning vectors, $d^{(t)} = d(\mathbf{x}^{(t)}, \mathbf{x})$ is a distance (to be determined) between $\mathbf{x}$ and $\mathbf{x}^{(t)}$ and C_i is the class of $\mathbf{x}^{(t)}$. φ_i is a distance function, which verifies:

$$\begin{cases} \varphi_i(0) = 1 \\ \lim_{d \to +\infty} \varphi_i(d) = 0 \end{cases} \quad (14)$$

Many functions can be used, in [11] Denoeux proposes as an Euclidian distance:

$$\varphi_i(d) = \exp(-\nu_i d^2), \quad (15)$$

where ν_i is a positive parameter according to the class C_i. We will use this function. The distance calculation $d^{(t)} = d(\mathbf{x}^{(t)}, \mathbf{x})$ can take time if the training database is important, but we can consider only the k nearest neighbors.

The fundamental difference between both approaches is that in the first case we have to estimate the probabilities $p(q_j / C_i)$ and in the second case the distance d. For the decision stage the estimation of $p(q_j / C_i)$ is very easy, but it is quite difficult to choose an appropriate distance in this case (symbolic distance). On the contrary, for the characteristic level, the estimation of $p(q_j / C_i)$ can be difficult if the distribution is unknown, and the Euclidian distance, for example, can be chosen for d. In this paper, we will apply the probabilistic approach for the decision stage (*i.e.* C_{kq} outputs of the MLP), and the distance approach for the characteristic level (*i.e.* o_k outputs of the MLP).

The combination of mass functions is based on the orthogonal non-normalized Dempster-Shafer's rule:

$$m(.) = \bigoplus_{i=\overline{1,N}} \left(\bigoplus_{j=\overline{1,m}} m_j^i(.) \right) \quad (16)$$

In the case of the distance approach, this combination can be rewritten very simply [11]. Other combination rules can be used, but we keep this rule as it gives the best result on our data.

The last stage of the fusion process is the decision. In the evidence theory, we can use the maximum of plausibility, maximum of belief or maximum of pignistic probability [12]. We make a compromise by keeping the maximum of pignistic probability in this article.

5 Experiments

Our database has been randomly divided into three equal parts. The first one is used for the MLP learning, the second one for the fusion learning and the third for tests. We repeat this random division 10 times in order to achieve a good estimation of the classification rate, and we analyze the mean percentage of good classification rates defined as the number of good classified small-images dived by the total of small-images.

The classification rates for the four texture analysis with MLP are given in table 2. We have also tested a global MLP with in input all texture features given by the four methods, referred in table 2 as the MLP global. The classification rates are given with the confidence interval at 95%.

Classification	%	Variance
Co-occurrence matrices with MLP	70.0 ± 2.46	6.69
Run length with MLP	50.3 ± 2.68	1.08
Wavelet with MLP	68.9 ± 2.48	34.85
Gabor filter with MLP	66.4 ± 2.53	2.35
MLP global	50.0 ± 2.68	6.95

Table 2. Classification performance.

On table 2, notice that co-occurrence matrices, wavelet transform and Gabor filter give the best results, but there are not significantly different. Using the run length representation on this kind of images, give very bad results. The global multilayer perceptron is not robust to the bad feature coming from the run length. We have also performed the fusion of the fourth classification approaches by a MLP with 24 unit inputs. For the same raisons than explained before, we have obtained 49.6 ± 2.68% with a variance of 0.95. All small-image are classified in sand as using the run length representation.

In the case of majority vote, we have observed a conflict of 18.59% estimated by the number of small-images for which the decision cannot be determined. Hence we choose a weighted vote principle. The weights α_{jk} are estimated on the second part of the database, with the confusion matrices. The classification rate with the fusion of the four texture analysis by weighted vote gives 62.0 ± 2.60 %. Notice that results with this approach are statistically worse than classification by co-occurrence, wavelets or Gabor filter alone. However weighted vote is more robust than global MLP.

In the context of the possibility theory we have tested different operators. The classification rates given in table 3 show that there is no significant difference between the tested operators except for the Sugeno integral, which gives worst results. The result of the Sugeno integral is conformable to the result given in [4] for another application. The best result is reached for the *t*-norm max (with 69.9 ± 2.46 %). The best classification rate for the possibility theory is comparable to the best classification with co-occurrence only. Possibility theory in this application is also more robust to bad features than the global MLP.

Operator	%	Variance
t-norm: min	66.1 ± 2.54	29.77
t-conorm: max	69.9 ± 2.46	4.11
Mean	67.8 ± 2.51	5.29
Median	67.5 ± 2.51	4.18
Sugeno integral	61.8 ± 2.61	2.78

Table 3. Classification performance for possibility theory.

Sediment	%	Variance
Rock	87.3 ± 1.79	5.15
Sand	84.9 ± 1.92	3.32
Ripple	61.3 ± 2.61	8.36
Silt	4.9 ± 1.16	27.58
Cobbles	0.9 ± 0.51	8.26
Shadow	71.5 ± 2.42	66.86
No patch worked	91.3 ± 1.51	1.58
Patch worked	63.1 ± 2.59	3.18

Table 4. Detailed classification performance for evidence theory with mass function based on distance.

For the evidence theory based method we have tested two different estimation approaches for mass functions: one based on a probability and one based on a distance. We have obtained a classification rate of 69.9 ± 2.5 % for the mass function based on probability (equation (10)) with a variance measurement equal to 3.01 and a classification rate of 79.5 ± 2.17 % for the mass function based on distance (equation (11)) with a variance measurement equal to 2.03. Probability-based approach is comparable to the best classification rate with co-occurrence only. Moreover the distance-based approach outperforms all the classifications with only one texture analysis method and all other fusion methods. The difference between classification rates is statistically significant. In the table 4, we present the detailed classification performance for the mass function based on distance.

Note that the best rates are reached for rock and sand sediments, which are the most numerous in the database. For silt and cobble sediments, the classification rates are bad. This can be explained by the fact that the more a sediments is represented in the database, the best it will be classified by the MLP. Inversely, the classifier based on texture analysis do not behave well for sediments less represented in the databases. Notice also that classification rates are very good for small-images that are not patch-worked (91.3 ± 1.51 %). We observe similar results for the other fusion methods, but with classification rates not so good.

6 Conclusions

We have proposed here a comparative study between several information fusion strategies for sonar images classification. We have presented fusion methods in three different contexts: weighted vote, possibility theory and evidence theory. We have seen that weighted vote and possibility theory with several operators give classification rates comparable to the one based on co-occurrence. However these fusion methods are more robust to badly extracted features whereas the global MLP is not. The best performance is obtained with the evidence theory approach based on distance. This method allows a significant improvement.

Weighted vote is very simple to apply in a context of classifier fusion. The learning of the weights by the confusion matrices allows a reduction of conflict. The advantage of our possibility-based method is the non-learning phase due to the interpretation of classifier outputs as a possibility degree for the class (except for Sugeno integral). In general, we can interpret the classifier outputs as a possibility degree for the class if outputs are in [0,1]. Thus these approaches are very simple to apply. Both presented methods using the evidence theory are also very simple to apply for classifier fusion. The distance-based one outperforms other fusion methods but takes more processing time (especially for the learning step).

The interest of all presented fusion methods for classifier fusion is that we do not need any knowledge on data. We can integrate these methods for any automatic classification process.

In order to improve the performance of the sediment classification, there are two problems to resolve. An important problem for a multilayer perceptron classifier comes from the effective difference of the kind of sediments in our database. The learning for the type of sediment few represented is bad. Another problem is the patch-worked small-images. Therefore, we are working on the realization of a new repartition of the data with a previous manual segmentation of the sediment.

References

[1] Thomas H., Collet C., Yao K., and Burel G., *Some improvements of a rotation invariant autoregressive method. Application to the neural classification of noisy sonar images*, European Signal Processing Conference (Eusipco), Rhodes 1998, Vol. 4, pp. 2001-2004.

[2] Martin A., Sévellec G., and Leblond I., *Characteristics vs decision fusion for sea-bottom characterization,* In Situ Sea Bed Characterization Conference, Brest 2004.

[3] Haralick R., *Statistical and textural approaches to textures*, Proceedings of the IEEE, Vol 67, No 5, pp 786-804, 1979.

[4] Martin A., and Radoi E., *Effective ATR Algorithms Using Information Fusion Models*, Information Fusion Conference, Stockholm, Sweden 2004.

[5] Xu B.V., Krzyzak A., and Suen C.Y., *Methods of Combining Multiple Classifiers and Their Application to Handwriting Recognition*, IEEE Transactions on Systems, Man and Cybernetics, Vol 22, No 3, pp. 418-435, 1992.

[6] Lam L., and Suen C.Y., *Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance*, IEEE Transactions on Systems, Man and Cybernetics, Vol 27, No 5, pp. 553-568, 1997.

[7] Zadeh L.A., *Fuzzy Sets as a Basis For a Theory of Possibility*, Fuzzy Sets and Systems, 1: pp 3-28, 1978.

[8] Dubois D., Prade H., *Possibility Theory*, Plenum Press, New York, 1988.

[9] Bloch I., *Some Aspects of Dempster-Shafer Evidence Theory for Classification of multi-modality medical images Taking Partial Volume Effect into Account,* Pattern Recognition Letter, Vol. 17, pp. 905-919, 1996.

[10] Appriou A., *Situation Assessment Based on Spatially Ambiguous Multisensor Measurements*, International Journal of Intelligent Systems, Vol 16, No 10, pp. 1135-1166, 2001.

[11] Denoeux T., *A k-nearest neighbor classification rule based on Dempster-Shafer Theory*, IEEE Transactions on Systems, Man and Cybernetics, Vol 25, No 5, pp. 804-913, 1995.

[12] Smets P., *The Combination of the Evidence in the Transferable Belief Model*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 12, No 5, pp. 447-458, 1990.