A Comparative Study of The Performance of Population Mixture Models for Image Thresholding

Sho Kikkawa[1], Kazumi Murata[2], Tetsuya Masumoto[1],
Susumu Ekawa[1], and Hisashi Yoshida[1]

Abstract

There have been two major mixture models used for image thresholding, the Gaussian mixture (G-G model) and the Poisson mixture (P-P model). In this report, we quantitatively compare the performance of three mixtures, the G-G model, the P-P model, and a new additional Poisson and Gaussian mixture (P-G model) proposed here. The minimum description length (MDL) is used to assess the performance of the models. The images used here are 76 infrared images from National Oceanic and Atmospheric Administration (NOAA) satellites and 62 images from a visible channel of the Japanese Geostationary Meteorological Satellite (GMS, Himawari). It was found that the G-G model is generally the most excellent and the P-G model is the next, whereas the P-P model may not be applicable to the 256-grey level images. However, the G-G model is not always so good when the EM algorithm is used to estimate the mixtures. It is because the algorithm is very sensitive to the choice of the initial parameters values.

1. Introduction

Thresholding is an important approach to segment gray level images[1, 2, 3, 4, 5] . In the global thresholding methods based on the gray level histogram, it is usually assumed that the histogram is obtained from a mixture population[6] . These methods typically use a Gaussian mixture (G-G model)[7, 8, 9, 10] .

On the other hand, criticizing the G-G model for not having any justification, Pal et al. modeled the image histogram as a mixture of two Poisson distributions (P-P model) based on the image formation theory and developed several thresholding methods[11, 12, 6] . They compared the performance of their P-P model-based methods with G-G model-based methods and judged that their methods are better than the others[11, 12] . Their judgment, however, was given by inspecting with their eyes whether the images were suitably segmented or not, and by using several real 32-grey levels images.

It is necessary to evaluate histogram models more quantitatively and by using much more real images with more gray levels, e.g. 256 levels images which we come across much more often than 32-gray-levels images. To our knowledge, however, there have not been such studies. One of the main reasons may be that the true threshold of the histogram of an actual image is unknown in most cases.

Is there any quantitative way to evaluate the performance of the mixture model without precise knowledge of the true threshold? It must be reasonable to think that we will get better segmentation when we find a better estimation of the population mixture. If so, we can assess the performance of mixture models for image segmentation quantitatively by measuring how closely the histogram fits the estimated distribution of the model.

In this report, we quantitatively compare the performance of the three models, that is, the G-G model, the P-P model and a new additional Poisson-Gaussian mixture (P-G model) proposed here, using the MDL as a measure how well mixture models fit the histogram. The images used here are 76 NOAA infrared images and 62 images from a GMS visible channel of which the number of gray levels is 256.

A problem of the P-G model is that it is a mixture of a discrete distribution and a continuous one. However, the problem is not so crucial, because the Gaussian distribution is always discretized in the process of finding the best mixture the histogram fits and we don't have to restore it to the continuous one. Further, the mixture model is not for explaining image formation mechanism but just for representing histograms.

Received December 9, 2005.
This work was supported in part by Project Research of the School of Biology-Oriented Science and Technology, Kinki University (No.03-IV-4)
[1]Department of Electronic Systems and Information Engineering, Kinki University, Wakayama 649-6493, JAPAN
[2]Osaka Development Center of Digital Vision Solution Co., Ltd., Osaka, 550-0005, JAPAN

We note that there are reports in which other mixtures, the gamma mixture[13] and the binomial mixture[14], are used for thresholding. However, they are omitted from our consideration because it is difficult to decide one of the two parameters of the gamma distributions coherently[13] and a binomial distribution is virtually the same as a Poisson distribution when the number of the pixels of an image is large.

2. Population mixture models

As shown in Figure 1, not a few NOAA or GMS image histograms are positively skewed in the lower part of gray levels, and symmetrical in the upper part. The lower part does not always approximate a Gaussian distribution and the upper part is not so simple as a one-parameter Poisson distribution. Therefore, we may need a heterogeneous mixture model which consists of different kinds of components for expressing the histograms. However, any heterogeneous mixture model has not been proposed. Therefore, we add the heterogeneous P-G model to the existing G-G and P-P models and compare their performance.

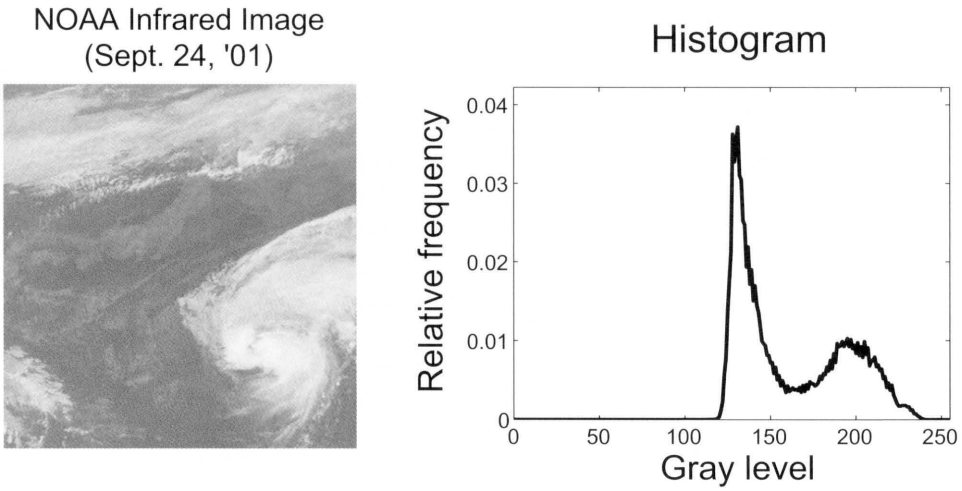

Figure 1: A NOAA infrared image and its gray level histogram.

Letting the components of a mixture be $p_i(x|\boldsymbol{\theta}_i)$ $(i = 1, 2)$ where $\sum_x p_i(x|\boldsymbol{\theta}_i) = 1$, the mixture is written as

$$p(x) = c_1 p_1(x|\boldsymbol{\theta}_1) + c_2 p_2(x|\boldsymbol{\theta}_2), \tag{1}$$

where $c_1, c_2 \geq 0$, $c_1 + c_2 = 1$ and $\boldsymbol{\theta}_i$ denotes the parameter vector of p_i. The components of the G-G model are given by

$$p_i(x|\boldsymbol{\theta}_i) = \exp[-(x - \mu_i)^2/(2\sigma_i^2)]/(\sqrt{2\pi}\sigma_i), \tag{2}$$

where $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)^t$ $(i = 1, 2)$, those of the P-G model are

$$p_1(x|\theta_1) = \nu^x \exp[-\nu]/x!, \tag{3}$$

$$p_2(x|\boldsymbol{\theta}_2) = \exp[-(x - \mu)^2/(2\sigma^2)]/(\sqrt{2\pi}\sigma), \tag{4}$$

where $\theta_1 = \nu$, $\boldsymbol{\theta}_2 = (\mu, \sigma)^t$, and those of the P-P model are

$$p_i(x|\theta_i) = \nu_i^x \exp[-\nu_i]/x!, \tag{5}$$

where $\theta_i = \nu_i$, $(i = 1, 2)$.

3. Methods

We use two methods, i.e. the round robin method (RB) and the EM algorithm (EM), to obtain the optimum estimation of a mixture. Then we use the MDL to assess the performance of the models.

3.1 Round robin method

We call the next method the RB method. First, we define a divergence $D(h||p)$ between a histogram $h(x)$ and the mixture distribution $p(x)$ of a model by

$$D(h||p) = \sum_x h(x) \log \frac{h(x)}{p(x)} \tag{6}$$

and assume that the approximation below holds

$$\begin{aligned} D_a(h||p) \approx & \sum_{x \leq T} h(x) \log \frac{h(x)}{c_1 p_1(x|\boldsymbol{\theta}_1)} \\ & + \sum_{x > T} h(x) \log \frac{h(x)}{c_2 p_2(x|\boldsymbol{\theta}_2)}, \end{aligned} \tag{7}$$

where $0 \leq T \leq 255$ is a threshold. Next, we get the optimum parameters c_i and $\boldsymbol{\theta}_i$ by solving the simultaneous equations $\partial D_a/\partial c_i = 0, \partial D_a/\partial \boldsymbol{\theta}_i = 0$ so that the resultant $p(x)$ gives the minimum D_a. Calculating $D(h||p)$ by substituting the $p(x)$ into Eq. (6), we rewrite the $D(h||p)$ as $D(T)$. We repeat the procedures for all the possible values of $T(= 0, 1, \cdots, 255)$, and find $\hat{T} = \arg\min_T D(T)$. The resultant $p(x)$ is the optimum estimation of the mixture. We denote it as $\hat{p}(x)$.

In the RB method, the overlap between the components is ignored by assuming that the truncation errors are neglected. In other words, it is assumed that all the pixels in one cluster divided by a threshold T are produced by one component distribution and those in the other cluster are produced by the other component. Therefore, the estimation error gets larger as the overlap increases.

3.2 EM algorithm

The EM algorithm used for estimating the components associated with a mixture[15, 16] is useful for image segmentation[17, 18, 19] . Although the EM method is powerful even when the components of a mixture overlap each other, it is very sensitive to the choice of the initial values of parameters and tends to fall into a local minimum. Therefore, the estimation error can be fairly large when it falls into a local solution.

3.3 MDL for assessing the performance of the models

The maximum likelihood principle used in the EM algorithm is equivalent to the minimum divergence principle used in the RB method. Therefore, it is appropriate to use the divergence $D(h||\hat{p})$ between the histogram $h(x)$ and the estimated mixture $\hat{p}(x)$ of a model to assess the performance of the model. At the same time, we should consider the number, k, of parameters to assess the performance because, in general, the more parameters the model has, the better performance it will have. Hence, we assess the performance of the models with Rissanen's MDL[20, 21] which consists of the divergence and the terms including k:

$$MDL \approx D(h||\hat{p}) + \frac{k \log N}{2N} + \frac{k}{2N}, \tag{8}$$

where N is the number of data, i.e. the number of the pixels.

4. Experiments

We here compare the performance of the three models using 76 NOAA images and 62 GMS images. Considering its sensitivity to the initial parameter values, when we use the EM methods, we repeat the EM procedure 10 times per image changing the initial values at random.

4.1 With or without 0-cutting operation

Like as an example shown in Figure 1, there are fairly many histograms of NOAA and GMS images whose values below a certain gray level vanish. The performance of the models might be affected by cutting the region of 0's (0-cutting operation). Therefore, we assess the performance of the models both for the original histogram and for the histogram after the 0-cutting operation (referred to as the 0-cut histogram).

Figure 2 shows, as an example, the optimum mixture obtained by RB method by using the original histogram (in column (a)) and by using the 0-cut histogram (in column (b)) of the image used in Figure 1.

In both columns, from the top, the results of the G-G model, of the P-G model, and of the P-P model are shown. The solid lines show the optimum mixture and the dotted lines the histograms. The MDL value is shown in each graph.

In this figure, the performance of the G-G model is the best, that of the P-G model is the next and that of the P-P model is the last for the original histogram, whereas the G-G model gives the place to the P-G model for the 0-cut histogram.

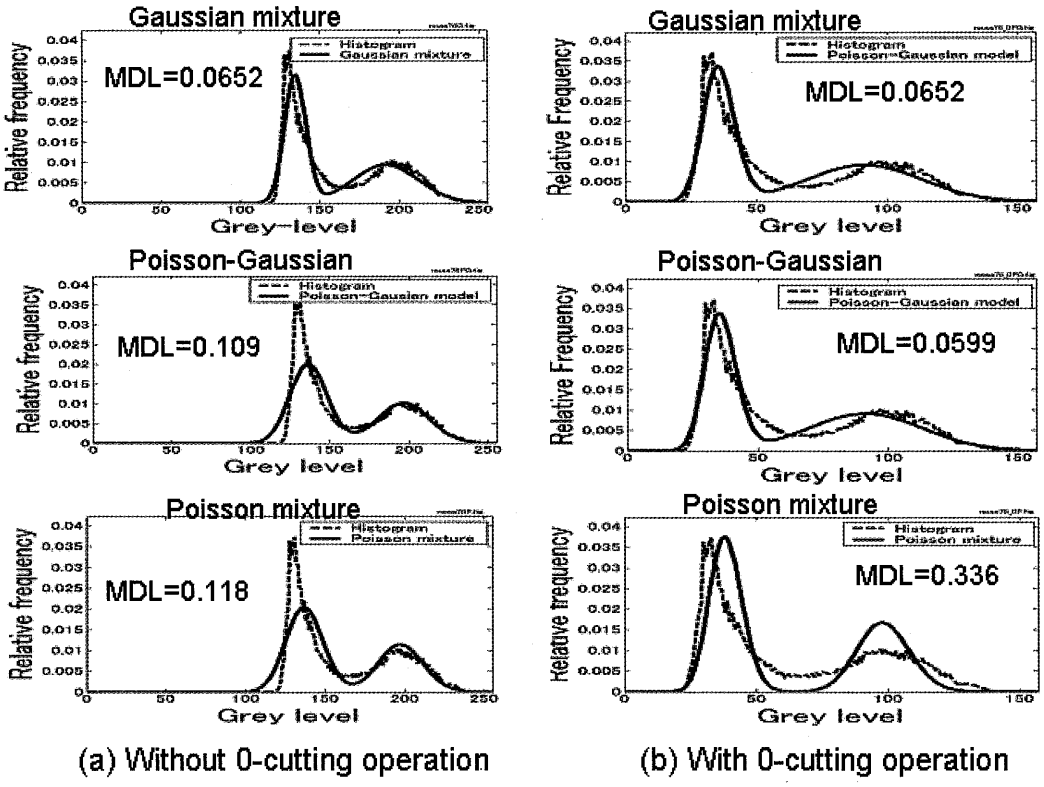

Figure 2: The estimated mixture from the image histogram in Figure 1

Hereafter, we add a suffix '1' to RB and EM as RB_1 and EM_1 when the methods are used without 0-cutting operation and '2' as RB_2 and EM_2 when they are used with 0-cutting operation.

4.2 Comparison of the models

We measured the goodness of fit between the histogram and the optimum mixture of a model using the MDL. In each figure, the model having the minimum MDL value was graded A. The number of times that each model got the grade A are shown in Table 1 (when the RB methods were used) and in Table 2 (when the EM methods were used).

In the significance tests conducted in this report, the hypotheses are rejected when the statistic is significant at a level less than .05.

First, we set up the null hypothesis that $p_{GG} = p_{PG} = p_{PP}$, where p_{GG}, p_{PG}, or p_{PP} is the population proportion that the G-G model, the P-G model, or the P-P model gets the grade A, respectively. In every method of RB_1, RB_2, EM_1 and EM_2 in both Table 1 and Table 2, the χ^2 statistic was clearly significant at a significance level less than .001. This means that the performance of the G-G model and the P-G model are, in general, better than that of the P-P model in every method.

Next, we consider the null hypothesis that $p_{GG} = p_{PG}$. In Table 1 and Table 2, the hypothesis was rejected at a level less than .001 in every RB method and EM method. Therefore, we find that the performance of the G-G model is the best in all of the cases except in the EM methods for GMS images, where the P-G model is the best.

Table 1: The performance of the models for the RB methods.

Model	Frequency (%)			
	NOAA images		GMS images	
	RB_1	RB_2	RB_1	RB_2
G-G	58 (76 %)	48 (63 %)	62 (100 %)	57 (92 %)
P-G	16 (21 %)	28 (37 %)	0 (0 %)	5 (8 %)
P-P	2 (3 %)	0 (0 %)	0 (0 %)	0 (0 %)

Table 2: The performance of the models when the EM procedure was repeated 10 times.

Model	Frequency (%)			
	NOAA images		GMS images	
	EM_1	EM_2	EM_1	EM_2
G-G	412 (54 %)	436 (57 %)	228 (37 %)	248 (40 %)
P-G	245 (32 %)	317 (42 %)	392 (63 %)	372 (60 %)
P-P	103 (14 %)	7 (1 %)	0 (0 %)	0 (0 %)

Furthermore, we examined the effects of the 0-cutting operation on the performance of the models. The null hypotheses of homogeneity that $H_{\mathrm{RB}} : p_{\mathrm{GG}}(\mathrm{RB}_1) = p_{\mathrm{GG}}(\mathrm{RB}_2)$, $p_{\mathrm{PG}}(\mathrm{RB}_1) = p_{\mathrm{PG}}(\mathrm{RB}_2)$, $p_{\mathrm{PP}}(\mathrm{RB}_1) = p_{\mathrm{PP}}(\mathrm{RB}_2)$, and $H_{\mathrm{EM}} : p_{\mathrm{GG}}(\mathrm{EM}_1) = p_{\mathrm{GG}}(\mathrm{EM}_2)$, $p_{\mathrm{PG}}(\mathrm{EM}_1) = p_{\mathrm{PG}}(\mathrm{EM}_2)$, $p_{\mathrm{PP}}(\mathrm{EM}_1) = p_{\mathrm{PP}}(\mathrm{EM}_2)$ were set up, where $p_{\star\star}(\bullet\bullet)$ is the population proportion that the model $\star\star$ gets the grade A when the method $\bullet\bullet$ was used. By χ^2 testing, the hypothesis H_{RB} was rejected at a level less than .02 for the NOAA images and was rejected at a level less than .05 for the GMS images. The hypothesis H_{EM} was rejected at a level less than .001 for the NOAA images but was not rejected for the GMS images.

Therefore, we can say that the 0-cutting operation is effective in the RB methods in such a way that the operation relatively lowers the performance of the G-G model and relatively raises the performance of the P-G model when the RB method was used. In the EM methods, on the other hand, the operation is effective for the NOAA images, and raises the performance of both the G-G model and the P-G model, but is not effective for the GMS images.

The inconsistency between the RB methods and the EM methods or within the EM methods may be because the EM algorithm reaches fairly often local solutions. We consider this problem in the next section.

4.3 Reduction of the effects of the local solutions

Here, in order to reduce the effect of the local solutions of the EM algorithm, we repeated the EM procedure changing the initial values of the parameters until the EM algorithm was judged to have reached the global solution. The results are shown in Table 3.

Table 3: The performance of the models when the EM algorithm was judged to have reached the global solution.

Model	Frequency (%)			
	NOAA images		GMS images	
	EM_1	EM_2	EM_1	EM_2
G-G	60 (79 %)	53 (70 %)	49 (79 %)	48 (77 %)
P-G	16 (21 %)	23 (30 %)	13 (21 %)	14 (23 %)
P-P	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)

The same tests in Section 4.2 were conducted. The results here are more clear-cut than those in Table 2. The hypothesis $P_{GG} = P_{PG}$ was rejected at a level less than 0.001 in every EM methods and the effects of the 0-cutting operation was not significant for both NOAA images and GMS images.

5. Summation and Conclusion

In Table 4, we showed the best model for each method and in Table 5, we showed the effects of the 0-cutting operation. In both the tables, EM_B and EM_C mean the EM method used in Section 4.2 and Section 4.3, respectively. In Table 5, signs + and − denote an increase and a decrease in performance of the model, respectively and ** denotes no significant effects of the operation on the model.

Table 4: The best performance models . † denotes $\alpha < .001$.

Image	NOAA images			GMS images		
Method	RB	EM_B	EM_C	RB	EM_B	EM_C
Best model	G-G †	G-G †	G-G †	G-G †	P-G †	G-G †

Table 5: The effects of the 0-cutting operation. †, †† and ††† denote $\alpha < .001, \alpha < .02$, and $\alpha < .05$, respectively.

Image	NOAA images			GMS images		
Method	RB	EM_B	EM_C	RB	EM_B	EM_C
G-G	− ††	+ †	**	− †††	**	**
P-G	+ ††	+ †	**	+ †††	**	**
P-P	**	− †	**	**	**	**

The conclusive remarks are then as follows.

1. Generally speaking, the G-G model has the best performance, the P-G model is the next, and the P-P model is the last for the total 138 meteorological images with 256-grey levels.

2. However, the G-G model gave the place to the P-G model when the EM algorithm was used for the GMS images. It must be due to local solutions.

3. The 0-cutting operation lowers the performance of the G-G model and raises that of the P-G model when the RB method was used. On the other hand, when the EM method was used, the results were complicated. It suggests that there is still a possibility that the EM algorithm falls into a local solution even when we judged the EM algorithm to have reached the global solution. Therefore we need further experiments to say some conclusive remarks about the effect of the operation when the EM method is used.

Acknowledgments

The work of S. Kikkawa was supported in part by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (C)(No.15500191) and School of Biology-Oriented Science and Technology, Kinki University for the Project Research(No.03-IV-4).

The work of H. Yoshida was supported in part by Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research (C)(No.17560381).

References

(1) A. D. Brink and N. E. Pendock. Minimum cross-entropy thereshold selection. *Pattern Recognition*, 29(1):179–188, 1996.

(2) N. Otsu. A threshold selection method from gray-level histogram. *IEEE Trans. Sytems, Man, and Cybernetics*, SMC-9(1):62–66, 1979.

(3) Kapur J. N., P.K.Sahoo, and A.K.C.Wong. A new method for gray level picture thresholding using the entropy of the histogram. *Computing Vision Graphics Image Process.*, 29:273–285, 1985.

(4) M. Fleury, L. Hayat, and A. F. Clark. Parallel entropic auto-thresholding. *Image and Vision Computing*, 41:247–263, 1996.

(5) J. S. Weszka. A survey of threshold selection techniques. *Comput. Graphics Image Process*, 7:259–265, 1978.

(6) Nikhil R. Pal and Sankar K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

(7) Stanley L. Sclove. Application of conditional population-mixture model to image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-5(4):428–433, 1983.

(8) J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern recognition*, 19(1):41–47, 1986.

(9) T. Kurita, N. Otsu, and N. Abdelmalek. Maximum likelyhood thresholding based on population mixture models. *Pattern recognition*, 25(10):1231–1240, 1992.

(10) C. H. Li and C. K. Lee. Minimum cross entropy thresholding. *Pattern Recognition*, 26(4):617–625, 1993.

(11) Nikhil R. Pal and Sankar K. Pal. Image model, Poisson distribution and object extraction. *International Journal of Pattern Recognition and Artificial Intelligence*, 5(3):459–483, 1991.

(12) Nikhil R. Pal. On mimimum cross-entropy thresholding. *Pattern Recognition*, 29(4):575–580, 1996.

(13) Samadani Ramin. A finite mixture algorithm for finding proportions in SAR omages. *IEEE Trans. Image Processing*, 4(8):1182–1186, 1995.

(14) Eiji. Nakamura, Akio. Shio, and Hiroshi. Kaneko. An adaptive thresholding technique based on two-mixture distributions. *IPSJ SIG Technical Report, Computer Vision and Image Media*, 97(70(CVIM-106)):1–8, 1997. (in Japanese).

(15) Richard A. Render and Homer F. Walker. Mixture densities, maximum likelyhood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

(16) Martin A. Tanner. *Tools for statistical inference*. Springer, 1996.

(17) Yair Weiss and Edward H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. *Proc. IEEE Conf. CVPR*, pages 321–326, 1996.

(18) Nir Friedman and Stuart Russell. Image segmentation in video sequences:A probabilistic approach. *Proc. 13th Conference on Uncertainty and Artificial Intelligence*, pages 175–181, 1997.

(19) Zhihua Zhang, Chibiao Chen, Jian Sun, and Kap Luk Chan. EM algorithm for Gaussian mixture with split-and-merge operation. *Pattern Recognition*, 36:1973–1983, 2003.

(20) J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–472, 1978.

(21) J. Rissanen. Stochastic complexty in statistical inquiry. *World Scientific Series in Comp. Sci.*, 15, 1989.

和文抄録

画像２値化のための混合分布モデルの性能比較

吉川昭、 村田和水、 増本哲也、 江川進、 吉田久

画像２値化のための混合分布モデルとしては、一般的にガウス混合分布モデルとポアッソン混合分布モデルの２種類がある。本論文では、このガウス混合分布モデル(G-G モデル) とポアッソン混合分布モデル(P-P モデル) に加えて、新たに提案するポアッソン-ガウス混合分布モデル(P-G) の性能を定量的に比較した。なお、モデルの評価には最小記述長基準(MDL 基準) を用いている。本論文で使用したデータは、気象衛星NOAA から得た赤外画像76 枚と、気象衛星ひまわりから得た可視画像62 枚である。比較の結果、一般にG-G モデルが最も良く、続いてP-G モデルの順であった。他方、256 階調の白黒画像に対して、P-P モデルは適当でないと思われた。しかしながら、分布の推定にEM アルゴリズムを用いる場合は、必ずしもG-G モデルが最良モデルであるとは言えない。なぜなら、EM アルゴリズムは初期値に対して非常にセンシティブなアルゴリズムであり、推定した分布が必ずしも良い結果を与えないためである。