A non parametric approach for histogram segmentation

Julie DELON
CMLA, ENS de Cachan. E-mail: delon@cmla.ens-cachan.fr
Agnès DESOLNEUX
MAP5 - UFR Maths-Info Université Paris 5. E-mail: desolneux@math-info.univ-paris5.fr
José-Luis LISANI
Univ. Illes Balears, Spain. E-mail: joseluis.lisani@uib.es
Ana-Belén PETRO*
Univ. Illes Balears, Spain. E-mail: anabelen.petro@uib.es

***Abstract*— We propose a method to segment a 1D-histogram without *a priori* assumptions about the underlying density function. Our approach considers a rigorous definition of an admissible segmentation, avoiding over and under-segmentation problems. A fast algorithm leading to such a segmentation is proposed. The approach is tested both with synthetic and real data and an application to the segmentation of written documents is presented. We shall see that this application requires the detection of very small histogram modes, which can be accurately detected with our method.**

***Index Terms*— SEG-STAT, OTH-DOCU**

I. INTRODUCTION

Histograms have been extensively used in image analysis, and more generally in data analysis, mainly for two reasons: they provide a compact representation of large amounts of data and it is often possible to infer global properties of the data from the behavior of their histogram. One of the features that better describes a 1D-histogram is the list of its *modes*, *i.e.* the intervals of values around which data concentrate. For example the histogram of hues or intensities of an image made of different regions shall exhibit different peaks, each one of them ideally corresponding to a different region in the image. In this case, a proper segmentation of the image can be obtained by computing the appropriate thresholds that separate the modes in the histogram. However, it is not always easy to quantify the amount of "data concentration" in an interval, and hence to separate modes.

Among the algorithms proposed for 1D histogram segmentation, we can distinguish between parametric and non-parametric approaches. The first ones (see [13]) assume the set of data as samples of mixtures of k random variables of given distributions, as in the Gaussian Mixture Models. If k is known, optimization algorithms such as the EM algorithm [11] can estimate efficiently the parameters of these distributions. The estimated density can then be easily segmented to classify the original data. The main drawback of this approach is that histograms obtained from real data cannot always be modeled as mixtures of Gaussians, for example, luminance histograms of natural images, as we shall see in the experimental section. Non-parametric approaches give up any assumption on the underlying data density. Among them, bi–level or multi–level thresholding methods, such as [1], [7], [18], [22], divide the histogram into several segments by minimizing some energy criterion (variance, entropy, *etc*...).

In all cases, the number of modes in the final segmentation must be estimated. This number can be specified *a priori* and becomes a method parameter. It can also be estimated if its *a priori* distribution is hypothesized. The selection of this parameter is crucial since a wrong choice leads to an over or under segmentation of the data. Generally, *ad hoc* procedures are used to estimate the actual number of modes.

Other non-parametric approaches (for instance, mean shift [9]) find peaks (local maxima) of the histogram without estimating the underlying density. These methods tend to detect too many peaks in histograms coming from real noisy data. Some criterion is therefore needed to decide which of these peaks correspond to true modes ([25]). Indeed, one of the main challenges of histogram analysis is the detection of small modes among big ones (see, for example, Fig. 3).

An different approach has been recently proposed in [14]. The authors propose to fit the simplest density function compatible with the data. Such a method is globally convincing but the choice of the data-compatibility threshold is not formalized, only justified by experiments.

The limitations observed in the previous methods have motivated the development of a new non-parametric approach, robust to small variations in the histogram due to the limited number of samples, and local enough to detect isolated small modes.

In the following section the theoretical framework of the proposed approach is described in detail. Several tests are displayed in Section III, with applications to document segmentation.

II. A NEW APPROACH TO HISTOGRAM ANALYSIS

A density function f is said to be **unimodal** on some interval $[a, b]$ if f is increasing on some $[a, c]$ and decreasing on $[c, b]$. It seems appropriate to segment a histogram by looking for segments on which it is "likely" that the histogram is the realization of a unimodal law. On such intervals we

will say that the histogram is "statistically unimodal" (this expression will be precisely defined later). Obviously, such a segmentation is generally not unique. In particular the segmentation defined by all the local minima of the histogram has this property. However, small variations due to the sampling procedure should clearly not be detected as modes. In order to get a "minimal" division of the histogram, these fluctuations should be neglected. We arrive at two requirements for an admissible segmentation:

- in each segment, the histogram is "statistically unimodal",
- there is no union of several consecutive segments on which the histogram is "statistically unimodal".

What are the right tests to decide whether a histogram is "statistically unimodal" on an interval or not? In a non parametric setting, any unimodal density on the considered interval should be hypothesized and the compatibility between this density and the observed histogram distribution should be tested. Unfortunately, this leads to a huge number of tests and this is therefore impossible. There is, however, a way to address this question by testing a small number of adequate unimodal laws. In [16], this problem was solved for the case of decreasing laws. Our purpose here is to extend this method to the segmentation of any histogram into meaningful modes. We shall treat the problem in three stages in the next three sections:

- Step A: testing a histogram against a fixed hypothesized density,
- Step B: testing a histogram against a qualitative assumption (decreasing, increasing),
- Step C: segmenting a histogram and generating an estimate of its underlying density.

A. Distribution hypothesis testing

Consider a discrete histogram $h = (h_i)_{i=1\dots L}$, with N samples on L bins $\{1, \dots L\}$. The number h_i is the value of h in the bin i. It follows that

$$\sum_{i=1}^{L} h_i = N. \tag{1}$$

For each discrete interval $[a, b]$ of $\{1, \dots L\}$, let $r(a, b)$ be the proportion of points in $[a, b]$,

$$r(a, b) = \frac{1}{N} \left(\sum_{i=a}^{b} h_i \right). \tag{2}$$

Assume that an underlying discrete probability law $p = (p_i)_{i=1\dots L}$ is hypothesized for h. One would like to test the adequacy of the histogram h to this given density. For each interval $[a, b]$ of $\{1, \dots L\}$, let $p(a, b)$ be the probability for a point to fall into the interval $[a, b]$,

$$p(a, b) = \sum_{i=a}^{b} p_i. \tag{3}$$

Consider the hypothesis $\mathcal{H}_0$ that h originates from p. In other words, the N samples of the histogram h have been sampled independently on $\{1, \dots L\}$ with law p. A simple way to accept or to reject $\mathcal{H}_0$ is to test for each interval $[a, b]$ the similarity between $r(a, b)$ and $p(a, b)$. Under the hypothesis $\mathcal{H}_0$, the probability that $[a, b]$ contains at least $Nr(a, b)$ samples among N is given by the binomial tail $\mathcal{B}(N, Nr(a, b), p(a, b))$, where

$$\mathcal{B}(n, k, p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}. \tag{4}$$

In the same way, the probability that $[a, b]$ contains less than $Nr(a, b)$ samples is $\mathcal{B}(N, N(1 - r(a, b)), 1 - p(a, b))$. If one of these probabilities is too small, the hypothesis $\mathcal{H}_0$ can be rejected. Define for each interval $[a, b]$ its number of false alarms,

$$\text{NFA}_p([a, b]) = \begin{cases} \frac{L(L+1)}{2} \mathcal{B}(N, Nr(a, b), p(a, b)) \\ \qquad\qquad \text{if } r(a, b) \geq p(a, b), \\ \frac{L(L+1)}{2} \mathcal{B}(N, N(1 - r(a, b)), 1 - p(a, b)) \\ \qquad\qquad \text{if } r(a, b) < p(a, b). \end{cases} \tag{5}$$

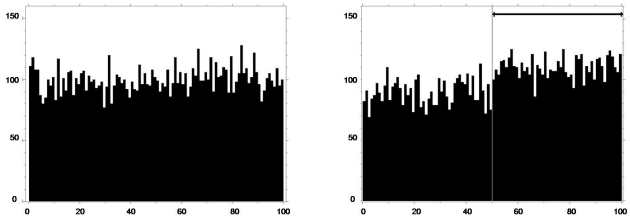
Definition 1 *An interval $[a, b]$ is said to be an ε-**meaningful rejection** of $\mathcal{H}_0$ if*

$$\textit{NFA}_p([a, b]) \leq \frac{\varepsilon}{2}. \tag{6}$$

Proposition 1 *Under the hypothesis $\mathcal{H}_0$, the expectation of the number of ε-meaningful rejections among all the intervals of $\{1, \dots L\}$ is smaller than ε.*

The proof of proposition 1 is obvious [12] and uses a Bonferroni argument, taking into account the number of tests $\frac{L(L+1)}{2}$ (the number of different intervals in $\{1, \dots L\}$). This means that testing a histogram h following a law p will lead on the average to less than ε wrong rejections. It may be asked how reliable this estimate is. In [17], Grompone and Jakubowicz have shown that the expectation of ε-meaningful events could be approximated by $\varepsilon/100$. This will be confirmed in section III (see table I). Thus in practice we fix $\varepsilon = 1$, and just talk about meaningful rejections.

Definition 2 *We say that a histogram h* **follows the law** *p on $[1, L]$ if h contains no meaningful rejection for $\mathcal{H}_0$.*



(a) (b)

Fig. 1. *Histograms of $N = 10000$ samples distributed on $L = 100$ bins, tested against the unifom law on $[1, 100]$. (a) Realization of the uniform law on $[1, 100]$. (b) Realization of a mixture of two uniform laws: $[1, 50]$ with a weight 0.45, and $[51, 100]$ with weight 0.55.*

Figure 1 shows two histograms which have been tested against the uniform law on $[1, 100]$. The first one is a realization of this law, and no rejection is found. The second is a mixture of two uniform laws on different intervals. In this case, several rejections of the uniform law on $[1, 100]$ are found. The rejection with the lowest NFA_p (the interval $[50, 100]$) is shown in Figure 1(b).

B. Testing the monotone hypothesis

Next we test if a histogram h follows a decreasing hypothesis on $[1, L]$ (the increasing case can be deduced by symmetry). This test will be useful later to give a suitable meaning to the expression "being statistically unimodal on an interval". The aim of a ideal test is to examine the adaptation of h to any decreasing density on $[1, L]$. This operation is obviously impossible but can be circumvented by using an estimate of the most likely decreasing law that fits h.

Let $\mathcal{P}(L)$ be the space of discrete probability distributions on $\{1, \dots L\}$, *i.e.,* the vectors $r = (r_i)_{i=1,\dots L}$ such that

$$\forall i \in \{1, 2, ..., L\}, \quad r_i \geq 0 \qquad \text{and} \qquad \sum_{i=1}^{L} r_i = 1. \tag{7}$$

Let $\mathcal{D}(L) \subset \mathcal{P}(L)$ be the space of all decreasing densities on $\{1, \dots L\}$. If $r = \frac{1}{N}h \in \mathcal{P}(L)$ is the normalized histogram of our observations, let $\bar{r}$ be the Grenander estimator of r. Introduced by Grenander in 1956 ([16]), this estimator is defined as the non-parametric maximum likelihood estimator restricted to decreasing densities on the line.

Definition 3 *The histogram $\bar{r}$ is the unique histogram which achieves the minimal Kullback-Leibler distance from r to $\mathcal{D}(L)$,* i.e.

$$KL(r||\overline{r}) = \min_{p \in \mathcal{D}(L)} KL(r||p), \tag{8}$$

where $\forall p \in \mathcal{D}(L)$, $KL(r||p) = \sum_{i=1}^{L} r_i \log \frac{r_i}{p_i}$.

Grenander shows in [16] (see also [3]) that $\bar{r}$ is merely "the slope of the smallest concave majorant function of the empirical repartition function of r". $\bar{r}$ also achieves the minimal L^2-distance from r to $\mathcal{D}(L)$. It can easily be derived from r by an algorithm called "Pool Adjacent Violators" (see [2], [4]).

Pool Adjacent Violators

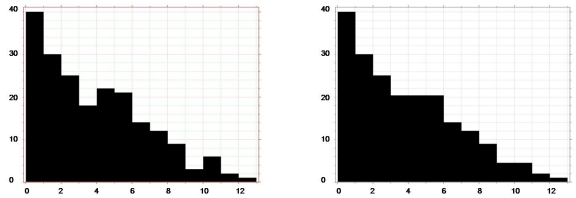
Consider the operator $D : \mathcal{P}(L) \to \mathcal{P}(L)$ defined by: for $r = (r_i)_{i=1,\dots L} \in \mathcal{P}(L)$, and for each interval $[i, j]$ on which r is increasing, i.e. $r_i \leq r_{i+1} \leq ... \leq r_j$ *and* $r_{i-1} > r_i$ *and* $r_{j+1} < r_j$*, set*

$$D(r)_k = \frac{r_i + ... + r_j}{j - i + 1} \text{ for } k \in [i, j], \quad \text{and } D(r)_k = r_k \text{ otherwise.} \tag{9}$$

This operator D replaces each increasing part of r by a constant value (equal to the mean value on the interval). A finite number (less than the size L of r) of iterations of D yields a decreasing distribution denoted $\overline{r}$:

$$\overline{r} = D^L(r). \tag{10}$$

An example of discrete histogram and its Grenander estimator are shown in Fig. 2.



(a) (b)

Fig. 2. *(a) Original histogram, (b) Grenander estimator obtained from the "Pool Adjacent Violators" algorithm.*

The previous definitions of meaningful rejections can obviously be applied to this case by taking $p = \bar{r}$ in the hypothesis $\mathcal{H}_0$, with $\bar{r}$ the Grenander estimator of $r = \frac{1}{N}h$.

Definition 4 *Let h be a histogram of N samples and $\overline{r}$ the Grenander estimator of $r = \frac{1}{N}h$. An interval $[a, b]$ is said to be a* ***meaningful rejection for the decreasing hypothesis*** *if*

$$\textit{NFA}_{\bar{r}}([a, b]) \leq \frac{1}{2}, \tag{11}$$

where $\textit{NFA}_p([a, b])$ is defined for any density law p in (5).

Definition 5 *We say that a histogram h* **follows the decreasing hypothesis** (*resp. the increasing hypothesis*) *on an interval $[a, b]$ if the restriction of the histogram to $[a, b]$* (i.e. $h_{|[a,b]} = (h_a, h_{a+1}, \dots h_b)$) *contains no meaningful rejection for the decreasing (resp. increasing) hypothesis.*

C. Piecewise unimodal segmentation of a histogram

Definition 6 *We say that a histogram h* ***follows the unimodal hypothesis*** *on the interval $[a, b]$ if there exists $c \in [a, b]$ such that h follows the increasing hypothesis on $[a, c]$ and h follows the decreasing hypothesis on $[c, b]$.*

We call segmentation of h a sequence $1 = s_0 < s_1 < \dots < s_n = L$. The number n is termed length of the segmentation. Our aim is to find an "optimal" segmentation S of h, such that h follows the unimodal hypothesis on each interval $[s_i, s_{i+1}]$ of S. If S is the segmentation defined by all the local minima of h, h follows obviously the unimodal hypothesis on each of its segments. But this segmentation is not reasonable in general (see Fig. 3 (a)). A segmentation following the unimodal hypothesis on each segment is generally not unique. In order to be sure to build a minimal (in terms of number of separators) segmentation, we introduce the notion of "admissible segmentation".

Definition 7 *Let h be a histogram on $\{1, \dots L\}$. A segmentation S of h is* ***admissible*** *if it satisfies the following properties:*

- *h follows the unimodal hypothesis on each interval $[s_i, s_{i+1}]$,*
- *there is no interval $[s_i, s_j]$ with $j > i + 1$, on which h follows the unimodal hypothesis.*

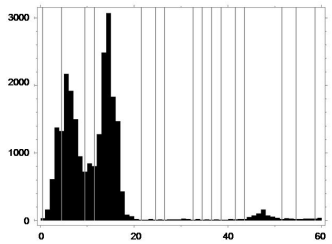
The first requirement avoids under-segmentations, and the second one avoids over-segmentations. It is clear that such a segmentation exists. Starting from the segmentation defined by all the local minima of h, merge recursively the consecutive intervals until both properties are satisfied.

Fine to Coarse (FTC) Segmentation Algorithm:

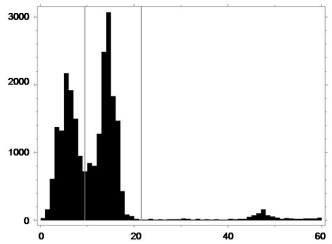
1) *Define the finest segmentation (i.e. the list of all the local minima, plus the endpoints 1 and L) S ={s_0,... , s_n} of the histogram.*
2) *Repeat:*
   Choose i randomly in $[1, length(S)-1]$. If the segments on both sides of s_i can be merged into a single interval $[s_{i-1}, s_{i+1}]$ following the unimodal hypothesis, group them. Update S.
   Stop when no more pair of successive intervals follows the unimodal hypothesis.
3) *Repeat step 2 with the unions of j segments, j going from 3 to $length(S)$.*

It must be remarked that step 3 is necessary since it can happen that the union of j segments follows the unimodal hypothesis while $k < j$ successive intervals contained in this union do not. For this reason all the possible combinations of successive intervals must be tested.

The result of this algorithm on a histogram is shown on Fig. 3.



(a)



(b)

Fig. 3. *(a) Initialization of the algorithm (all the local minima of the histogram). The histogram presents small oscillations, which create several local minima. (b) Final segmentation after FTC algorithm. Three modes are detected in this histogram, one is very small.*

In the histogram of Fig. 3, an energy–minimizing algorithm (for example the one presented in [1]) gives similar results if is specified that 3 segments are required. The separator between the second and third modes is not located exactly at the same place, but this variation has a negligible effect on the classification, since very few points are represented in this zone of the histogram. If only 2 segments are required, the second and third modes are united. It is interesting to note that for the energy defined in [1], the bimodal segmentation has almost the same energy as the three-modal segmentation. This implies that with a term penalizing the number of segments in the energy, the bimodal segmentation would certainly be chosen instead of the three-modal one. Therefore, the small mode cannot be found by this kind of method.

Figure 6 shows the result of the FTC algorithm on a more oscillating histogram. Popular techniques of histogram analysis such as mean shift [9] would over-segment this histogram, as noticed in [25], since many of the observed small oscillations would be detected as peaks.

III. Experiments

The experimental section is organized as follows: First, some experiments on synthetic data are performed to test the ability of the method to segment mixtures of laws without *a priori* assumption. Then, some experiments on image segmentation are displayed, and the validity of modeling real data histograms by Gaussian mixtures is discussed. The section ends with experiments on document segmentation, and the robustness of the method is tested.

A. Some results on synthetic data

>From a given probability law, 100 distributions, represented by $N = 2000$ samples each, were generated and quantized on 50 bins. For each distribution the number of segments found by the FTC algorithm was noted. Table I shows for different classical laws the number of distributions among the 100 leading to 1, 2 or 3 segments. The laws used here are the uniform law, a gaussian distribution of standard deviation 10 and mixtures of two gaussian functions $\frac{1}{2}\mathcal{N}(\mu,\sigma)+\frac{1}{2}\mathcal{N}(\mu+d,\sigma)$, with $\sigma = 5$ and $d = 2\sigma$, 3σ or 4σ.

For a uniform or a Gaussian law the number of segments is almost always found to be 1. For Gaussian mixtures, the results are of course closely related to the distance d between the means of the Gaussian distributions. When $d = 2\sigma$, the FTC algorithm always finds a single segment. It begins to find two segments when $d \simeq 2.5\sigma$, and finds two segments in 99% of the cases as soon as $d \geq 3.4\sigma$. Figure 4 shows that these results correspond to intuition. When $d = 2\sigma$, the two Gaussian functions cannot be distinguished, whereas the mixture clearly shows two modes when $d \geq 3\sigma$. These results also obviously depend on the number N of points. The larger N is, the more each distribution looks like the real mixture law, and the sooner the algorithm finds two segments. Of course, segmenting Gaussian mixtures can be made more efficiently

TABLE I

Number of segments found by the FTC algorithm for 100 samples of size 2000 of different laws.

	unif.	gauss.	mix. of 2 Gaussian laws		
			$d = 2\sigma$	$d = 3\sigma$	$d = 4\sigma$
1 segment	99	100	100	24	0
2 seg.	1	0	0	76	100
3 seg.	0	0	0	0	0

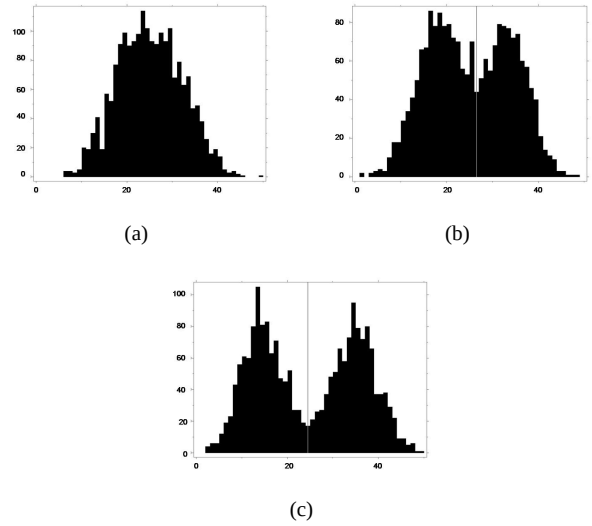


Fig. 4. *Examples of Gaussian mixtures of the form* $\frac{1}{2}\mathcal{N}(\mu,\sigma)+\frac{1}{2}\mathcal{N}(\mu+d,\sigma)$*, where* $\sigma=5$*. (a) Case* $d=2\sigma$*, (b)* $d=3\sigma$*, and (c)* $d=4\sigma$*. For* $d=2\sigma$*, the FTC algorithm finds no separator. For the other mixtures, the vertical line indicates the segmentation found.*

by dedicated algorithms if we really know they are Gaussian. In practice, observed mixtures are seldom Gaussian mixtures.

B. Some experiments on image segmentation

Figure 5 displays an image that contains a uniform background and a set of small objects of different intensities. The intensity histogram shows a large peak corresponding to the background and very small groups of values corresponding to the objects in the foreground. The FTC algorithm segments the histogram into four modes. The associated regions are shown in Fig. 5, each one of them corresponding to a different object in the scene.

Modes of a gray level histogram do not necessarily correspond to semantical visual objects. In general modes simply correspond to regions with uniform intensity and segmenting the histogram boils down to quantizing the image on the so-defined levels. This is the case in Fig. 6. The histogram of 'Lena' is automatically divided into 7 modes, and the corresponding image quantization is shown in Fig. 6 (c). Remark that no information about the spatial relations between image pixels is used to obtain the segmentation. Some authors (e.g. [23], [8]) propose the use of such information to improve the results of histogram thresholding techniques.

As mentioned in the beginning of section II, segmenting a histogram consists of two steps: 1. choosing a set of possible densities; 2. looking for the simplest of these densities which better adapts itself to the histogram for some statistical test. In the FTC algorithm, the densities proposed are a set of mixtures of unimodal laws, constructed from local Grenander estimators of the histogram. The test consists in looking for meaningful rejections. Another option is to use an EM algorithm to look for the best mixture of k Gaussian laws fitting the histogram. For each k, the adequacy of the mixture to the histogram is measured by a Kolmogorov-Smirnov test. The final segmentation is then defined by all the local minima

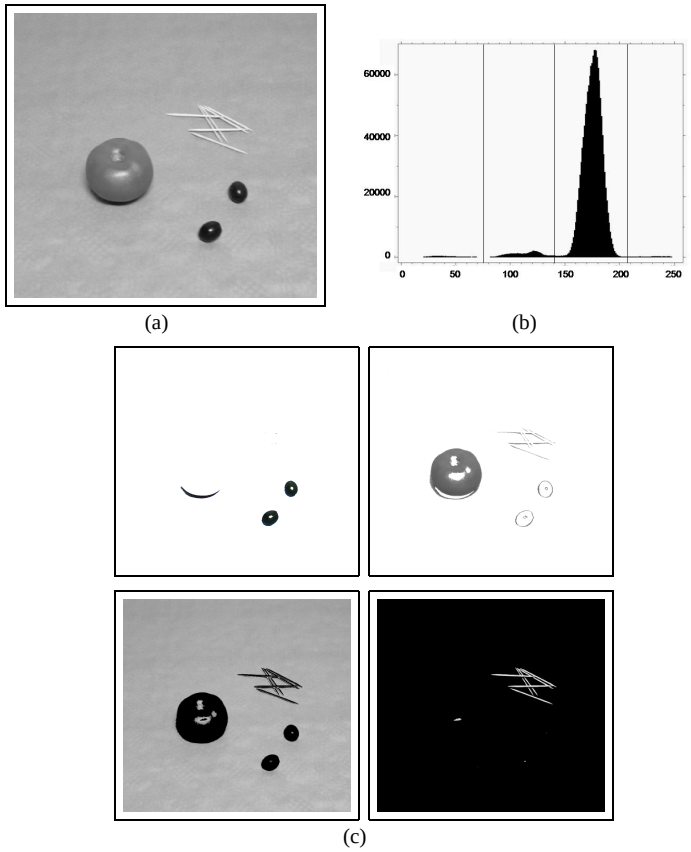
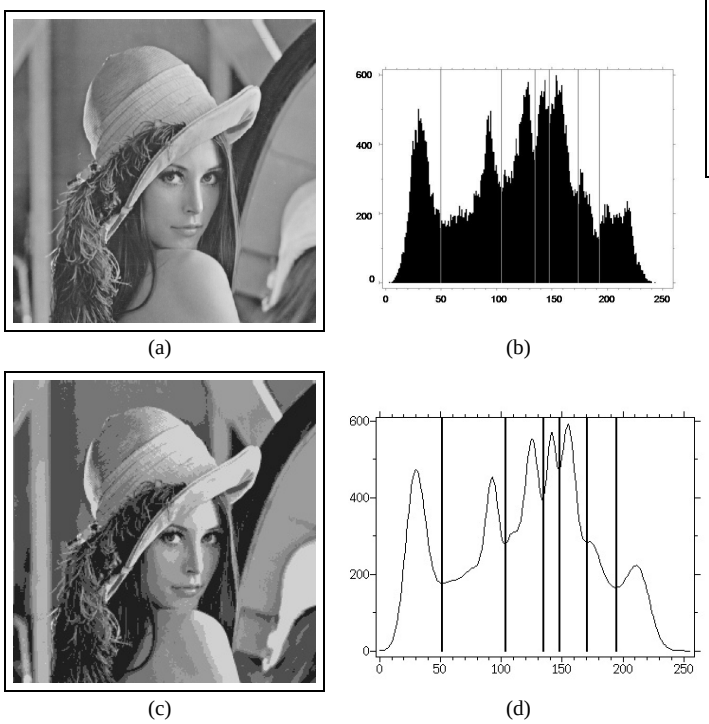


Fig. 5. *(a) original image (399 × 374 pixels), (b) its intensity histogram segmented into 4 modes. (c) Regions of the image corresponding to the 4 obtained histogram modes (in decreasing level of intensity, the background is either white or black depending on the mean intensity of the represented mode).*

of the selected mixture. This can be tested on the 'Lena' histogram. The EM algorithm is initialized by a k-means algorithm. For a significance level of 5% the first value of k leading to an accepted mixture is $k=14$ (the p-value for this mixture is 0.053). The adaptation between this mixture and the histogram is confirmed by a Cramer von Mises test at a significance level of 5% (p-value = 0.0659). Figure 6 (d) shows this best mixture of 14 Gaussian laws and indicates its local minima. Observe that this density is constituted of 7 modes that correspond exactly to the modes found previously.

With more demanding tests (a Chi-square test, or the search of meaningful rejections presented in section II-A), all mixtures are rejected until $k=20$ (the modes found in this case are still the 7 same modes). This illustrates the discrepancy between the number of Gaussians needed to correctly represent the law and the actual number of modes. This discrepancy can be explained by the following observation: when a digital picture is taken, the sensors of the camera and the post-processing that is used to store the picture into a file are non-linear. Even if the real intensity values of a given object followed a Gaussian law, the intensity distribution of this object on the picture would not be well represented by a Gaussian function. In particular, the corresponding mode on the histogram can be highly non-symmetric (see e.g. Fig.6). Such a mode needs several Gaussian laws to be well represented, whereas a unique

unimodal law fits it. As a consequence, looking for a mixture of unimodal laws is more adapted in this case than looking for Gaussian mixtures.



(a) (b) (c) (d)

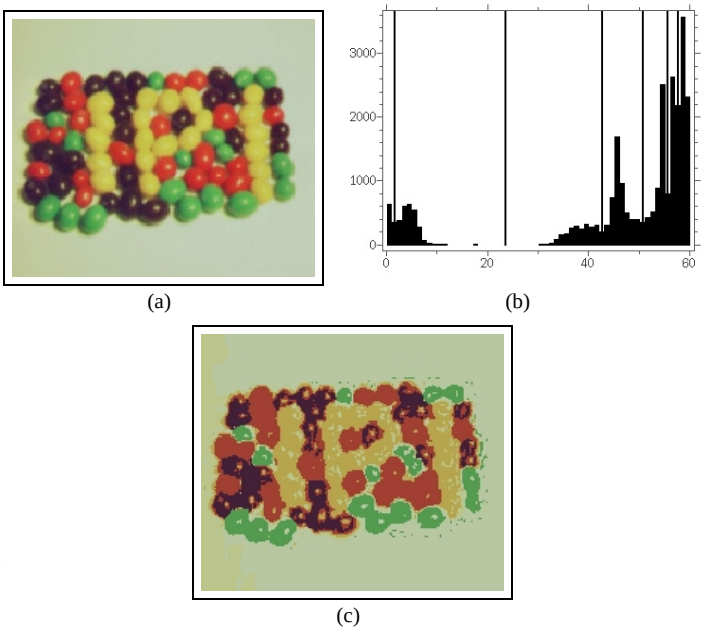
Fig. 6. *(a) Image (256 × 256 pixels) Lena. (b) Its intensity histogram segmented into 7 modes by the FTC algorithm. Observe that this histogram presents strong oscillations. In the initialization, the histogram presented 60 local minima among 256 bins. The segments have been merged until they follow definition 7 of an admissible segmentation. (c) Image Lena quantized on the 7 levels defined by the histogram segmentation shown in (b). (d) Best mixture of 14 Gaussian laws for the histogram (b), found by an EM algorithm. The local minima of this mixture, indicated by the vertical lines, correspond almost exactly to the separators found in (b).*

Figure 7 shows an example of image segmentation using the hues instead of the gray levels. Remark that hue histograms are circular. The FTC algorithm is perfectly adapted to this case.

C. Some experiments in document image analysis

Histogram thresholding is widely used as a pre-processing step for document understanding and character recognition. Its main use in this domain is to sort out the background and the characters in scanned documents. In this kind of documents, the intensity histogram generally presents two different modes: one large mode that represents the background, and another one, much smaller, corresponding to the text. Many different binarization methods have been proposed (see [21] and [15]) to find the best histogram thresholds for grayscale images. Nowadays, different methods are still studied, using simple spatial features ([10]), texture features ([20]) or mathematical morphology information ([6]).

However, binarization methods present two drawbacks. First, when the foreground region (the text here) is too small in comparison to the background (see Fig. 8), the position of the threshold becomes arbitrary, and the foreground may not be well detected. Second, binarization methods are not adapted



(a) (b) (c)

Fig. 7. (a) 'Beans' image. (b) Hue histogram of the image and corresponding segmentation in 6 modes (remark that the hue histogram is circular). (c) Corresponding segmentation of the image.

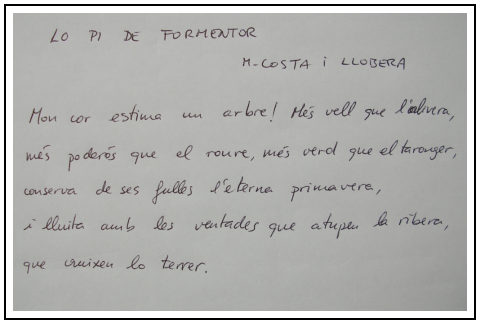
to any kind of written documents (see Fig. 10). When the background pattern is complicated, or when different inks are used in the text, segmenting the histogram in only two modes is not a good solution. Finding automatically the number of modes in the histogram allows one to get more than two modes when necessary.

In simple written documents, where only one ink has been used, the segmentation found by the FTC algorithm is bimodal. This is the case of the example shown in Fig. 8. The histogram is segmented into two modes, one of them corresponding to the text characters (see Fig. 8(c)). This example shows that the FTC algorithm is able to find very small modes when they are isolated enough. In Fig. 9, although the image presents several different gray shades, the FTC algorithm also segments the histogram into two modes (Fig. 9(b)), separating clearly the characters in the check from the background, as we can see in Fig. 9(c). In these experiments, it must be underlined that the size of the images can interfere in the results. In a text image, the larger the proportion of white pixels is, the more difficult it becomes to extract a black mode from the histogram, since it tends to become negligible in front of the white one. In these cases, it is interesting to narrow the image around the text.

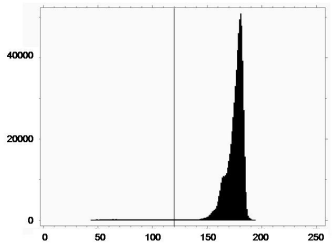
In the case of the histogram of the image shown in Fig. 10, the algorithm finds three different modes, corresponding to three intensity regions in the image. The first mode represents the band in the bottom of the image, the second mode corresponds to the text and the stars of the image, and the third one is the background. A bi–level histogram thresholding method could not yield this separation (e.g. [22], [21]).

D. Sensibility of the method

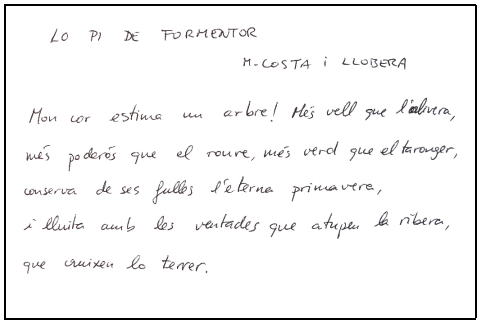
Generally, two factors can influence the segmentation: the noise in images and the histogram quantization noise.



(a)



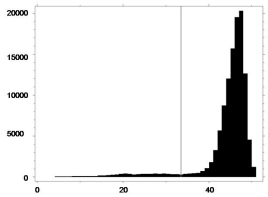
(b)



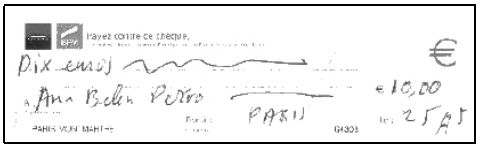
(c)

Fig. 8. *(a) Original image (1010 × 661 pixels). (b) Intensity histogram and the threshold obtained. (c) Pixels corresponding to the left segment of the histogram.*

(a)



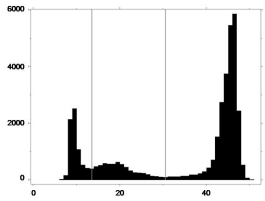
(b)



(c)

Fig. 9. *(a) Original image (755 × 201 pixels), (b) Intensity histogram of the original image and corresponding segmentation. This histogram presents several local minima, but the final segmentation is bimodal. (c) Pixels corresponding to the first segment of the histogram.*

(a)



(b)

Fig. 10. *(a) Original image (246 × 156 pixels), (b) Intensity histogram with the 3 modes obtained. The first mode on the left corresponds to the lower and darker band of the image, the middle mode corresponds to the text and the stars, and the last mode is the background one.*

Theoretically if we add a noise b to an image u, its intensity distribution becomes $h_u * h_b$, where h_u is the gray level histogram of u, and h_b the noise histogram. This results in a blur in the histogram. If the image has N pixels, and if the noise is an impulse noise, added to $p\%$ of the pixels, then $h_u * h_b = (1-p)h_u + p\frac{N}{256}\mathbf{1}_{[0,255]}$, which is not really disturbing for the FTC algorithm (adding a uniform noise on a histogram does not change its unimodality property on a given interval). Moreover this kind of noise can be easily removed by a median filter on the image. In the case of gaussian noise, the operation smoothes the shape of the histogram h_u. As a consequence, the number of modes found can decrease when the standard deviation of the noise increases too much. However, this kind of image noise can be efficiently handled by NL-means algorithms [5] before computing the histogram.

The performance of the FTC algorithm in the presence of additive noise can be evaluated as follows ([24], [19]): (1) create a synthetic image (Figure 11, top) and segment it manually; (2) add increasing quantities of uniform noise (Figure 11, middle and bottom); (3) segment the histograms using FTC and evaluate the probability of error by applying:

$$P(\text{error}) = \sum_{j=1}^{N} \sum_{i=1, i\neq j}^{N} P(R_i|R_j)P(R_j) \qquad (12)$$

where N is the number of regions in the manually segmented image ($N = 4$ in our example), $P(R_j)$ is the proportion of pixels in the jth region and $P(R_i|R_j)$ is the proportion of pixels in the jth region assigned to the ith region by the FTC algorithm.

The results of this evaluation are shown in Table II. Since the histogram of the original synthetic image was composed of 4 gaussians, the EM algorithm was also used to estimate this mixture. Remark that when SNR decreases the gaussian mixture hypothesis no longer holds and the EM algorithm gives poor results. The FTC method, however, is able to cope with this distortion down to lower SNR values.

The real noise in histograms is the quantization noise, coming from the fact that the histograms have a finite number

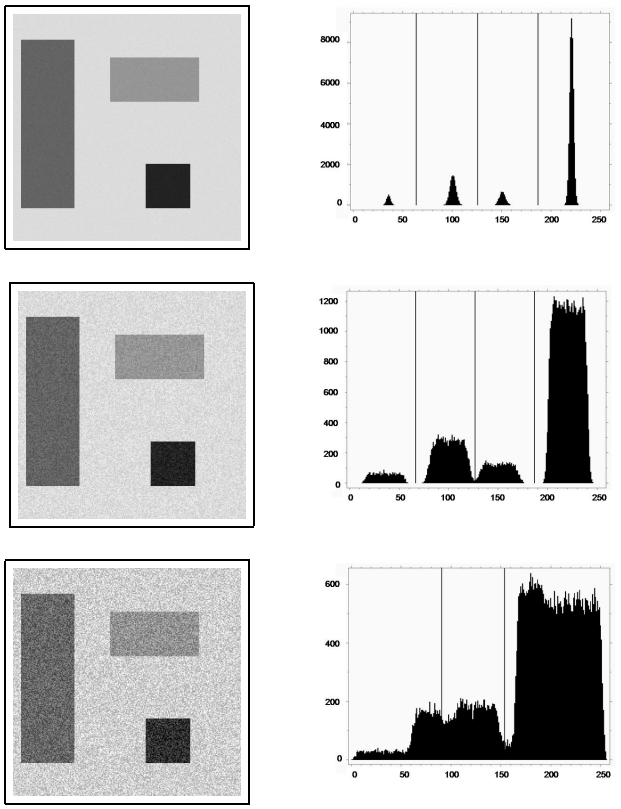
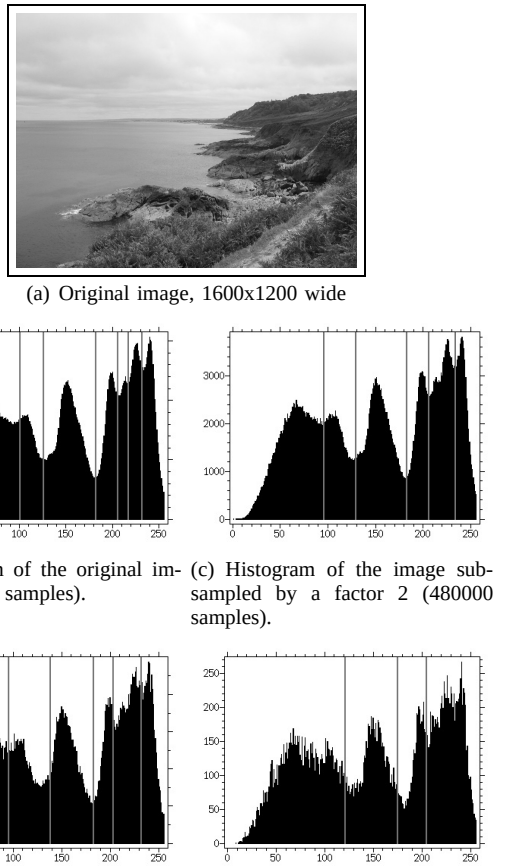


Fig. 11. Performance evaluation of the FTC algorithm in the presence of additive noise. Top, reference image (256×256) and its histogram. Middle and bottom, images corrupted with uniform noise (SNR=24dB and SNR=17dB, respectively), and their corresponding histograms. The segmentation results are marked on the histograms.



(a) Original image, 1600x1200 wide

(b) Histogram of the original image (1920000 samples).

(c) Histogram of the image subsampled by a factor 2 (480000 samples).

(d) Histogram of the image subsampled by a factor 4 (120000 samples).

(e) Histogram of the image subsampled by a factor 8 (30000 samples).

Fig. 12. Sensibility of the method to quantization. The larger the number of samples is, the more certain the segmentation is. It follows that the histogram is more and more segmented when N increases. The segmentation tends towards the segmentation of the deterministic histogram of the continuous underlying image.

N of samples. If a histogram h originates from an underlying density p, the larger N is, the more h looks like p. When $N \to \infty$, a segmentation algorithm should segment the histogram at each local minima of its limit p. Consider the example of Fig. 12. An image of size 1600x1200 is subsampled several times by a factor 2. Each intermediate image yields an histogram. These histograms can all be considered as realizations of the density given by the histogram of the original image. The smaller the number of samples is, the less information we have, and the less the histogram can be segmented with certainty. Figure 12 shows that the number of segments found by the FTC algorithm increases with N. The separators tend towards the separators of the deterministic histogram of the continuous underlying image.

IV. Conclusion

This papers presents a new approach to segment a histogram without *a priori* assumptions about the number or shape of its modes. The central idea is to test the simplest multimodal law fitting the data. The proposed adequacy test, called "meaningful rejections", is a multiple test which presents the advantage of being simultaneously local and global. This method is more generic than looking for Gaussian mixtures and avoids overestimating the number of modes. The corresponding algorithm is able to detect very small modes when they are isolated, which makes it well adapted to document analysis. The statistical aspect of the approach makes it robust to quantization noise: the larger the number of samples is, the more the histogram can be considered as deterministic, and the more it is segmented. Several tests on histograms computed from real or synthetic data endorse the efficiency of the method. Now, it is clear that such a method should be extended to higher dimension in order to segment color histograms. First results have been obtained by segmenting hierarchically color histograms in the HSV space. A direct adaptation of the method to any dimension is currently studied.

TABLE II

Performance evaluation.

SNR (dB)	Inf.	36	30	27	24	22	17
P(error) FTC	0	0	0	0	0	0.08	0.14
P(error) EM	0	0	0	0.08	0.44	0.72	0.71

Acknowledgement This work has been partially financed by the Office of Naval research under grant N00014-97-1-0839, the Direction Générale des Armements (DGA), the Ministère de la Recherche et de la Technologie (projet RNRT ISII) and the Ministerio de Ciencia y Tecnologia under grant

TIC2002-02172.

REFERENCES

[1] ALVAREZ, L., AND ESCLARÍN, J. Image quantization using reaction-diffusion equations. *j-SIAM-J-APPL-MATH 57*, 1 (Feb. 1997), 153–175.

[2] AYER, M., BRUNK, H., EWING, G., REID, W., AND SILVERMAN, E. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics 26*, 4 (1955), 641–647.

[3] BARLOW, R., BARTHOLOMEW, D., BREMNER, J., AND BRUNK, H. *Statistical Inference Under Order restrictions*. Wiley, New York, 1972.

[4] BIRGÉ, L. The Grenander estimator: A nonasymptotic approach. *The Annals of Statistics 17*, 4 (1989), 1532–1549.

[5] BUADES, A., COLL, B., AND MOREL, J. A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation (SIAM interdisciplinary journal) 4*, 2 (2005), 490–530.

[6] CALDAS, J., BANDEIRA, L., DA COSTA, J., AND PINA, P. Combining fuzzy clustering and morphological methods for old documents recovery. *IbPRIA Proceedings*, 2 (2005), 387–394.

[7] CHANG, C., CHEN, K., WANG, J., AND ALTHOUSE, M. L. G. A relative entropy based approach to image thresholding. *Pattern Recognition 27(9)* (1994), 1275–1289.

[8] CHENG, H., AND SUN, Y. A hierarchical approach to color image segmentation using homogeneity. *IEEE Transactions on Image Processing 9*, 12 (2000), 2071–2082.

[9] COMANICIU, D., AND MEER, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24* (2002), 603–619.

[10] DAWOUD, A., AND KAMEL, M. Iterative mutimodel subimage binarization for handwritten character segmentation. *IEEE Image Processing 13*, 9 (2004), 1223.

[11] DEMPSTER, A., LAIRD, N., AND RUBIN, D. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B 39* (1977), 1–38.

[12] DESOLNEUX, A., MOISAN, L., AND MOREL, J.-M. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 4 (2003), 508–513.

[13] DUDA, R., HART, P., AND STORK, D. *Pattern Classification*. John Wiley and Sons, 2000.

[14] FREDERIX, G., AND PAUWELS, E. A statistically principled approach to histogram segmentation. Tech. rep., CWI, 2004.

[15] GLASBEY, C. An analysis of histogram-based thresholding algorithms. *CVGIP: Graphical Models and Image Processing 55*, 6 (1993), 532–537.

[16] GRENANDER, U. *Abstract Inference*. Wiley, New York, 1980.

[17] GROMPONE, AND JAKUBOWICZ. *Forthcoming*.

[18] KAPUR, J. N., SAHOO, P. K., AND WONG, A. K. C. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing 29* (1985), 273–285.

[19] LIM, Y., AND LEE, S. On the color image segmentation algorithm based on the threshold and the fuzzy c-means technique. *Pattern Recognition 23*, 9 (1990), 935–952.

[20] LIU, Y., AND SRIHARI, N. Document image binarization based on texture features. *IEEE PAMI 19*, 5 (1997), 540.

[21] OTSU, N. A threshold selection method from grey-level histograms. *IEEE Transactions on Systems*, 19 (1979), 62–66.

[22] PUN, T. A new method for grey-level picture thresholding using the entropy of the histogram. *Signal Processing 2* (1980), 223–237.

[23] SAHA, P., AND UDUPA, K. Optimum image thresholding via class uncertainty and region homogeneity. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23* (2001), 689–706.

[24] TOBIAS, O., AND SEARA, R. Image segmentation by histogram thresholding using fuzzy sets. *IEEE Transactions on Image Processing 11*, 12 (2002), 1457–1465.

[25] WANG, H., AND SUTER, D. False-peaks-avoiding mean shift method for unsupervised peak-valley sliding image segmentation. In *Proc. VIIth Digital Image Computing: Techniques and Applications* (2003), pp. 581–590.