Application of Gaussian Mixture Model in Identification of Oil Spill on Sea

Jin Weiwei & Zhao Yupeng & An Wei & Li Jianwei

China Offshore Environmental Service Ltd., Tanggu, Tianjin, 300452 Pei Jianxin

Ocean University of China, Qingdao, Shandong, 266100

Keywords: oil spill identification, Gaussian mixture model, image segmentation and unsupervised clustering

Abstract. Aiming at the phenomena of evaporating, emitting, dripping or leaking of offshore oil platform, optical image acquisition device is used to carry out continuous unattended monitoring and effective oil spill identification algorithm is utilized to monitor and identify offshore oil spill. This paper focuses on exploring the study on application of Gaussian mixture model in segmentation and recognition of offshore oil spill image, describes specific algorithm and build an offshore oil spill model by using the expectation-maximization (EM) algorithm and minimum description length (MDL) principle of Gaussian mixture model and combines sequential maximum a posteriori (SMAP) algorithm to segment and identify oil spill image. The research result shows that this method can be used to effectively acquire oil spill information and effectively segment and identify oil spill image.

1. Introduction

With the continuous expansion of offshore oil and gas development scale, marine environments under which such development is carried out are complicated and diversified; in particular, many offshore oil spill accidents occur in recent years, seriously destroying the marine ecological environment. In order to deal with pollution caused by offshore oil spill, countries in the world specially developed countries input a lot of money to build oil spill monitoring system. In the world, more and more countries mainly utilize spectral image monitoring technology. Countries like Norway, Canada, United States, the Netherlands and Germany mount spectral information collection equipment such as visible light camera, video camera, forward looking infrared (FLIR)/CCD camera, infrared/ultraviolet scanner, infrared camera and downward vision camera on propeller-driven aircraft and unmanned aerial vehicle to routinely monitor offshore oil spill. It can be seen that using spectral image monitoring technology to collect offshore oil spill information has become an important means to monitor offshore oil spill and oil spill identification algorithm based on image segmentation and identification technology is an important technical means for processing oil spill image.

Offshore oil spill is characterized by complicated distribution pattern and instability and the surface smoothness of the same type of oil spilled varies due to emulsification. Therefore, the spectral signatures, textural features, geometrical characteristics and the like of offshore oil spill can be effectively utilized to carry out segmentation of oil spill image and target identification. The paper will build a Gaussian mixture model (GMM) for offshore oil spill, which achieves good application result [2] in the segmentation of multi-texture and multi-spectral image, estimate [3,4] the number of class clusters of Gaussian mixture model based on offshore oil spill sample library by adopting minimum description length (MDL) principle and utilize expectation-maximization (EM) [5] algorithm to carry out parameter estimation of each sub class of Gaussian mixture model and build a Gaussian mixture model (GMM) for offshore oil spill. An oil spill image is segmented and identified based on Gaussian mixture model by using sequential maximum a posteriori (SMAP) algorithm [6].

2. Cluster segmentation algorithm based on Gaussian mixture model

Image segmentation algorithm based on Gaussian mixture model is an unsupervised clustering algorithm which can realize clustering through establishing decision making rule according to the statistical characteristic of classified samples without the prior knowledge of cluster. A sample set without cluster mark is divided into several clusters according to a certain rule to maximize the similarity of samples in the same cluster and minimize the similarity of samples in different samples, so as to realize the effective aggregation of data sets. This is consistent with the image segmentation method of human visual system. Therefore, clustering algorithm is more and more widely applied [1] in image segmentation in recent years.

2.1 Gaussian mixture model

Gaussian mixture model is a model which can accurately quantify things by using Gaussian probability-density functions and break things into several Gaussian probability-density functions. In most cases, it is impossible to use a certain Gaussian model to realize accurate quantification and the linear combination of several peak Gaussian distributions is often used to build a model. Any curve can be well fitted when the number of Gaussian distributions is high enough. In this case, Gaussian mixture model is a clustering algorithm aided model. Formula (1) shows one-dimensional Gaussian probability-density function, where, mathematical expectation μ determines position and standard deviation σ determines distribution range. Formula (1) represents Gaussian mixture density function composed of three Gaussian density functions, where, a_1, a_2 and a_3 represent the weighting coefficient of these Gaussian density functions. Figure 1 shows a Gaussian mixture curve obtained through the linear combination of three different Gaussian curves described by μ and σ.

$$g(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{(2\pi)}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Where μ represents mathematical expectation and σ represents standard deviation.

$$p(x)=a_1 g(x;\mu_1,\sum_1)+a_2 g(x;\mu_2,\sum_2)+a_2 g(x;\mu_2,\sum_2) \quad (2)$$

Where $a_1 + a_2 + a_3=1$.

2.2 Minimum description length (MDL)

MDL principle was put forward by Rissenen in 1997 and requires the selection of model with minimum total description length. Total description length is considered to be the sum of description length of network structure and description length of sample data set under given structure. Rissanen proposes a new approximate expression for estimation method based on some assumptions:

$$MDL(K,\theta)=-\log p_y(y|k,\theta)+\frac{1}{2}L\log(MN) \quad (3)$$

Where $L=K(1+M+\frac{M(M+1)}{2})-1$

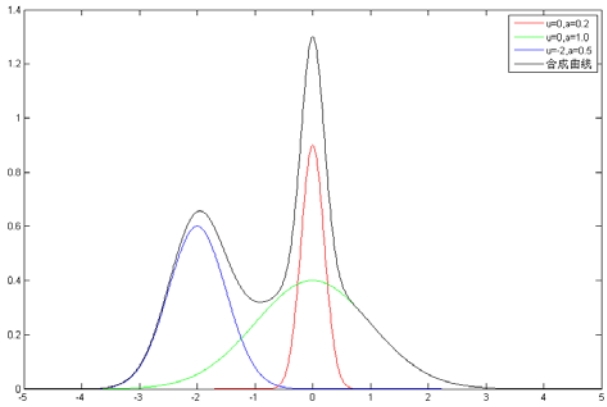

Figure 1 Gaussian Mixture Curve

2.3 EM

EM algorithm is a method proposed by Dempster, Laind and Rubin in 1977 and used for obtaining the maximum likelihood estimation of parameters. This algorithm can be widely used to process so called incomplete data such as imperfect data, censored data and dreaded data. During data processing by EM algorithm, we assume that parameter set θ has been initialized to be $\theta^{(i)}$. In order to enhance MDL principle, EM algorithm will result in an iterative processing process. The probability that pixel y_n belongs to class cluster k can be represented as follows according to Bayes Rule:

$$p_{x_n|y_n}(k|y_n,\theta^{(i)}) = \frac{p_{x_n|y_n}(y_n|k,\theta^{(i)})\pi_k}{\sum_{l=1}^{K} p_{y_n|x_n}(y_n|l,\theta^{(i)})\pi_l} \qquad (3)$$

2.4 Sequential maximum a posteriori (SMAP)

Bayesian image segmentation method often combines maximum a posteriori (MAP) and Markov Random Field (MRF). This method achieves good result but also has come disadvantages. In practical application, it is difficult to accurately compute MAP estimate and the computation of approximate MAP estimate may result in considerable loss. Charles A. Bouman and Michael Shapiro put forward a new method, i.e. SMAP (Sequential Maximum a Posterriori) in Reference [7] and recommended the use of Bayesian image segmentation to eliminate the disadvantages of MAP.

It is assumed that we wants to segment Image Y, we must accurately estimate the pixel label of each pixel point X. Based on Bayesian estimation technique, we assume that prior probability p(x) is known. Generally, Bayesian estimation follows the rule of minimum average segmentation error cost to carry out image segmentation and can be described by the following formula.

$$\hat{x} = \arg\min_{x} E\left[C(X,x)|Y=y\right] \qquad (5)$$

Where C(X, x) represents cost function dividing x into cluster X. The selection of function C is important since this function determines the relevant importance of classification error.

The use of MAP estimation to solve the problem of Formula (3) can be expressed by the formula below:

$$C_{MAP}(X,x) = 1-\delta(X-x) \qquad (6)$$

It can be seen that in Formula (6), $\delta(X-x)$ is equal to 1 when X=x and equal to 0 under other circumstances. When $C_{MAP}(X,x)=1$, pixel will be labeled in a wrong way and MAP estimation maximizes the probability that all pixels are correctly classified. The estimation principle of MAP algorithm is too strict [8,9].

When MAP is used to estimate MSRF, Formula (4) may have the following form:

$$C_{MAP}(X,x) = 1-\delta(X-x) = 1-\prod_{n=0}^{L}\delta(X^{(n)}-x^{(n)}) \qquad (7)$$

It can be seen that in the Formula (5), when classification error occurs under any grey level n, the value of MAP cost function is equal to 1. This kind of cost distribution is irrational. Ideally, a rational cost function should gradually increase with the expansion of pixel area impacted by classification error. In order to achieve this goal, a new cost function is put forward in the reference.

$$C_{SMAP}(X,x) = \frac{1}{2} + \sum_{n=0}^{L} 2^{(n-1)} C_n(X,x) \qquad (8)$$

Where $C_n(X,x) = 1-\prod_{i=n}\delta(X^{(i)}-x^{(i)})$

It is assumed that K is a constant and $X^{(K)} \neq x^{(K)}$, but $X^{(i)} = x^{(i)}, \forall i > K$. Therefore, C_n can be simplified as follows:

$$C_n(X,x) = \begin{cases} 1 \ \ if \ \ n \le K \\ 0 \ \ if \ \ n > K \end{cases} \qquad (9)$$

Formula (5) can be transformed into the following formula:

$$\begin{aligned}\hat{x} &= \arg\min_x E[C_{SMAP}(X,x)|Y=y] \\ &= \arg\min_x \sum_{n=0}^{L} 2^{n-1}\left\{1-P(X^{(i)}=x^{(i)}i\geq n|Y=y)\right\} \\ &= \arg\min_x \sum_{n=0}^{L} 2^{n} P(X^{(i)}=x^{(i)}i\geq n|Y=y)\end{aligned} \tag{10}$$

Compared with MAP and MPM algorithms, SMAP classifier has many advantages. The former algorithms requires considerable computing loss to obtain the optimal value in an iterative manner while SMAP algorithm can directly obtain the value without iteration. In addition, MAP algorithm only can be used to obtain minimum error probability of each pixel point without considering spatial distribution error. However, SMAP algorithm tries to minimize spatial distribution error, so it can produce a satisfactory subjective classification result.

3. ALGORITHM IMPLEMENTATION AND VALIDATION

Through the above analysis, we can know that the automatic offshore oil spill identification accuracy is substantially determined by the building of oil spill model. Here we use Gaussian mixture model (GMM) to build an offshore oil spill model. In order to validate the GMM parameter calculation accurateness of the algorithm we use, we will use a known GMM to generate a data set, utilize this data set to calculate the parameter of GMM in an inverse manner and compare the calculated result with known parameters to validate the accurateness of the algorithm.

First of all, we have known a Gaussian mixture model composed of three Gaussian models and each Gaussian model is described by three parameters: relative proportion π_i, average value μ_i and covariance matrix R_i. Then we use this GMM to generate two-dimensional random vectors containing 200 samples and the data distribution diagram is shown in Figure 2. The sample data distribution diagram shows that there are only two clusters clearly classified and two clusters overlap with each other and cannot be clearly classified. Now we use MDL and EM algorithms described above to calculate the parameters of GMM. Some outputs of MDL algorithm are shown in Figure 3. It can be seen from the figure that when subclass=3, MDL algorithm obtains the minimum value, indicating that the optimal number of clusters of the data set is 3. This result is consistent with the class number of the model generating the data, indicating that MDL algorithm can achieve a good result even under the condition of data overlapping. Then EM algorithm is used to calculate the parameters of GMM and the result is shown in Table 1. It can be seen from the table that maximum likelihood estimation result is quite approximate to true value.

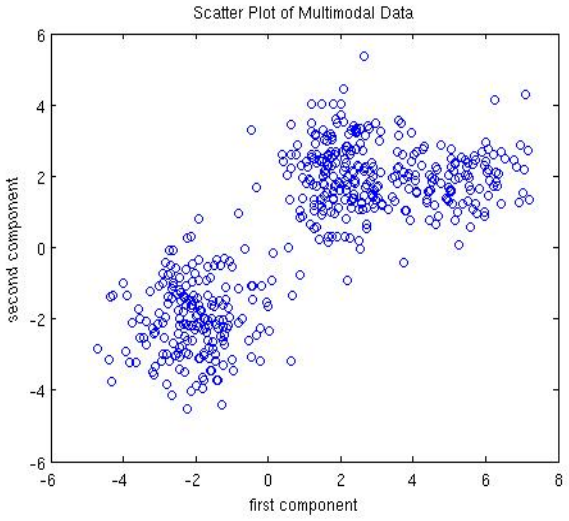

Figure 2 Gaussian Mixture Sample Data Distribution Diagram

Subclasses = 6; Rissanen = 1984.862735; Combining Subclasses (3,4)
Subclasses = 5; Rissanen = 1972.446690; Combining Subclasses (1,3)
Subclasses = 4; Rissanen = 1956.088972; Combining Subclasses (0,3)
Subclasses = 3; Rissanen = 1939.444535; Combining Subclasses (0,2)
Subclasses = 2; Rissanen = 1949.218131; Combining Subclasses (0,1)
Subclasses = 1; Rissanen = 2130.086314;

Figure 3 Part of Cluster Number Estimation Process Based on MDL Principle

Table 1 GMM Parameters Obtained by Using EM Algorithm

	Parameter	True value	Estimated value
Cluster1	π_1	0.4	0.385141
	μ_1	[2.0 2.0]	[1.968770 1.908531]
	R_1	$\begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.089314 & 0.291038 \\ 0.291036 & 1.034361 \end{bmatrix}$
Cluster2	π_2	0.4	0.433581
	μ_2	[-2.0 -2.0]	[-2.096070 -1.960741]
	R_2	$\begin{bmatrix} 1 & -0.1 \\ -0.1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1.089534 & -0.091958 \\ -0.091958 & 0.928903 \end{bmatrix}$
Cluster3	π_3	0.2	0.181277
	μ_3	[5.5 2.0]	[5.333640 1.904941]
	R_3	$\begin{bmatrix} 1 & 0.2 \\ 0.2 & 0.5 \end{bmatrix}$	$\begin{bmatrix} 0.870289 & 0.188575 \\ 0.188515 & 0.578056 \end{bmatrix}$

4. IMAGE SEGMENTATION RESULT

Through the above algorithm process, it can be seen that MDL and EM algorithms can be effectively used to calculate the number and parameters of GMM. In the following text, we will use GMM to build an offshore oil spill model and use SMAP algorithm for segmentation. Different targets in the image correspond to different GMM. Image information is transformed into data information in the format of ASCII; maximum a posteriori estimation analysis is carried out to data corresponding to each pixel point; the probability that a pixel belongs to a GMM is calculated and segmentation is conducted according to the probability value to effectively segment image targets. In order to facilitate subsequent processing, segmented pixel value (color) selected should be consistent with the sequential number of cluster and segmented image should be shown in the form of color. Two-frame offshore oil spill image after being processed by the above method is shown below:

Figure 4 Offshore Oil Spill Image and Corresponding Segmented Image

5. CONCLUSIONS

This paper uses EM algorithm and MDL principle of GMM to extract targets in offshore oil spill image. The image segmentation result shows that image segmentation technology using unsupervised clustering algorithm based on rigorous mathematical theory and GMM can segment targets on the basis of collecting global and local information of an image; this technology has strong advantages in terms of segmentation performance improvement and anti-noise capacity and can effectively segment offshore oil spill image without man-machine interaction, thereby laying the foundation for the subsequent processing of oil spill image.

REFERENCES

[1]. C.A. Bouman & M. Shapiro. 1994. A multiscale random field model for bayesian image segmentation. Image Processing, IEEE Transactions on, 3(2):162–177.

[2]. G.J. McLachlan, K.Krishnan & S.K. NG. 2009. The em algorithm. The Top-Ten Algorithms in Data Mining, X. Wu and V. Kumar (Eds.). Boca Raton, Florida: Chapman and Hall/CRC, 93–115.

[3]. H. Permuter, J. Francos & I. Jermyn. 2006. A study of gaussian mixture models of color and texture features for image classification and segmentation. Pattern Recognition, 39(4): 695–706.

[4]. J. Marroquin, S. Mitter & T. Poggio. 1987. Probabilistic solution of ill-posed problems in computational vision. Journal of the American Statistical Association, 82(397):76–89.

[5]. P.D. Grnwald. 2007. The minimum description length principle. The MIT Press.

[6]. RC Dubes, AK Jain, SG Nadabar & CC Chen. 1990. Mrf model-based algorithms for image segmentation. In Pattern Recognition, Proceedings. IEEE 10th International Conference on, volume1, 808–814.

[7]. V. Kumar, J. Heikkonen, J. Rissanen & K. Kaski. 2006. Minimum description length denoising with histogram models. Signal Processing. IEEE Transactions on, 54(8):2922–2928.

[8]. W.M. Campbell, DE Sturim & DA Reynolds. 2006. Support vector machines using gmm supervectors for speaker verification. Signal Processing Letters, IEEE, 13(5):308–311.

[9]. Xiang Rihua & Wang Runsheng. 2003. A Range Image Segmentation Algorithm Based on Gaussian Mixture Model. Journal of Software, 14(7): 1250-1257.