INFORMATION CRITERIA FOR HISTOGRAM THRESHOLDING TECHNIQUES

Pierre Courtellemont [1] *Christian Olivier* [2] *Frédéric Jouzel* [1]

[1] PSI-La3i,
University of Rouen, France
Pierre.Courtellemont@univ-rouen.fr

[2] SIC-IRCOM, UMR CNRS 6615,
University of Poitiers, France
Olivier@sic.univ-poitiers.fr

ABSTRACT

This paper deals with grey-level images histograms. In a first part, we show how possible it is to reduce the number of levels with the minimum of information loss, thanks to information criteria. The same criteria allow to threshold these histograms, giving the optimal number of thresholds.

1 INTRODUCTION

In this paper, we consider multi-level images, for that numerous techniques of optimal thresholds searching have been proposed. Among the most usual methods, let us quote the methods of Otsu (1979) and Kittler and Illingworth (1986) then generalized by Kurita *et al.* (1992). They all appear as Maximum Likelihood (ML) methods, but with different hypothesizes on the parameters of the considered distributions. Iterative algorithms to find the thresholds, avoiding an exhaustive search, have been proposed, but they suppose an *a priori* knowledge on the number of thresholds. We propose the use of Information Criteria (IC) penalizing the log-likelihood function, allowing to obtain the optimal number of thresholds. Different criteria are discussed. For its properties of consistency, the criterion φ^*_β is proposed. The results are illustrated on synthetic images, for different noise levels, and on some real images. Criteria with the same terms of penalty allow also to optimally reduce the number of bins of grey-scale histograms, supplying an image dependent compression and giving also, by another way, the number of thresholds.

This paper is made of three parts. The first one gives a historical review of the several IC suggested for the search of the order of a parametrized model or a set of probability density functions (PDF) and we prove how these criteria can be applied to supply histogram based laws approximation. In a second part, we show the use of IC for the optimal reduction of the number of grey-levels in an image. The third part shows how to use the same IC for the search of the optimal number of thresholds, and their location, in a problem of segmentation based on multi-thresholding.

2 PARAMETRIZED MODELS ORDER ESTIMATION

2.1 Information criteria

Historically, the first IC was obtained in the autoregressive (AR) modeling context [1]. More generally, let us consider a set of PDF $f(.,\lambda)$ associated with a random variable X observed via a sample $X^N = X_1, \ldots, X_N$. Let $\widehat{\theta}_{k,N}$ be an asymptotically normal estimate (when $N \to +\infty$) of the k-dimensional vector λ defined from the sample X^N. An estimation of the expectation of the KL information between $f(.,\lambda)$ and $f(.,\theta_{k,N})$ leads to the Akaike´s criterion [1]:

$$AIC\,(k) = 2k - 2\sum_{i=1}^{N} \log f\left(X_i, \widehat{\theta}_{k,N}\right) \qquad (1)$$

where N is the number of observations available, and $2k$ the penalty term of the criterion. The parameter dimension k is thus estimated by minimizing (1). To avoid the non-consistency of this estimate, other model selection criteria were suggested, where $2k$ is replaced by $c_N k$, a penalty term depending on the number of free parameters in the model. For example, we give some c_N values corresponding to classical IC: $c_N = \log N$ corresponds to BIC (for Bayesian Information Criterion) [12] or MDL (Minimum Description Length) [9], allowing a strongly consistent estimator of k , $c_N = \log\log N$ corresponds to Hannan and Quinn's φ criterion [3]. The later criterion ensures a weak consistency. One can derive other criteria, which appear as combinations of AIC, φ and BIC. From asymptotic approximations of KL information and *stochastic complexity* introduced by J. Rissanen [10], one can obtain the following criteria [2]:

$$AIC^*\,(k) = 2k + k\log N - 2\sum_{i=1}^{N} \log f\left(X_i, \widehat{\theta}_{k,N}\right) \qquad (2)$$

$$\varphi^*_\beta\,(k) = 2k + kN^\beta \log\log N - 2\sum_{i=1}^{N} \log f\left(X_i, \widehat{\theta}_{k,N}\right), \beta \in]0,1[\qquad (3)$$

AIC^* and φ^*_β are strongly consistent order estimators.

2.2 Histogram based law estimation

We are going to show how to approximate a law λ by a histogram, according to a sequence X^N of observations. The number of bins k of this histogram and the width of these bins must be optimal in the sense of a cost approximated of the expectation of the KL information. From an

initial partition M with m classes of the set Ω on which is defined the random variable X of density $f(.,\lambda)$, we search a subset K with k classes of M. Several authors were also interested with the approximation of histograms based upon AIC-type criteria; let us quote J. Rissanen [11] or C. C. Taylor [14]. In this paper, we use the method proposed in [7], that allows to give the number and the bins width. The obtained bins have not the same width, allowing an optimization of the approximation. One obtain the following AIC-type estimator of the number of bins of a histogram from N observations [7]:

$$AIC\,(k) = 2k - 2N \sum_{B\in K} \widehat{\theta}_{k,N}(B) \log \frac{\widehat{\theta}_{k,N}(B)}{\mu(B)} \qquad (4)$$

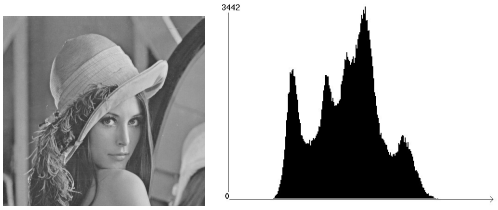
where $\widehat{\theta}_{k,N}(B)$ is the frequency associated with bin B and μ is a *a priori* law. We can also extend the criteria (2) and (3) which ensures the estimator consistency:

$$AIC^*\,(k) = 2k + k\log N - 2N \sum_{B\in K} \widehat{\theta}_{k,N}(B) \log \frac{\widehat{\theta}_{k,N}(B)}{\mu(B)} \qquad (5)$$

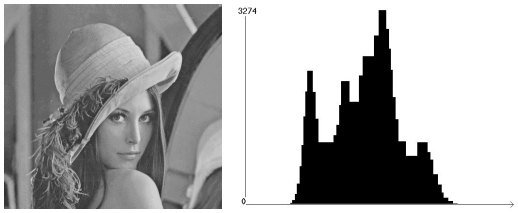
$$\varphi^*_\beta(k) = 2k + kN^\beta \log\log N - 2 \sum_{B\in K} \widehat{\theta}_{k,N}(B) \log \frac{\widehat{\theta}_{k,N}(B)}{\mu(B)} \qquad (6)$$

3 IMAGE APPLICATION

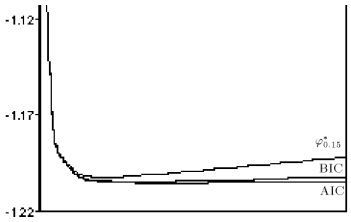
The initial partition M is a partition with the m equal-width bins given by the initial number of grey levels (generally 256). The form of the bins for a subset K' with $k' = (k-1)$ bins is obtained from the subset K with k bins, $k < m$, by a merging process of adjacent bins. At each step, we search the two adjacent bins to merge giving the minimal $IC(k-1)$ value. The process iterates until $IC(k-1) > IC(k)$. This method allows a smoothing of the curve avoiding false increasing. We present the histogram (Fig. 1(b)) on 256 grey levels of the test image *Lena* of size $N = 512 \times 480 = 245,760$ pixels (Fig. 1(a)). The initial histogram has numerous local valleys and we show on this example, the results of the merging process obtained thanks to the different IC. Figure 1(e) shows different curves $IC(k)$: AIC, BIC and φ^*_β, with $\beta = 0,15$. Indeed, for these values of β and N, the criteria AIC^* and φ^*_β are equivalent. One can see the curve smoothing obtained thanks to the merging process. The smallest number of bins is obtained using φ^*_β criterion. The number of bins reduction by AIC is weak: 73 bins from the 256 initial levels. If we consider this histogram obtained by AIC as reference, the one obtained by φ^*_β (Fig. 1(d)) keeps the main peaks, but with an important data reduction. The number of classes being important, the difference between the penalties of φ^*_β and BIC (or MDL) is interesting. Figure 1(c) shows the image *Lena* obtained by the 40 classes extracted by this criterion. The effects of the compression are not visually sensitive, and less than with a reduction to 40 same width bins. The effect on compression is significant: if a LZW-compressed TIFF file of the initial image



(a) original image (b) grey levels histogram



(c) Image Lena encoded using the 40 levels (d) Merging by φ^*_β (k=40)



(e) Behaviors of the criteria

Figure 1: image coding by the proposed method

Lena needs around 240 kB, only 97 kB are necessary for the resulted image.

4 APPLICATION TO THRESHOLDING

4.1 Maximum Likelihood thresholding

Let us recall the principles of histogram thresholding methods, classical techniques for image segmentation when the different objects can be distinguished by their grey-level values [5]. Let $X^N = X_1, \ldots, X_N$ be an observation sequence, with discrete values in $[1, m]$. The histogram $h(i), i = 1, \ldots, m$, is then built, as also the normalized histogram $p(i) = h(i)/N$ when $N = \sum_{i=1}^m h(i)$ is the number of observations. The image segmentation is a classification problem of the observations into k classes $c_1, \ldots, c_k$ where k is *a priori* given, thanks to $k-1$ thresholds $t_j, j = 1, \ldots, k-1$, and $t_0 = 0$ and $t_k = m$. Let us consider the following mixture model:

$$f\left(X^N \mid c_1, \ldots, c_k\right) = \prod_{j=1}^{k} \pi_j^{N_j} \prod_{i=1}^{N_j} f_j(X_i) \qquad (7)$$

where π_j is the prior for class c_j, that can be estimated, in the ML sense, by : $\widehat{\pi}_j = \sum_{i=t_{j-1}}^{t_j-1} p(i)$ and where N_j is the number of observations in the interval $[t_{j-1}, \ldots, t_j[$. The PDF associated with each mixture component f_j will be considered as gaussian. The mean and the variance of each class $c_j, j = 1, \ldots, k$, noted respectively μ_j and σ_j^2 can be estimated in the ML sense by $\widehat{\mu}_j = \widehat{\pi}_j^{-1} \sum_{i=t_{j-1}}^{t_j-1} ip(i)$ and $\widehat{\sigma}_j^2 = \widehat{\pi}_j^{-1} \sum_{i=t_{j-1}}^{t_j-1} (i - \widehat{\mu}_j)^2 p(i)$.

One obtains the following expression of the maximized log-likelihood of the model defined in (7):

$$L(k) = N \sum_{j=1}^{k} \widehat{\pi}_j \log \widehat{\pi}_j - \frac{N}{2} \log(1 + 2\pi) - \frac{N}{2} \sum_{j=1}^{k} \widehat{\pi}_j \log \widehat{\sigma}_j^2 \tag{8}$$

The location of the $k-1$ thresholds maximizing $L(k)$ allows the image thresholding. More accurately, let us consider the three following hypothesizes:

- Hypothesis H_1 (general case): no restriction on the parameters. Missing out the parameters that do not depend on k, one obtains the following criterion: $L_1(k) = N \sum_{j=1}^{k} \widehat{\pi}_j \log \frac{\widehat{\pi}_j}{\widehat{\sigma}_j}$ that is equivalent to the Kittler and Illingworth´s criterion [4].
- Hypothesis H_2: same variance σ^2 for each class. We propose: $L_2(k) = N \sum_{j=1}^{k} \widehat{\pi}_j \log \widehat{\pi}_j - \frac{N}{2} \log \widehat{\sigma}^2$.
- Hypothesis H_3: same variance σ^2, and *a priori* equiprobability of the classes. In this case, the obtained criterion is equivalent to Otsu´s criterion [8]: $L_3(k) = -\frac{N}{2} \log \widehat{\sigma}^2$.

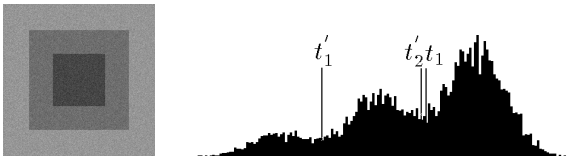
4.2 Application of the information criteria

Some IC have been already used in image processing [6, 13, 15]. We here propose to add in multi-thresholding methods, the contribution of consistent information criteria to obtain the optimal number of thresholds. These criteria always have the following form:

$$IC(k) = -2L_i(k) + n_k c_N \tag{9}$$

where $L_i(.), i = 1, 2, 3$, is one of the previous log-likelihood functions and n_k is the number of free parameters of the model. The value k minimizing (9) is taken as class number estimator. The usual criteria differ by the form of c_N. Under the hypothesis H_1, the number of free parameters n_k will be $3k-1$ (see [6]) since $\sum_{j=1}^{k} \pi_j = 1$. If we consider the hypothesis H_2 with a common variance, the number of free parameters is $2k$. For hypothesis H_3, the number of free parameters is $k+1$. We then obtain three expressions of the criteria where the $t_j, j = 1, \ldots, k-1$, are such that the log-likelihood $L_i(k)$ is maximum for a given k. Figure 2(a) presents a synthesized noised image coded on 256 grey levels ($N = 256 \times 256$) and $\sigma = 10$.

We can see on figure 2(b) the location of the single threshold t_1 ($k = 2$) or the two thresholds $t_1^{'}$ and $t_2^{'}$ under hypothesis H_1, for the noise level value $\sigma = 10$. The correct number ($k = 3$ components) of classes is found by the



(a) synthetized image (b) histogram and thresholds location

Figure 2: location of the thresholds on a synthetic image

criterion AIC^*. With the other penalties, the criteria values decrease when the number of classes increases. Table 1 gives, under this hypothesis H_1, the values of the criteria AIC and AIC^*. Different authors [6, 13] have been

Table 1: values of the criteria AIC and AIC^*

k	$AIC(k)$	$AIC^*(k)$
1	156,73	156,75
2	149,36	149,41
3	145,01	145,09
4	145,01	145,11

remarked that the penalties of classical criteria (AIC or MDL) are not sufficient and proposed empirical greater values. The criterion AIC^* offers an improvement of these criteria and is theoretically justified. We can see on the example given in table 1 that the penalty just allows to increase the criterion AIC^* from the number $k = 3$, but with a weak value in relation to the value of the criterion, under this hypothesis H_1. We have noted that in real situations involving document images, the histograms have few classes but often noised by the presence of little peaks. The Kittler and Illingworth´s method is sensitive to these noises, avoiding the convergence of criteria. On all the treated documents, hypothesis H_2 (common variance) gives the best results for the location of thresholds. Moreover, it is under this hypothesis that the criteria converge the easiest. But in some cases, the greatest penalty (AIC^*) with the best hypothesis (H_2) are not sufficient.

Figure 3 shows an image and AIC^* value. Table 2 gives the values of each penalty term and likelihood for the different number of classes. The analysis of the results shows

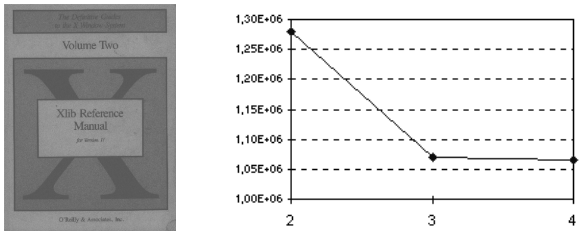


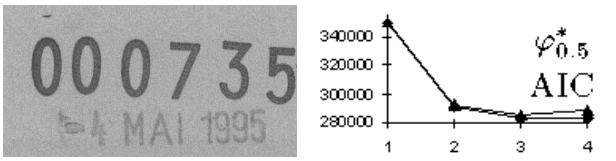
Figure 3: original image and AIC^* value

that the penalty term keeps weak in relation to likelihood

Table 2: behaviors of AIC^* according to each hypothesis

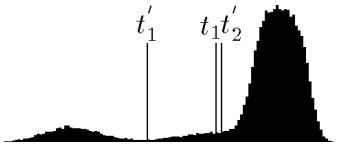
	$k=2$		$k=3$		$k=4$	
	L_1	$c_N n_k$	L_1	$c_N n_k$	L_1	$c_N n_k$
H_1	1214366	70	902850	112	445307	154
H_2	1280117	56	1070192	84	1065250	112
H_3	1020187	42	655668	56	579604	70

terms. So, the criterion φ^*_β should bring a solution to this problem under any hypothesis. Figure 4 shows the behaviour of φ^*_β on an example of a part of document, with $\beta = 0.5$. The three regions ($k = 3$) are well segmented for φ^*_β criterion by the method, under the hypothesis H_2.

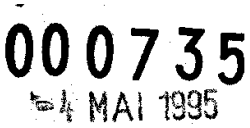


(a) Initial image

(b) Behaviors of two criteria



(c) Location of the thresholds



(d) Thresholded image

Figure 4: Part of a document image

5 CONCLUSION

In this paper, we have introduced the notion of information criteria and their definition within the framework of histogram based law approximation. The effect of this approximation on an image is an optimal reducing of the number of levels: the reduction factor equals 6 on the image *Lena*, allowing a strong compression without any visually sensitive degradation. The compression rate is important, 2.5 on the image *Lena*. In the last part of this paper, information criteria are again used to estimate the number and the location of thresholds in grey-level histograms. On the two kinds of applications treated here, we can observe the overparametrisation tendency of AIC, φ or even BIC. In that sense, the penalty term in φ^*_β seems to be a good improvement from AIC and BIC penalties. A problem is not solved: the choice of β value ($0 < \beta < 1$) is empirical.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Cáski, editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kaidó.

[2] A. El Matouat and M. Hallin. Order selection, stochastic complexity and Kullback-Leibler information. In P.M. Robinson and M. Rosenblatt, editors, *Time Series Analysis*, volume 2, in memory of E.J. Hannan, pages 291–299. Springer Verlag, New York, 1996.

[3] E .J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Roy. Stat. Soc. B*, 41(2):190–195, 1979.

[4] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19:41–47, 1986.

[5] T. Kurita, N. Otsu, and N. Abdelmalek. Maximum likelihood thresholding based on population mixture models. *Pattern Recognition*, 25(10):1231–1240, 1992.

[6] Z. Liang, J. Jaszack, and R. E. Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. *IEEE Trans. On Nuc. Sci.*, 39(4):1126–1131, 1992.

[7] C. Olivier, P. Courtellemont, O. Colot, D. de Brucq, and A. El Matouat. Comparison of histograms: A tool for detection. *Journal of Diagnosis and Safety in Automation*, 4(3):335–355, 1994.

[8] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. on Systems, Man and Cybernatic*, 9(1):62–66, 1979.

[9] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[10] J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[11] J. Rissanen, T.P. Speed, and B. Yu. Density estimation by stochastic complexity. *IEEE Trans. on Information Theory*, 38(2):315–323, March 1992.

[12] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[13] S. Sclove. Applications of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343, 1987.

[14] C.C. Taylor. Akaike's information criterion and the histogram. *Biometrika*, 74:636–639, 1987.

[15] Y. Wang, T. Lei, and J. M. Morris. Detection of the number of image regions by minimum bias/variance criterion. In Proc. Of the SPIE., editor, *Visual Communication and Image Processing´94*, volume 2308, pages 2020–2029, Chicago, USA, September 1994.