Image Segmentation by Nonparametric Clustering Based on the Kolmogorov-Smirnov Distance

Eric J. Pauwels[1,2] and Greet Frederix[2,3]

[1] Centre for Mathematics and Computer Science (CWI),
Kruislaan 413, 1098 SJ Amsterdam, The Netherlands
Eric.Pauwels@cwi.nl
[2] ESAT-PSI, K.U.Leuven, K. Mercierlaan 94,
B-3001 Heverlee, Belgium
Eric.Pauwels@esat.kuleuven.ac.be
[3] Dept. of Mathematics, K.U.Leuven, Celestijnenlaan 200 B,
B-3001 Heverlee, Belgium
Greet.Frederix@esat.kuleuven.ac.be

Abstract. In this paper we introduce a non-parametric clustering algorithm for 1-dimensional data. The procedure looks for the simplest (i.e. smoothest) density that is still compatible with the data. Compatibility is given a precise meaning in terms of the Kolmogorov-Smirnov statistic. After discussing experimental results for colour segmentation, we outline how this proposed algorithm can be extended to higher dimensions.

1 Motivation and Overview

The quest for robust and autonomous image segmentation has rekindled the interest of the computer vision community in the generic problem of *data clustering* (see e.g. [3,6,15,1,2,16]). The underlying rationale is rather straightforward: As segmentation algorithms try to divide the image into regions that are fairly homogeneous, it stands to reason to map the pixels into various feature-spaces (such as colour- or texture-spaces) and look for clusters. Indeed, if in some feature-space pixels are lumped together, this obviously means that, with respect to these features, the pixels are similar. By the same token, image regions that are perceptually salient will map to clusters that (in at least some feature-spaces) are clearly segregated from the bulk of the data.

Unfortunately, the clustering problems encountered in segmentation applications are particularly challenging, as neither the number of clusters, nor their shape is known in advance. Moreover, clusters are frequently unbalanced (i.e. have widely different sizes) and often distinctly non-Gaussian (e.g. skewed). This heralds serious difficulties for most "classical" clustering algorithms that often assume that the number of clusters is known in advance (e.g. K-means), or even that the shape of the data-density is explicitely specified up to a small number of parameters that can be estimated from the data (e.g. Gaussian Mixture Models (*GMM*)).

D. Vernon (Ed.): ECCV 2000, LNCS 1843, pp. 85–99, 2000.

Furthermore, strategies to estimate the number of clusters prior to, or concurrent with, the actual clustering are of limited value as they tend to be biased towards solutions that favour spherical or elliptical clusters of roughly the same size. The root for this bias is to be found in the fact that almost all cluster-validity criteria compare variation *within* to variation *between* clusters (for more details we refer to standard texts such as [9,11,10]).

To circumvent the problems outlined above, we focus on clustering based on **non-parametric density estimation** (for prior work, see e.g. [3,15]). In contradistinction to *parametric* density estimation (such as *GMM*), no explicit parametric form of the density is put forward, and the data-density is obtained by convolving the dataset by a density-kernel. More precisely, given an d-dimensional dataset $\{\mathbf{x}_i \in I\!R^d;\ i = 1 \ldots n\}$ a density $f(\mathbf{x})$ is obtained by convolving the dataset with a unimodal density-kernel $K_\sigma(\mathbf{x})$:

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(\mathbf{x} - \mathbf{x}_i), \tag{1}$$

where σ is the size-parameter for the kernel, measuring its spread. Although almost any unimodal density will do, one typically takes K_σ to be a (rotation-invariant) Gaussian density with σ^2 specifying its variance:

$$K_\sigma(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{d/2} e^{-\|\mathbf{x}\|^2/2\sigma^2}. \tag{2}$$

After convolution we identify clusters by using *gradient ascent* (hill-climbing) to pinpoint local maxima of the density f. This procedure ends up assigning each point to a nearby density-maximum, thus carving up the data-set in compact and dense clumps.

However, it is obvious that unless the width σ is judiciously picked within a fairly narrow range, this procedure will result in either too many (if σ is chosen too small) or too few clusters (if σ is set too large). Although a huge bulk of the work on density-estimation concerns itself with this problem of choosing an "optimal" value for σ (e.g. see the book by Thompson and Tapia [18]), it is fair to say that it remains extremely tricky to try and estimate optimal (or even acceptable) clustering parameters.

For this reason we propose a different approach: We start from a sub-optimal (too small) choice for σ, and then modify the resulting density f directly. The proposed modification (which will be detailed in Section 3) is based on the *Kolmogorov-Smirnov statistic* and the resulting criterion has therefore a precise and easy to grasp meaning, which does not involve arbitrarily chosen parameters.

The rest of this paper is organised as follows. In Section 2 we will argue that performance of clustering is improved if the dimensionality of the problem can be meaningfully reduced. Rather than trying to combine all the information in one huge feature-vector, we will champion the view that it makes sense to look at as simple a feature as reasonable. This amounts to projecting the high-dimensional data-set on low-dimensional subspaces and is therefore similar in

spirit to *Projection Pursuit*, a technique used in data analysis, where projections on low-dimensional subspaces (1- or 2-dimensional) are used to gain insight into the structure of high-dimensional data.

One particularly interesting and useful case of the aforementioned dimension reduction is that of clustering one-dimensional data, which boils down to partitioning the corresponding histogram. This topic is discussed extensively in Section 3 for several reasons. First, although one can argue that this is just a special case of the general n-dimensional clustering problem, the topology of a 1-dimensional (non-compact) space (such as $\mathbb{R}$) is unique in that it allows a *total order*. As a consequence, the mathematical theory is well understood and yields sharp results. Furthermore, the 1-dimensional case furnishes us with a useful stepping stone towards the more complex high-dimensional case that will be discussed in Section 5. Finally, Section 4 will report on results obtained for colour segmentation.

2 High-Dimensional Versus Low-Dimensional Clustering

Like most statistical procedures, clustering in high-dimensional spaces suffers from the dreaded *curse of dimensionality*. This is true in particular, for density estimation, as even for large data sets, high-dimensional space is relatively empty.

As a consequence the reliability and interpretability of the resulting clustering may be improved whenever it is possible to reduce the dimensionality of the problem. In particular, this argument indicates that it is often ill-advised to artificially increase the dimensionality of the problem by blindly concatenating feature-vectors into high-dimensional datapoints. More precisely, if there is no theoretical or prior indication that features are mutually dependent, it is advisable to cluster them separately. The reason for this is straightforward: if features $x_1, x_2, \ldots, x_n$ are independent, then their joint probability density function factorizes into a product of 1-dimensional densities:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \ldots f_n(x_n), \tag{3}$$

and interesting structure in the joint density f will also be apparant in (one of) the marginal densities f_i. For instance, computing the mean and variance of the gray-values in a small window about every pixel produces two features at each pixel. However, for an unconstrained image there is no reason why these two features would be dependent. Therefore, it makes sense to cluster them separately, rather than confounding the problem by focussing exclusively on their joint distribution.

In particular, there are a number of perceptually relevant dichotomies (e.g. dark versus bright, horizontal versus vertical, direction versus randomness, coloured versus gray, textured versus flat, etc.) that can be captured mathematically in a relatively straightforward fashion, but that nevertheless yield important clues for segmentation. This means that it makes sense to start studying 1-dimensional densities (simple histograms) and this will be our main point of focus for most of this paper.

Indeed, one of the motivations for this work is the observation that lots of effort in computer learning and artificial intelligence focuses on ways of finding transformations (often non-linear ones) that vastly reduce the *dimensionality* of the problem. The assumption is that in many cases there is a relatively small set of so-called *latent variables* that capture the intrinsic structure of the problem and by determining the intrinsic dimensionality of the data, these (hidden) variables are brought to the fore. Exponents of this approach are classical methodologies such as *principal component analysis* (PCA) and multi-dimensional scaling, but also more recent developments of similar flavour such as *projection pursuit* (PP), *generative topographic mapping* (GTM), Kohonen's *self-organising maps* (SOM) and *independent component analysis* (ICA). The latter is actually looking for transformations that decouple different components such that the factorisation in eq.(3) is — at least approximately — realised.

3 Histogram Segmentation and 1-Dimensional Clustering

3.1 The Empirical Distribution Function

In this section we will concentrate on finding clusters in a sample $x_1, \ldots, x_n$ of 1-dimensional data. In principle, clustering 1-dimensional data by segmenting the histogram should be fairly straightforward: all we need to do is locate the peaks (local maxima) and valleys (local minima) of the data density (for which the histogram is an estimator) and position the cluster boundaries at the local minima. However, the problem is that the number and position of these local minima will strongly depend on the width of the histogram bins. An appropriate choice for this parameter is difficult to make.

For this reason we have decided to use the *cumulative density function* (also called the *distribution function*) as the tool of choice for segmentation, since it allows a *non-parametric approach* (see below). We recall that for a stochastic variable X with density function f, the cumulative density (distribution) F is defined in terms of the probability P by

$$F(x) := P(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

Of course, in most cases of interest the underlying density f is unknown and we proceed by using the *empirical distribution* F_n, which for a sample $X_1, \ldots, X_n$ is given by

$$F_n(x) = \frac{\#\{i \,:\, X_i \leq x\}}{n} \tag{4}$$

One can prove (see eg.[12]) that F_n is an adequate estimator of F, as for instance

$$F_n(x) \longrightarrow F(x) \qquad \text{as} \qquad n \longrightarrow \infty.$$

at every *continuity point* x *of* F.

Compared to the histogram, the empirical distribution has a number of advantages. First, it is parameter-free as it is completely determined by the data itself and there is no need to judiously pick values for critical parameters such as bin-width. Second, working with the cumulative density rather than with the density itself has the added benefit of stability. Indeed, the integration operation which transforms f into F smooths out random fluctuations, thereby highlighting the more essential characteristics. And last but not least, using the distribution allows us to invoke the Kolmogorov-Smirnov statistic, a powerful non-parametric test that can be used to compare arbitrary densities. This theme will be elaborated further in the next section.

3.2 Non-parametric Density Estimation Using Kolmogorov-Smirnov

To make good on our promise to proceed in a non-parametric fashion, we proceed by asking ourselves the question: *What is the smoothest density g that is compatible with the data, in the sense that the corresponding cumulative distribution G is not significantly different from the empirical distribution F_n?* This is basically a reformulation of *Occam's razor* and in that sense akin to the MDL-principle that has made several appearances in this context. To tackle this question we note that, recast in the appropriate mathematical parlance, it reads as follows (see Fig. 1): Find the density g that solves the following constrained minimisation problem:

$$\text{minimize} \quad \Phi(g) \equiv \int_{\mathbb{R}} (g'(x))^2 \, dx, \qquad \text{subject to} \qquad \sup_x |G(x) - F_n(x)| \leq \epsilon_n, \tag{5}$$

where ϵ_n is the critical value for the Kolmogorov-Smirnov statistic at an appropriate significance level, e.g. 5% (details regarding the Kolmogorov-Smirnov statistic can be found in section 3.3).

As there is no straightforward closed form solution to this problem, we proceed by invoking a *gradient descent* procedure,

$$\frac{\partial g}{\partial t} = -D\Phi(g), \tag{6}$$

but this calls for a precise definition of the gradient of a functional. This concept is studied extensively in funtional analysis and we briefly remind the reader of the relevant definition (for more details, see e.g. Troutman[19], p. 44). To motivate the approach we recall that in classical calculus, the rate of change of a function in a specified direction is obtained by taking the inner-product of the gradient and the unit-vector in the specified direction. Exactly the same procedure can be used for functionals: The standard inner product on function spaces is given by

$$< f, g > = \int_{\mathbb{R}} f(x) g(x) \, dx$$

and the functional equivalent of a directional derivative is provided by the important concept of the *Gâteaux derivative of Φ at g in the direction of v:*

$$D_v\Phi(g) := \lim_{\epsilon\to 0} \frac{\Phi(g+\epsilon v) - \Phi(g)}{\epsilon} = \frac{\partial}{\partial \epsilon}\Phi(g+\epsilon v)\bigg|_{\epsilon=0} \tag{7}$$

Under quite mild regularity conditions one can prove that for each g there is a unique function w_g such that for all v, $D_v\Phi(g) =< w_g, v >$. This function is called the *gradient of Φ at g* and denoted by $D\Phi(g)$, resulting in the suggestive formula

$$D_v\Phi(g) =< D\Phi(g), v > \qquad \text{(for all } v\text{)} \tag{8}$$

which is formally identical to the corresponding formula in standard vector calculus relating the gradient to an arbitrary directional derivative.

It is now straightforward to compute the gradient for the functional in (5). Plugging the explicit form of the functional Φ into eq.(7) yields:

$$\begin{aligned} D_v\Phi(g) &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_{\mathbb{R}} [(g' + \epsilon v')^2 - g'^2]\, dx \\ &= \lim_{\epsilon\to 0} \frac{1}{\epsilon} \int_{\mathbb{R}} [(2\epsilon g' v' + \epsilon^2 v'^2]\, dx \\ &= 2\int_{\mathbb{R}} g' v'\, dx \end{aligned}$$

Next, using integration by parts and the assumption that the density function g and its derivatives vanish at infinity (a reasonable assumption for a density modelling a histogram), it immediately follows that

$$D_v\Phi(g) = -2 < g'', v > \quad \text{whence,} \quad D\Phi(g) = -2\frac{\partial^2 g}{\partial x^2}$$

Therefore the gradient-descent method for the functional Φ gives rise to the heat equation:

$$\frac{\partial g}{\partial t} = c\frac{\partial^2 g}{\partial x^2}, \qquad (c \text{ appropriate conductivity coefficient}) \tag{9}$$

which suggests the following **strategy** to search for a minimum in eq. (5): Take an initial (fine-grained, i.e. small bins) estimate $g = g_0$ for the density, e.g. by constructing a histogram with small bins, or using a kernel estimator (as in (1)) with σ sufficiently small. Next, subject g by plugging it into diffusion equation (9) with g_0 as initial condition. After each diffusion step, compute the cumulative density

$$G(x) = \int_{-\infty}^{x} g(u)du$$

by (numerically) integrating g. Now stop the diffusion the moment the constraint in (5) is violated and use the final g as the estimate for the density for which valleys and peaks can be determined.

Although this approach has been implemented and yields very satisfactory results, we hasten to point out that there is no guarantee that the evolution equation (9) actually ends up at a minimum (even a local one). The reason for this is that although the functional is quadratic, the diffusion is stopped as soon as it hits (domain-boundary specified by) the constraint. In most cases it will be possible to further reduce the functional Φ by sliding along the constraint.

In fact, one obvious way for doing this would be to make the diffusion coefficient c in eq(9) dependent on the Kolmogorov-Smirnov difference:

$$\rho(x) = |G(x) - F_n(x)|$$

yielding a non-linear diffusion:

$$\frac{\partial g}{\partial t} = c(\rho(x)) \frac{\partial^2 g}{\partial x^2}, \qquad \text{where e.g.} \quad c(\rho) = \exp\left(-\frac{\rho^2}{\epsilon_n^2 - \rho^2}\right) \quad (0 \leq \rho \leq \epsilon_n). \tag{10}$$

The conductivity coefficient c is engineered to behave like a Gaussian function near the origin, but to drop smoothly to zero when the difference ρ approaches the critical distance ϵ_n. This ensures that the diffusion is stopped wherever the smoothed density is about to violate the constraint, whereas it can proceed unhampered in locations where the Kolmogorov-Smirnov difference is still sufficiently small. In the actual implementation we used an even simpler computational scheme to guarantee the same effect: whenever the evolving distribution hits the KS-boundary the conductance-coefficient c in the region sandwiched between the two flanking minima was set to zero. This halts the smoothing in that region, but allows further reduction in complexity at other locations.

The sole drawback is that the diffusion tends to displace minima, so that for an accurate location it might be worthwhile to locally refit. Alternatively, one can simply pick the location of the actual minimal value (of the original data) in a small neighbourhood of the suggested minimum or trace it back to the original data.

3.3 Confidence Band Based on Kolmogorov-Smirnov Statistic

To implement the rationale underlying eq. (5) and amplified in the preceding section we still need to specify a principled way to determine the amount of acceptable deviation $|G(x) - F_n(x)|$. To this end we introduce the *Kolmogorov-Smirnov* statistic which directly compares distribution functions (eg. see [17]). More precisely, if $F_n(x)$ is the cumulative distribution for a sample of size n drawn from F, the Kolmogorov-Smirnov test-statistic is defined to be the L^∞-distance between the two functions:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{11}$$

for which the p-value can be computed using:

$$P(D_n > \xi) = Q_{KS}(\sqrt{n}\xi), \tag{12}$$

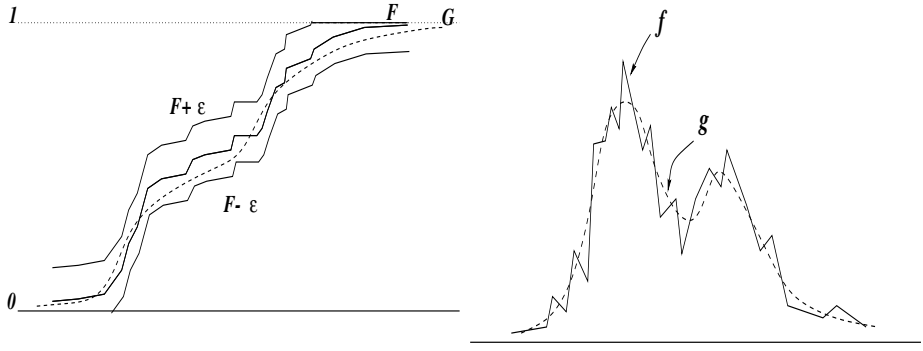

Fig. 1. Segmenting densities (histograms) using the cumulative density. *Left:* The empirical cumulative density F_n flanked by its Kolmogorov-Smirnov confidence bands $F_n \pm \epsilon_n$, together with the smoothed cumulative density G that fits within the band. *Right:* The corresponding densities (obtained by differentiation).

where

$$Q_{KS}(\xi) = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2\xi^2}.$$

(A reference can be found in Mood et.al. [12]). However, the alternating character makes this series expansion rather unwieldy to use, and we therefore hark back to Good [8] who proved the following approximation. First, define the one-sided difference

$$D_n^+ = \sup_x(F_n(x) - F(x)) \qquad \text{and} \qquad D_n^- = \sup_x(F(x) - F_n(x)),$$

then Good showed that under the null-hypothesis (i.e. if F_n does indeed correspond to a sample taken from the underlying distribution F), both statistics D_n^+ and D_n^- are identically distributed and tend to the following asymptottic distribution (for n sufficiently large):

$$4nD_n^{+2} \sim \chi_2^2. \tag{13}$$

This approximation is eminently useful as it provides us with an handle to compute the boundary ϵ_n in eq.(5). More precisely, we pick ϵ_n so that under the null-hypothesis, it is unlikely that the KS-distance exceeds ϵ_n:

$$P(D_n^+ > \epsilon_n) = \alpha \qquad \text{where e.g. } \alpha = 0.05 \text{ or } 0.1. \tag{14}$$

Selecting a critical point c_α for the χ_2^2-distribution such that $P(\chi_2^2 > c_\alpha) = \alpha$ we see that the probability in eq.(14) can be rewritten as $P(4nD_n^{+2} > 4n\epsilon_n^2) = \alpha$ whence $\epsilon_n^2 = c_\alpha/4n$. We therefore conclude that the bound ϵ_n in eq.(5) is determined by

$$\epsilon_n = \frac{1}{2}\sqrt{\frac{c_\alpha}{n}} \quad \text{where} \quad P(\chi_2^2 > c_\alpha) = \alpha. \tag{15}$$

The only point that needs further amplification concerns the fact that we are interested in statistics on the two-sided distance D, whereas eq. (15) yields bounds on the one-sided distances D^+ or D^-. However, since

$$\begin{aligned} P(D > \xi) &= P((D^+ > \xi) \text{ or } (D^- > \xi)) \\ &\leq P(D^+ > \xi) + P(D^- > \xi) \\ &= 2\,P(D^+ > \xi) \end{aligned}$$

Hence, we see that we get a (conservative) confidence bound if we set ϵ_n in eq.(5) to be equal to

$$\epsilon_n = \frac{1}{2}\sqrt{\frac{c_{\alpha/2}}{n}} \quad \text{where} \quad P(\chi^2_2 > c_{\alpha/2}) = \alpha/2. \tag{16}$$

3.4 Comparison to Fitting Gaussian Mixture Models

Fitting a *Gaussian Mixture Model* (GMM) is probably the most popular method to partition a histogram into a unknown number of groups. If the number of clusters is known in advance, one can take recourse to the well-known *Expectation-Maximisation algorithm* (EM) [4] to estimate the corresponding parameters (ie. mean, variance and prior probabilities of each group). However, caution is called for as the sensitivity of the EM-algorithm to its initialisation is well-documented: Initially assigning a small number of "outliers" to the wrong group (albeit with small probability) often lures the algorithm to an erroneous local likelihood minimum, from which it never recovers.

The second problem has to do with the fact that the number of groups isn't known in advance and needs to be determined on the fly. Obviously, maximum likelihood methods are unable to extract the number of clusters as the likelihood increases monotonically with the number of clusters. One possibility, proposed by Carson et.al. [2], is to use a criterion based on *Minimum Description Length* (MDL). The idea is combine the likelihood of the data with respect to a (Gaussian mixture) model with a penalty term that grows with the number of parameters that need to be determined to fit the model. More precisely, for a sample $\mathbf{x}$ of size n they choose the number K of components in the Gaussian mixture (determined by parameters θ) by maximisizing

$$L(\theta \mid \mathbf{x}) - \beta\,\frac{m_K}{2}\log n \tag{17}$$

where m_K is the number of free parameters needed for a model with K Gaussian (d-dimensional) mixture components:

$$m_K = (K-1) + Kd + K\frac{d(d+1)}{2}.$$

(The significance of the β-factor will be discussed presently).

There are two, potentially serious, problems. First, there are the aforementioned problems regarding the instabilities inherent to the EM-algorithm. But

even if the EM-algorithm is successful in identifying the underlying mixture, there is the need for an adhoc factor β to balance out the contribution from both cost-terms in eq.(17), as they may differ by an order of magnitude.

One could of course object that the fudge-factor β is comparable to the parameter α that needs to be fixed in the KS-approach. But there is an important difference: unlike β, the factor α specifying the confidence level has a clear and operational meaning in terms of the risk of committing a type-I error and this risk needs to be fixed in any statistical approach to data-analysis.

In all fairness we need to point out that there is one situation in which the EM-algorithm yields a more satisfying result than the non-parametric approach. Whenever we have two Gaussian densities that encroach on one another, there is a possibility that the global density shows two ill-separated bumps without a clearcut minimum. In such cases EM has little difficulty extracting the individual Gaussians (granted of course, that the number of Gaussians is specified beforehand). As there is no minimum in the original density, our method will have no alternative but to lump the Gaussians together in one cluster.

Having said that, it is also worthwhile to point out that there are situations where EM will fail to deliver the goods while the non-parametric approach has no difficulty whatsoever The simplest example is a uniformly distributed density. In an attempt to come up with a good approximation to this flat density, the EM-algorithm has no other option but to insert a variable number of Gaussians, resulting in a excessive fractioning of the cluster.

In **conclusion** we can say the EM-algorithm for GMM is a typical example of a parametric approach to density estimation. As such it enjoys an advantage over a non-parametric approach (such as the one detailed in this paper) whenever there is clear evidence that the underlying data-distribution is well modeled by the proposed parametrised density. However, in typical image-segmentation problems such an assumption is seldomly warranted and consequently, EM is almost invariably outperformed by the proposed non-parametric histogram segmentation.

4 Some Experimental Results

We also tested this strategy on a number of challenging colour images (see Figs. 2). In keeping with the spirit of our approach we project each image on the axes of a number of different colour-spaces (such as RGB, rgb, and opponent-colours). This yields for each image 9 histograms which are all segmented. The resulting histogram clusterings can easily be scored by marking whether there is more than one cluster (uninteresting) and if so, how well-separated and pronounced these clusters are (e.g. by comparing their mean distance to their variance). In the experiments reported below we display for each image the two most salient histograms. More precisely, the original colour images (left), together with two histograms obtained by projection on an appropriate colour-axis (the choice of which is image dependent) and the resulting image segmentation based on the

segmentation of the histogram. It is clear that combining the information from the different projections often yields very acceptable segmentation results.

To enhance the robustness of the segmentation we apply two simple preprocessing steps:

1. Slight diffusion of the colours in the original image; apart from reducing noise it introduces some sort of spatial correlation into the statistics and therefore compensates for the fact that spatial information is completely lost when mapping pixels into colour-spaces.
2. Global perturbation of the 1-dimensional data by adding independent Gaussian noise to all the datapoints:

$$\tilde{x}_i = x_i + \delta_i$$

where $\delta_i \sim N(0, \sigma^2)$ are independent and the standard deviation σ is taken to be a fraction of the data range R:

$$\sigma = \gamma R \qquad \text{(typically, } \gamma = 0.01\text{)}.$$

The reason for introducing this perturbation is that it resolves ties and removes artifacts due to quantisation, thus improving the final results.

It goes without saying that segmentation based on a single 1-dimensional histogram will only reflect a particular visual aspect (if any at all), and as such only has a very limited range of applicability. However, we contend that as different aspects are highlighted by different histograms, combinations of the regions thus obtained will yield complementary information.

This topic will be taken up in a forthcoming paper but for now, let us just point out that it is helpful to think of the segmentation results for the one-dimensional histograms as some sort of *spatial binding.* If for some feature pixels are mapped into the same region, then they are in effect "bound together" in the sense that, with respect to that particular feature, they are very similar. In this way, each different projection (feature) imposes its own binding-structure on the pixels and pixels that are often "bound together" in the same region therefore accrue a lot of mutual spatial correlation. This spatial correlation structure can be used to improve segmentation or to suggest to the user a number of different possible segmentations, the correlation structure detailing for each of them their statistical support.

5 Extensions to Higher Dimensions

The main thrust of the argument in this paper was based on the Kolmogorov-Smirnov distance, and it is therefore of interest to note that there is multi-dimensional extension of sorts for the KS-statistic. This opens up the possibility to extend this approach to higher dimensions, always bearing in mind of course that the dimension should not be inflated without proper reason.

The generalisation of the distribution function for a d-dimensional stochastic variable $\mathbf{X}$ is straightforward:

$$F(\mathbf{x}) := P(\mathbf{X} \leq \mathbf{x}) = P(X_1 \leq x_1, \ldots, X_d \leq x_d) \tag{18}$$

For the sake of brevity, we limit ourselves here to formulating the relevant theorem (for more details, see [20]):

For any $\epsilon > 0$ there exists a sufficiently large n_0 such that for $n > n_0$ the inequality

$$P\left\{\sup_{\mathbf{x}} |F(\mathbf{x}) - F_n(\mathbf{x})| > \epsilon\right\} < 2e^{-a\epsilon^2 n} \tag{19}$$

holds true, where a is any constant smaller than 2.

Notice how this result falls short of mathematical solidity and elegance enjoyed by the 1-dimensional result (12). First of all, having to deal with an inequality rather than an equality means that we are only given an upperbound for the probability. Furthermore, as stated above, the result is akward to use as it pontificates the existence of an appropriate sample size (n), given a KS-distance ϵ. However, in practice the sample size is fixed in advance and there is little scope for an asymptotic expansion. In fact, for most realistic sample sizes, the specified upper bound is much larger than 1 and therefore of little use.

These theoretical proviso's notwithstanding, there is no good reason why a strategy similar to the one expounded in section 3 cannot be explored in higher dimensions, if we are willing to shoulder a higher computational burden. More specifically we propose the following algorithm to cluster d-dimensional data.

Algorithm Given a sample $\mathbf{x}_1, \ldots, \mathbf{x_n}$ in $\mathbb{R}^d$;

1. Compute for each $\mathbf{x}_i$ the empirical distribution function $F_n(\mathbf{x}_i) = \#\{\mathbf{x}_k \mid \mathbf{x}_k \leq \mathbf{x}_i\}/n$ (the ordering relation is defined component-wise, as in eq.(18)). Next, pick a small initial value for σ;
2. Use eq.(1) to construct the kernel-estimate f_σ for the density. In order to evaluate the KS-statistic we need the corresponding cumulative density F_σ which can be obtained by integration:

$$F_\sigma(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\mathbf{x}} K_\sigma(\xi - \mathbf{x}_i)\, d\xi \tag{20}$$

   If the kernel K_σ is a rotation-invariant Gaussian (2) (actually the most common choice), then its integral can be straightfowardly expressed in terms of products of the error-function `erf(x)`, and (20) therefore yields an explicit expression.
3. Compute the KS-distance between the proposed distribution F_σ and the empirical one supported by the actual data:

$$D(F_\sigma) = \sup_i |F_\sigma(\mathbf{x}_i) - F_n(\mathbf{x}_i)|$$

4. To assess how (un)acceptable this result is we need to compute the p-value for $D(F_\sigma)$, ie. we need to compute the probability that a sample from F_σ will yield a value at least as large as $D(F_\sigma)$. To this end we draw M samples of size n from F_σ and construct for each of them the corresponding empirical $F_n^{(m)}, (m = 1, \ldots, M)$ and the associated distance $D^{(m)}$. Ranking $D(F_\sigma)$ relative to the sequence $\{D^{(m)};\ m = 1, \ldots, M\}$ yields an estimate for the required p-value. (Note that since F_σ is based on a convolution (1), sampling from this distribution is straigthforward: first pick a data-point $\mathbf{x}_i$ at random and next, sample from the Gaussian K_σ centered at $\mathbf{x}_i$.)
5. Finally, if the p-value thus obtained indicates that there is still room to further increase σ (ie. to further smooth f), do so and return to step 2. Notice how we can change σ *globally* (which amounts to a global smoothing), or *locally* at those locations where KS-difference indicates that there is further leeway for data-smoothing. This is the multi-dimensional equivalent of the non-linear smoothing proposed in eq.(10).

6 Conclusion and Outlook

In this paper we have introduced a non-parametric clustering algorithm for *1-dimensional* data. The procedure looks for the *simplest (i.e. smoothest) density that is still compatible with the data.* Compatibility is given a precise meaning in terms of the *Kolmogorov-Smirnov* statistic. This approach is therefore genuinely nonparametric and does not involve fixing arbitrary cost- or fudge-factors.

We have argued that it often makes sense to look for salient regions by investigating projections on appropriate 1-dimensional feature-spaces, which are inspected for evidence of clusters. We note in passing that this provides us with a operational tool for automatic and data-driven selection of promising features: a feature is deemed interesting (for the image under scrutiny) whenever it gives rise to a non-trivial clustering. Finally, we have outlined how the results obtained in the 1-dimensional case can be generalised to higher-dimensional settings.

Acknowledgement The authors gratefully acknowledges partial support by the Belgian Fund for Scientific Research (F.W.O. Vlaanderen), under grant G.0366.98 and KULeuven VIS-project.

References

1. C. Carson, S. Belongie, H. Greenspan, and J. Malik: *Region-Based Image Querying.* Proc. of CVPR'97 Workshop on Content-Based Access of Image and Video Libraries.
2. C. Carson, S. Belongie, H. Greenspan, and J. Malik: *Blobworld: Image Segmentation using Expectation-Maximization and its application to Image Querying.* Submitted to PAMI.
3. G. Coleman and H.C. Andrews: *Image segmentation by clustering.* Proc. IEEE 67, 1979, pp. 773-785.

4. A.P. Dempster, N.M. Laird and D.R. Rubin : *Maximum Likelihood from Incomplete Data via the EM Algorithm.* J. Royal Statist. Soc.Ser B, 39 (1977), pp. 1-38.
5. R.O. Duda and P.E. Hart: *Pattern Classification and Scene Analysis.* Wiley 1973.
6. H. Frigui and R. Krishnapuram: *Clustering by Competitive Agglomeration.* Pattern Recognition, Vol. 30, No. 7, ppe. 1109-1119, 1997.
7. K. Fukunaga: *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.
8. I.J. Good and R.A. Gaskins: *Nonparametric roughness penalties for probability densities.* Biometrika 58, 255-77, 1971.
9. A.K. Jain and R.C. Dubes: *Algorithms for Clustering Data.* Prentice Hall, 1988.
10. Leonard Kaufman and Peter J. Rousseeuw: *Finding Groups in Data: An Introduction to Cluster Analysis.* J. Wiley and Sons, 1990.
11. Brian S. Everitt: *Cluster Analysis.* Edward Arnold, 1993.
12. A. Mood, F. Graybill, D. Boes: *Introduction to the Theory of Statistics.* McGraw-Hill, 1974, 3rd Edition.
13. E.J. Pauwels, P. Fiddelaers and F. Mindru: *Fully Unsupervised Clustering using Center-Surround Receptive Fields with Applications to Colour-Segmentation.* Proc. of the 7th. International Conference on Computer Analysis of Images and Patterns. Kiel, Germany, Sept 10-12, 1997.
14. E.J. Pauwels and G. Frederix: *Non-parametric Clustering for Segmentation and Grouping.* Proc. of International Workshop on Very Low Bitrate Video Coding VLBV'98, Urbana, IL, Oct. 1998. pp. 133-136.
15. E.J. Pauwels and G. Frederix: *Finding Salient Regions in Images. Computer Vision and Image Understanding,* Vol. 75, Nos 1/2, July/August 1999, pp. 73-85.
16. J. Shi and J. Malik: *Normalized Cuts and Image Segmentation.* Proc. IEEE Conf. oon Comp. Vision and Pattern Recognition, San Juan, Puerto Rico, Jun
17. W. Press, B. Flannery, S. Teukolsky, W. Vetterling: *Numerical Recipes.* Cambridge University Press, 1989.
18. J.R. Thompson and R.A. Tapia: *Nonparametric Function Estimation, Modeling and Simulation.* SIAM, 1990.
19. John L. Troutman: *Variational Calculus with Elementary Convexity.* UTM, Springer-Verlag, 1983.
20. V.N. Vapnik: *The Nature of Statistical Learning Theory.* Springer, 1995.

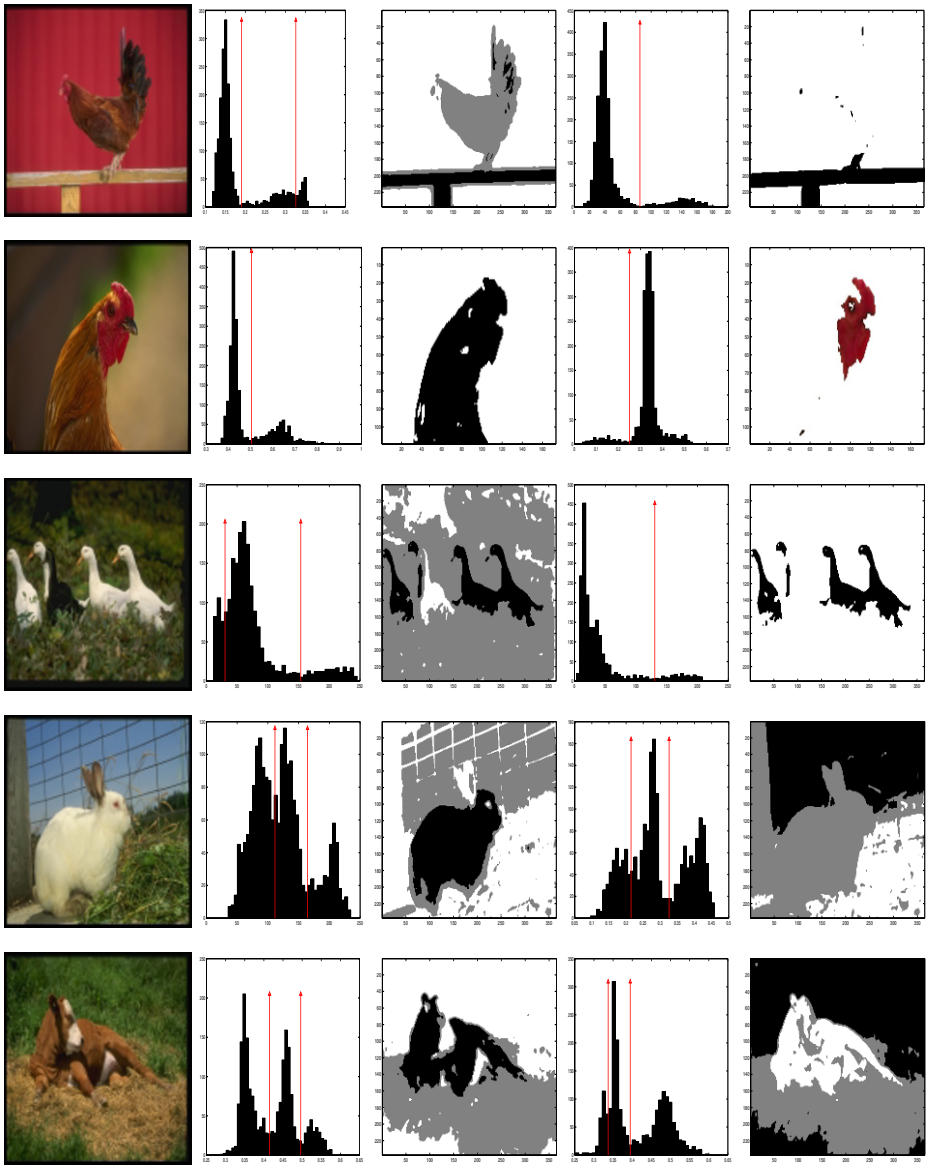

Fig. 2. Original colour images (left), together with two histograms obtained by projection on an appropriate colour-axis (the choice of which is image dependent) and the resulting image segmentation based on the segmentation of the histogram.