Proceedings of the Institute of Acoustics

Feature extraction from backscatter sonar data

Otto Milvang, Ragnar Bang Huseby, Katrine Weisteen, Anne Solberg.
Image Processing group
Norwegian Computing Center
Oslo, Norway

Abstract

In seabed classification from sonar data, it is important to extract reliable features which may be used to discriminate between different seabed types. In this paper, we evaluate a number of different methods for feature extraction from sonar data.

Raw backscatter data from the Simrad EM 1000 Multibeam Echo Sounder were selected from five areas of various seabed types. Each area was divided into 28 smaller regions. More than 50 features were selected from each region. The features may be divided into four categories, features based on the backscatter strength, features based on the backscatter probability density function, features based on the spectral distribution, and features based on texture. All features were evaluated based on their ability to differentiate between the various seabed types, and the sensitivity with respect to the number of samples used.

We have shown that features based on the backscatter strength and features based on the spectral distribution discriminate the seabed types very well, while some of the features based on the probability density function and one of the texture features may give additional separation between the seabed types.

1 Introduction

This work is a part of a seafloor mapping project. The project was started in December 1990 by Simrad Subsea A/S and Norwegian Computing Center (NR). At NR, the work has been concentrated on feature extraction and statistical classification [4]. In January 1992, Center for Industrial Research (SI) and NR started a work on neural network classification[5], and in June 1992 Simrad A/S and NR joined the ESMAC project.

Feature extraction is the foundation of successful classification. It is known that the backscatter strength is correlated with the seabed type. In [1], Pace & Gao suggested a method to extract spectral information. A texture measure was proposed by Czarnecki [2]. A set of texture measures were used by Pace and Dyer in [3]. All these works were done on side scan sonar, while our data were logged using a multibeam echo sounder. In this study, we compare the methods described in [1], [2] and [3] in addition to other methods.

2 Data

The data used in this project were recorded from an Simrad EM 1000 multibeam echo sounder, 95 kHz. It is designed for operation in water depths from 5-800 meters.

For each ping the EM 1000 returns a stream of backscatter values. These values are corrected according to angle of incidence assuming flat bottom. In addition to the backscatter data, the sonar also returns position information and depth data. The number of samples in a ping varies with the depth. A ping covers a sector of 150° or about 7.5 times the depth, and the sampling rate is about 6-7 samples/meter. A ping is divided into 60 beams, each covering 2.5° . In the evaluation of features we have only used samples from beam 4-23.

Due to a cruise in Oslofjorden, we got a large data set with a verbal description. By examination of the data in combination with the verbal description of the data we selected five regions which we assume represent five different seabeds. In this stage of the project it is not important to know exactly what seabed types the 5 regions represent so we label them type 1, ..., type 5. The classes may correspond to rock, sand, silt, clay, and mud respectively, but we stress that this is not confirmed. What is important, is that the five regions represent five different seabeds. Each region is divided into seven adjacent areas. For each area we have used three (overlapping) subareas, containing 2000, 4000, and 8000 backscatter samples respectively. (Table 1).

For each feature we want to examine, we will use these 5 regions × 7 adjacent areas × 3 subareas = 105 subareas.

Mean value and standard deviation are calculated for groups of 7 adjacent areas of the same size. The mean values and standard deviation for the areas containing 8000

Region	#ping	across(m)	along(m)
1	16	150	32
2	44	30	48
3	44	35	40
4	40	35	60
5	100	15	120

Table 1: The size of an area on the seafloor covered by 8000 samples. The values for 2000 samples and 4000 samples may be calculated by dividing along distance with 4 or 2 respectively.

samples are used to estimate the generalized Mahalanobis distance[6]. To achieve good classification the classes must be separated in the feature space. The Mahalanobis distance is used to give a numerical measure of the distance between two classes i and j. The generalized Mahalanobis distance $M_{i,j}$ is:

$$\begin{aligned} M_{i,j} &= \sqrt{\delta^2 + \gamma^2} \\ \delta^2 &= (\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) \\ \gamma^2 &= 4 \log \frac{|\Sigma|}{|\Sigma_i|^{\frac{1}{2}} |\Sigma_j|^{\frac{1}{2}}} \\ \Sigma &= \frac{1}{2}(\Sigma_i + \Sigma_j) \end{aligned}$$

where μ_i og μ_j are mean vectors, Σ_i, Σ_j are covariance matrixes for class i and j. If the distance between two classes is larger than 4.0, the probability of overlap between the two classes are 2.5 % or less provided that the classes are normal distributed. Experience shows that if the distance is greater than 10.0, hardly any misclassification occurs even if the classes are not normal distributed.

From a good feature we may also expect that the feature is (more or less) invariant to the number of backscatter samples used in the calculation.

3 Preprocessing

The sonar data contains a lot of artifact pings, beams or samples. These artifact data must be removed, otherwise it will result in wrongly computed features. All samples in a ping are treated as a continuous sequence of samples. That means that the last sample in one beam is the neighbor of the first sample in the next beam, and this neighbor relation is the same as for two subsequent samples in the same beam. The beams straight below the boat give an artificially strong echo and are not used. For the EM1000 with 60 beams we have used beam 4 to beam 23. Each ping is examined for bad samples. First, all samples with very low or high backscatter value are rejected. We have rejected backscatter > 0 dB and < -50 dB. Isolated rejected samples are replaced by the mean value of the predecessor and the successor. Groups of rejected samples will split the sample sequence into two subsequences. For each ping the longest subsequence of accepted samples is used. If this subsequence is shorter than a *minlimit* the entire ping is rejected.

4 Feature extraction

The backscatter values are denoted $X = \{x_1, \ldots, x_n\}$, and all features are tested for $n = 2000, 4000$, and 8000. The vector is formed by taking all samples from the first ping in the area and concatenate these with all samples from the second ping, and continue this concatenation until the vector contains n samples. The advantage of this is that all algorithms have an input of vector of length n and is independent of the number of pings used to build the vector. The disadvantage is the fact that sample x_i and x_{i+1} are either neighbors or are the last/first sample in two subsequent pings. The latter may give rise to discontinuos jumps in the data if x_i and x_{i+1} represent different seabed types. However, since the ratio "not neighbor/neighbor" is max $100/8000 = 1.25\%$, this influence on the result is small.

The features are divided into four categories, features based on the backscatter strength, features based on the backscatter probability density function, features based on the spectral distribution, and features based on texture.

4.1 Backscatter strength

The most obvious feature to use is the mean value of the backscatter values.

$$mean = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The backscatter values fluctuate a lot and it is known that the mean value is sensitive to the existence of a few unusually large or small values. Order statistics is a set of methods which is usually more robust than classical statistics. A common method for summarizing the distribution of a random variable is by giving some selected *quantiles* of the random variable. The term "quantile" is not as well known as the terms "median", "quartile", "decile", and "percentile", yet these latter terms are popular names given to particular quantiles. The median of a random variable, for example, is the number the random variable will exceed

with probability 0.5 or less and will be smaller than with probability 0.5 or less. This definition may be extended as follows.

The number Q_p for a given value of p between 0 and 1, is called the pth quantile of the vector X, if $P(x_i < Q_p) \leq p$ and $P(x_i > Q_p) <= 1 - p$. If more than one number satisfies the definition of the pth quantile, we will avoid confusion by adopting the convention that Q_p equals the average of all numbers that satisfy the definition.

The median is the 0.5 quantile, the third decile is the 0.3 quantile, the upper and lower quartiles are the 0.75 and 0.25 quantiles respectively, and the sixty-third percentile is the 0.63 quantile.

For a small area of the seabed all raw backscatter data inside the area were sorted and the $0.1, 0.2, \ldots, 0.9$ quantiles were extracted.

4.2 Backscatter probability density function

The probability density function (PDF) of the raw backscatter measurements has been shown to be sensitive to changes in the roughness of the seabed.

In addition to locating the center of the PDF, described in the previous section, it is useful to measure the extent of variation around the center. A common used measure is the standard deviation.

$$stdev = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Another measure of spread is the mean deviation:

$$meandev = \frac{1}{n}\sum_{i=1}^{n} \mid x_i - \overline{x} \mid$$

Other measures useful in comparing shapes of PDFs are the k'th-order central moments:

$$M_k = s^{-k}\sum_{i=1}^{n}(x_i - \overline{x})^k,\ k = 3,4,5$$

So far we have only considered features which are not based on any assumption on the shape of the PDF. If one assumes that the PDF is a normal density the mean and the standard deviation completely characterize the PDF. However, in many cases, the PDF seems to be asymmetric, contradicting the normal density assumption. A class of asymmetric PDFs is the gamma family. The PDF is given by

$$f(y; r, \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)}, y > 0$$

where r and λ are parameters. We have used two methods for estimating the parameters, the method of moments and the method of maximum likelihood[7].

The moment estimates, $\hat{r}_{MOM}$ and $\hat{\lambda}_{MOM}$, satisfy the equations:

$$m(\hat{r}_{MOM}, \hat{\lambda}_{MOM}) = \overline{x}$$
$$v(\hat{r}_{MOM}, \hat{\lambda}_{MOM}) = s^2$$

where $m(r, \lambda)$ is the expectation and $v(r, \lambda)$ is the variance of the gamma distribution. Because $m(r, \lambda) = \frac{r}{\lambda}$ and $v(r, \lambda) = \frac{r}{\lambda^2}$ [7] we get the r-moment:

$$m(r, \lambda) = \frac{\overline{x}^2}{s^2}$$

and the λ-moment:

$$v(r, \lambda) = \frac{\overline{x}}{s^2}$$

Using the method of maximum likelihood the estimates $\hat{\lambda}_{ML}$ and $\hat{r}_{ML}$ will be those values of λ and r which maximizes

$$\prod_{i=1}^{n} f(x_i; \lambda, r)$$

Then we get r-maximum likelihood:

$$rml = \psi^{-1}(\frac{1}{n}\sum_{i=1}^{n}\log x_i - \log \overline{x})$$

and λ-maximum likelihood:

$$\lambda ml \frac{\hat{r}_{ML}}{\overline{x}}$$

Here ψ^{-1} denotes the inverse of ψ where ψ is the derivative of the logarithm of the Gamma function.

4.3 Spectral distribution

The work on power spectrum methods[1] has been concentrated around exploring the nature of the power spectra and the features extracted from them.

The power spectrum is obtained by first applying a window function to the input data, performing a Fourier transform on the windowed data and then taking the sum square of the resulting amplitudes. The windowing is done by first adjusting the data to have a zero mean value, then applying the window function and then re-adjusting the data.

$$P_i(f) = \left| F\left[g_i(t)\right]\right|^2, \qquad 1 \leq i \leq n$$

where i is the ping number and $g_i(t)$ is the windowed input data.

The power spectrum is then averaged over n pings, giving

$$\overline{P}(f) = \frac{1}{n}\sum_{i=1}^{n} P_i(f).$$

In [1], the "log-power spectrum" is defined as

$$P_L(f) = log\left(\frac{A\,\overline{P}(f)}{P_m} + 1\right) / log(A+1),$$

where P_m is the maximum value of $\overline{P}(f)$ and A is a constant multiplier. Finally, applying normalization, we get :

$$P_{NL}(f) = P_L(f) \Big/ \int_0^{f_{NY}} P_L(f)\,df$$

From the normalized log-power spectrum, Pace & Gao define three features, D_{f_1}, D_{f_2} and D_{f_3}. These parameters have been modified somewhat compared to the description given in [1]. In our version data from several pings are merged together into a long array. From this array, segments of 64 elements are Fourier transformed. The segments are overlapping such that the start elements are 0, 32, 64, and so on. The mean of all Fourier transforms are calculated and the D_f features are computed. We will call these features the "Pace features".

The zero frequency is not used in the computation.

$$D_{f_1} = \int_1^{f_{BA}} P_{NL}(f)\,df \Big/ \int_{f_{BA}}^{f_{NY}} P_{NL}(f)\,df$$

$$D_{f_2} = \int_1^{1/2 f_{BA}} P_{NL}(f)\,df \Big/ \int_{f_{BA}}^{f_{NY}} P_{NL}(f)\,df$$

$$D_{f_3} = \int_1^{1/2 f_{BA}} P_{NL}(f)\,df \Big/ \int_{3/4 f_{NY}}^{f_{NY}} P_{NL}(f)\,df$$

The method was also used in connection with a smoothing filter. The data were smoothed before the Fourier transform were calculated. For smoothing of the input data we have tried the median filter with different sizes and the Crimmins filter [8] with different numbers of iterations. Several values for the BA frequency were also tested.

The results can be seen in Figures 1 and 2. The two figures show the power spectra for the five types of seabed before and after smoothing. The DC value is by far larger than the rest and is therefore omitted.

The smoothing operation obviously removes a lot of noise in the higher frequencies and furthermore, judging from the results (see below), it does not seem to remove significant information.

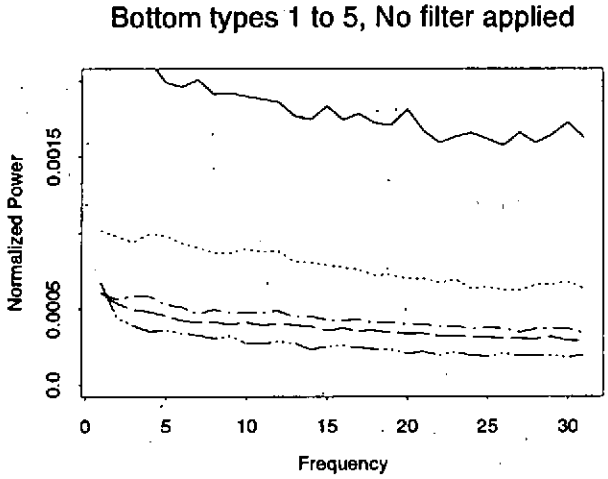

Figure 1: Normalized power spectrum without filtering of the five seabed types

4.4 Texture

An important approach to image analysis is to quantify the texture content [9][10][11]. Although no formal definition of texture exists, we intuitively view this descriptor as providing a measure of properties such as smoothness, roughness, coarseness, and regularity. Most surfaces in nature are not smooth. The texture of the material corresponds to the surface of the material and may therefore be used to identify structures. In sonar images the variation in reflectivity may correspond to structures on the seabed. The three principal approaches used in image processing to describe the texture of a region are spectral, statistical, and structural. Spectral techniques are based on properties of the Fourier spectrum and are used primarily to detect global periodicity in the images. Statistical approaches yield characterizations of textures as smooth, coarse, granularity, and so on. Structural techniques are based on regular repeated patterns in the images. For the sonar imaging these methods may be applied to reflectivity data or depth data.

4.4.1 Gray-Level Co-Occurrence (GLCM)

Textural features can be calculated from the gray level spatial co-occurrence matrix. If a ping is treated as a sequence of reflectivity data, the co-occurrence $P_d(i,j)$ of reflectivity i and j, is defined as the number of pairs of samples having reflectivity i and j, respectively, and which are in a fixed spatial relationship d. The co-occurrence matrix can be normalized by dividing each entry by the sum of all entries in the matrix giving $p_d(i,j)$. A lot af papers describe statistics of the co-occurrence matrix[9][10][11]. Some of the most important features are:

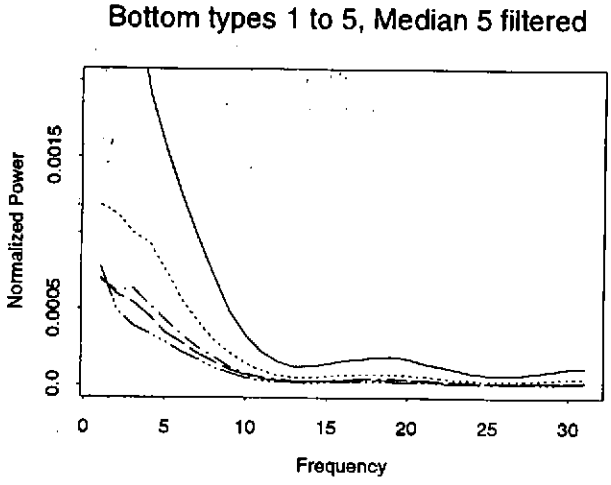

Figure 2: Normalized power spectrum after median filtering of the five seabed types

$$energy = \sum_i \sum_j p(i,j)^2$$

$$entropy = \sum_i \sum_j p(i,j) \log(p(i,j)$$

$$contrast = \sum_i \sum_j p(i,j)(i-j)^2$$

$$correlation = \sum_i \sum_j p(i,j)(i-\mu_i)(j-\mu_j)/\sigma_i\sigma_j$$

$$homogenity = \sum_i \sum_j \frac{p(i,j)}{1+|i-j|}$$

$$moment = \sum_i \sum_j p(i,j) i^a j^b$$

$$shade = \sum_i \sum_j p(i,j)(i+j-\mu_i-\mu_j)^3$$

$$prominence = \sum_i \sum_j p(i,j)(i+j-\mu_i-\mu_j)^4$$

In addition to these features we have introduced a new parameter

$$log-contrast = \sum_i \sum_j p(i,j)(|i-j|) \log(|i-j|+1)$$

This feature has almost the same properties as Contrast, but it is less affected by noise.

5 Evaluation of the features

The generalized Mahalanobis distance $M_{i,j}$ was estimated for all features and seabed types. In Table 2 we have listed $M_{1,2}$, $M_{2,3}$, $M_{3,4}$ and $M_{4,5}$ for some of the features. Only the most interesting of the features described in the previous sections are listed in the table.

Feature	$M_{1,2}$	$M_{2,3}$	$M_{3,4}$	$M_{4,5}$
Mean	15.4	40.8	5.2	4.3
$Q_{0.1}$	13.2	37.6	3.5	3.2
$Q_{0.5}$	13.5	40.6	6.7	4.5
$Q_{0.7}$	13.9	34.9	7.3	4.9
$Q_{0.8}$	15.1	32.8	7.8	5.2
$Q_{0.9}$	15.7	27.6	6.5	5.3
sd.dev	3.3	6.1	1.4	0.7
3. order mom	2.2	5.6	3.0	3.2
4. order mom	3.1	6.4	3.1	3.4
5. order mom	2.8	6.6	3.1	3.2
PaceN(10) D_{f_1}	1.2	1.0	1.2	3.2
PaceN(10) D_{f_2}	0.7	1.5	1.3	3.7
PaceN(10) D_{f_3}	0.8	1.6	1.2	3.5
PaceC(10) D_{f_1}	6.6	6.9	2.3	3.1
PaceC(10) D_{f_2}	6.1	6.7	2.6	3.3
PaceC(10) D_{f_3}	7.7	6.5	2.4	2.9
PaceM(5) D_{f_1}	6.7	5.5	2.5	3.8
PaceM(5) D_{f_2}	3.3	4.3	3.5	4.1
PaceM(5) D_{f_3}	5.2	5.6	2.1	3.2
PaceM(10) D_{f_1}	8.4	8.2	2.2	3.7
PaceM(10) D_{f_2}	8.2	7.6	2.4	3.8
PaceM(10) D_{f_3}	7.9	6.6	1.7	2.9
GLCM energy	3.1	3.9	1.3	0.8
GLCM cont	4.5	5.7	1.9	2.5
GLCM corr	2.9	1.2	2.4	1.6

Table 2: The generalized Mahalanobis distance $M_R i,j$ for some of the features calculated for 8000 samples. PaceX(n) means Pace feature calculated after (N) no filtering, (C) after crimmins filter, (M) after median filter, with BA-frequency= n

All features based on backscatter strength are well suited to discriminate the five seabed types. It is difficult to select one of them as the best. Nevertheless we select the $Q_{0.8}$ quantile as the best feature. It discriminates all five classes with less than 2.5 percent overlap. Another interesting feature is the PaceM(5) D_{f_2} feature (Pace feature calculated after filtering with a median filter, BA frequencey= 5). The PaceM(5) D_{f_2} feature was selected in preference to the other Pace features because of the ability to discriminate seabed type 4 and seabed type 5. Among the texture

features, the GLCM cont seems to be the best feature.

6 Correlation between features

In the seabed classification we will used two or three features. It is important that these features are uncorrelated. Some of the features are highly correlated. This is obvious since for example the mean value and the quantile features all are measures of the backscatter strength in the signal. A more interesting question is which correlation we will find between the quantile features, the Pace features, and the GLCM features.

In the computation of the correlations we have used Spearman's rank correlation[12]. In this test each number's rank is used rather than the value itself. The correlation between a data set X_i and a data set Y_i is given by:

$$r_{sp} = \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{n(n^2-1)/12}$$

where R_i is the rank of X_i and S_i is the rank of Y_i. The result r_{sp} is limited to $-1 \leq r_{sp} \leq 1$. r_{sp} near 1 indicates a tendency for the larger values of X to be associated with the larger values of Y. Values near -1 indicate the opposite relationship. The association need not be linear, only an increasing/decreasing relationship is required. All combinations of pairs of features were tested. If the correlation is calculated for data containing different seabed types we may obtain different correlation than for data covering the same seabed type. For this reason, we have divided the total data set of over 4400 samples into 400 groups of 11 samples for which the correlation was calculated. The mean of the 400 r_{sp}'s is presented in Table 3. From this table we read that the Pace-5-D_{F_2} feature is the Pace feature less correlated with the 0.8 quantile feature. The result may also indicate that the GLCM-contrast or Sd.dev feature may be used as a third parameter.

Feature	$Q_{0.8}$	Pace	GLCM
mean	0.9004	0.2879	-0.5305
0.5 quantile	0.8674	0.3032	-0.5325
0.7 quantile	0.8958	0.3154	-0.5016
0.8 quantile	-	0.3187	-0.4799
0.9 quantile	0.8665	0.3224	-0.4334
sd.dev	-0.3222	0.0074	0.5538
3. order mom	-0.3903	-0.3324	0.2529
4. order mom	-0.4847	-0.2498	0.2463
5. order mom	-0.4874	-0.3595	0.2608
PaceM(5) D_{f_1}	0.3910	0.8472	-0.4079
PaceM(5) D_{f_2}	0.3187	-	-0.3297
PaceM(5) D_{f_3}	0.3718	0.7660	-0.4001
PaceM(10) D_{f_1}	0.4224	0.6422	-0.4644
PaceM(10) D_{f_2}	0.4190	0.7604	-0.4643
PaceM(10) D_{f_3}	0.4088	0.6300	-0.4365
GLCM energy	0.1796	-0.0660	-0.3486
GLCM cont	-0.4799	-0.3297	-
GLCM corr	-0.0937	0.2546	-0.0520

Table 3: *The correlation between quant-0.8, PaceM(5) D_{f_2}, GLCM contrast and the other features.*

7 Conclusions

From the examination of the features we know that at least 18 of the features examined may be used in seabed discrimination (Table 3). A lot of these features are strongly correlated. Among the backscatter strength features we found the 0.8 quantile and the mean to be the most interesting. Among the features based on probability density the sd.dev and 3.order moments seem interesting. Among the spectral features PaceM(5) D_{f_2} is the best feature and among the texture features only GLCM contrast and GLCM correlation seems to be useful. If we should select 3 features for seabed classification we would select 0.8 quantile as the first feature, the PaceM(5)-D_{f_2} as the second feature, and either sd.dev, 3.order moment, or GLCM contrast as a third feature. Besides the fact that their discrimination ability are good, and they are not highly correlated features, they may also represent different physical properties from the seabed. If this is true, we may get quantitative measurements of the physical features rather than merely classifying the seabed to classes like rock, sand, silt, clay and mud.

Acknowledgements

We are grateful to Trym Eggen, Simrad Subsea A/S and to Geoffrey Shippey, Chalmers University, Göteborg for interesting and fruitful discussions.

This work has been supported by the Royal Norwegian Council for Scientific (NTNF) and Industrial Research and Nordic Industry Fund. The project is a part of the NTNF research program for Image Analysis and Pattern Recognition. It is also a subproject under the ESMAC (Environmental Seabed Mapping and Characterization) program.

References

[1] N.G.Pace and H.Gao, *Swathe seabed classification*, IEEE Journal of oceanic engineering, Vol 13, No 2, pp 83-90, April 1988.

[2] M.F. Czarnecki. An application of pattern recognition techniques to side scan sonar data. Oceans '79, San Diego 1979, p 112-119.

[3] N. Pace and C. Dyer. Machine Classification of Sedimentary Sea Bottoms. IEEE Transactions on geoscience electronics, Vol. GE-17, No 3, July 1979.

[4] R.B.Huseby, Milvang, Solberg, Weisteen. Seabed classification from backscatter sonar data using statistical methods. Acoustic Classification and Mapping af the Seabed, Bath 1993.

[5] T.Kavli, Carlin, Madsen. Seabed classification from backscatter sonar data using neural network. Acoustic Classification and Mapping af the Seabed, Bath 1993.

[6] N. L. Hjort. "Notes on the theory of statistical symbol recognition.", Technical Report 778, Norwegian Computing Center, 1986

[7] P.J.Bickel and K.A.Doksum, *Mathematical Statistics, Basic Ideas and Selected Topics*, Holden-Day, 1977.

[8] T.R.Crimmins, *Geometric filter for speckle reduction*, Applied Optics, Vol 24, No 10, pp 1438-1442, May 1985 .

[9] R. W. Conners et al. "Segmentation of a High-Resolution Urban Scene Using Texture Operators", Computer Vision, Graphics and Image Processing, Vol. 25, pp. 273-310, 1984.

[10] C. C. Gotlieb et al. "Texture Descriptors Based on Co-occurrence Matrices" Computer Vision, Graphics and Image Processing, Vol. 51, pp. 70-86, 1990.

[11] R.M.Haralick, K.Shanmugan and I.Dinstein, *Textural Features for Image Classification*, IEEE Trans. Systems, Man and Cybernetics, SMC-3, pp 610-621, 1973.

[12] G.K. Bhattacharyya and R.A. Johnson, *Statistical Concepts and Methods*, John Wiley & Sons, New York, 1977.