Pattern Analysis & Applications (1999)2:22–30

Distribution Free Decomposition of Multivariate Data*

Dorin Comaniciu and Peter Meer

Department of Electrical and Computer Engineering, Rutgers University, USA

Abstract: We present a practical approach to nonparametric cluster analysis of large data sets. The number of clusters and the cluster centres are automatically derived by mode seeking with the mean shift procedure on a reduced set of points randomly selected from the data. The cluster boundaries are delineated using a *k*-nearest neighbour technique. The proposed algorithm is stable and efficient, a 10,000 point data set being decomposed in only a few seconds. Complex clustering examples and applications are discussed, and convergence of the gradient ascent mean shift procedure is demonstrated for arbitrary distribution and cardinality of the data.

Keywords: Convergence; Gradient density estimation; Mean shift procedure; Mode seeking; Nonparametric cluster analysis; Range searching

1. INTRODUCTION

In image understanding, the feature spaces derived from real data most often have a complex structure and *a priori* information to guide the analysis may not be available. The significant featurs whose recovery is necessary for the solution of a task correspond to clusters in this space. The number of clusters, their shape and rules of assignment have to be discerned solely from the given data.

The feature space can be regarded as a sample drawn from an unknown probability distribution. Representing this distribution with a parametric model (e.g. Gaussian mixture) will introduce severe artifacts, since then the shape of the delineated clusters is predefined. Nonparametric cluster analysis, on the other hand, uses the *modes* of the underlying probability density to define the cluster centres and the *valleys* in the density to define the boundaries separating the clusters.

To estimate the probability density, several nonparametric techniques are available: multivariate histogram, the nearest neighbour method and kernel estimation [1–4]. For higher dimensional feature spaces, multivariate histograms are less useful due to their exponentially growing number of bins with the space dimension, as well as due to the artifacts introduced by the quantisation. The nearest neighbour method is prone to local noise (which makes the accurate detection of the modes difficult), and the estimate obtained is not a probability density, since it integrates to infinity [3, p. 96]. For low to medium data sizes, kernel estimation is a good practical choice: it is simple, and for kernels obeying mild conditions the estimate is asymptotically unbiased, consistent in a mean-square sense, and uniformly consisted in probability.

Kernel estimation based clustering essentially relies on two techniques. In the first technique, the underlying density is estimated and a hierarchical data structure is derived, based on which the data is decomposed. An example is the *graph theoretical* approach [1, p. 539]. In the second technique, density gradient estimation [5] is used, the modes being detected with the hill climbing *mean shift* procedure [6].

Both the density and the density gradient estimation require the search for the data points falling in the neighbourhood delineated by the employed kernel. This task is called *multidimensional range searching* [7, p. 373]. However, unlike the nearest neighbour search which can be performed in logarithmic time [8,9], the performance of the multidimensional range searching is difficult to predict for a particular data set [7, p. 385]. Therefore, for applications involving large data sets (e.g. multispectral image segmentation [10], image restoration, speech and image coding), both the kernel estimation and density gradient estimation become computationally expensive, their complexity being proportional to the square of the number of data points. The

Received: 7 October 1998
Accepted: 9 October 1998
* Presented at SPR '98

attempt to reduce computations by subsampling the data leads to inaccuracy, most notably in the tails [11].

As a solution to the problem described above, this paper presents a practical algorithm for unsupervised nonparametric clustering of large data sets. The algorithm is based on the mean shift procedure, being simple, efficient and easy to implement. In section 2 the principles behind the kernel density and the density gradient estimation are reviewed. The specific clustering techniques are discussed in section 3, and the convergence of the mean shift procedure is proved in section 4. The proposed algorithm is presented in section 5, with experimental results shown in section 6.

2. DENSITY AND DENSITY GRADIENT ESTIMATION

Let $\{\mathrm{X}_i\}_{i=1\ldots n}$ be an arbitrary set of n points in the d-dimensional Euclidean space R^d. The *multivariate kernel density estimate* obtained with kernel $K(\mathrm{x})$ and window radius h, computed in the point x is defined as [3, p. 76]

$$\hat{f}(\mathrm{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathrm{x} - \mathrm{X}_i}{h}\right) \qquad (1)$$

The kernel $K(\mathrm{x})$ is a scalar function which must satisfy the following conditions [5]

$$\sup_{\mathrm{x}\in R^d} |K(\mathrm{x})| < \infty, \quad \int_{R^d} |K(\mathrm{x})| d\mathrm{x} < \infty,$$

$$\lim_{\|\mathrm{x}\|\to\infty} \|\mathrm{x}\| K(\mathrm{x}) = 0, \quad \int_{R^d} K(\mathrm{x}) d\mathrm{x} = 1 \qquad (2)$$

where $\|\cdot\|$ is the Euclidean norm. For optimum performance, the window radius h has to be a function of the sample size n. Asymptotic unbiasedness, mean-square consistency and uniform consistency in probability of the density estimate are assured if the following conditions are satisfied, respectively

$$\lim_{n\to\infty} h(n) = 0, \ \lim_{n\to\infty} nh^d(n) = \infty, \ \lim_{n\to\infty} nh^{2d}(n) = \infty \qquad (3)$$

The optimum kernel yielding minimum Mean Integrated Square Error (MISE) is the Epanechnikov kernel

$$K_E(\mathrm{x}) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1 - \mathrm{x}^T\mathrm{x}) & \text{if } \mathrm{x}^T\mathrm{x} < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

where c_d is the volume of the unit d-dimensional sphere [3, p. 76]. Uniform and Gausian kernels are also frequently used. Note that a fast computation of Eq. (1) requires a fast multidimensional range search around x.

The use of a differentiable kernel allows us to define the estimate of the density gradient as the gradient of the kernel density estimate [1]

$$\hat{\nabla} f(\mathrm{x}) \equiv \nabla \hat{f}(\mathrm{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} \nabla K\left(\frac{\mathrm{x} - \mathrm{X}_i}{h}\right) \qquad (5)$$

Conditions on the kernel $K(\mathrm{x})$ and the window radius h to guarantee asymptotic unbiasedness, mean-square consistency and uniform consistency are derived in Fukunaga and Hostetler [5].

For the Epanechnikov kernel (4), the density gradient estimate (5) becomes

$$\hat{\nabla} f(\mathrm{x}) = \frac{1}{n(h^d c_d)} \frac{d+2}{h^2} \sum_{\mathrm{X}_i \in S_h(\mathrm{x})} [\mathrm{X}_i - \mathrm{x}]$$

$$= \frac{n_\mathrm{x}}{n(h^d c_d)} \frac{d+2}{h^2} \left(\frac{1}{n_\mathrm{x}} \sum_{\mathrm{X}_i \in S_h(\mathrm{x})} [\mathrm{X}_i - \mathrm{x}] \right) \qquad (6)$$

where the region $S_h(\mathrm{x})$ is a hypersphere of radius h having the volume $h^d c_d$, centred on x, and containing n_x data points. The last term in Eq. (6)

$$M_h(\mathrm{x}) \equiv \frac{1}{n_\mathrm{x}} \sum_{\mathrm{X}_i \in S_h(\mathrm{x})} [\mathrm{X}_i - \mathrm{x}]$$

$$= \frac{1}{n_\mathrm{x}} \sum_{\mathrm{X}_i \in S_h(\mathrm{x})} \mathrm{X}_i - \mathrm{x} \qquad (7)$$

is called the *sample mean shift*. Using a kernel different from the Epanechnikov kernel results in a weighted mean computation in Eq. (7). Note again that efficient mean shift computation requires efficient range searching.

The quantity $n_\mathrm{x}/(n(h^d c_d))$ is the kernel density estimate $\hat{f}(\mathrm{x})$ computed with hypersphere $S_h(\mathrm{x})$ (the uniform kernel), and thus we can write Eq. (6) as

$$\hat{\nabla} f(\mathrm{x}) = \hat{f}(\mathrm{x}) \frac{d+2}{h^2} M_h(\mathrm{x}) \qquad (8)$$

which yields

$$M_h(\mathrm{x}) = \frac{h^2}{d+2} \frac{\hat{\nabla} f(\mathrm{x})}{\hat{f}(\mathrm{x})} \qquad (9)$$

Expression (9) was first derived by Fukunaga and Hostetler [5], and shows that an estimate of the normalised gradient can be obtained by computing the sample mean shift in a uniform kernel centred on x. The mean shift vector has the direction of the gradient density estimate at x when this estimate is obtained with the Epanechnikov kernel. Therefore, the Epanechnikov kernel is also called the *shadow* of the uniform kernel [6].

Since the mean shift vector always points towards the direction of the maximum increase in the density, it can define a path leading to a local density maximum, i.e. to a mode of the density. The normalised gradient in Eq. (9) introduces a desirable adaptive behaviour: the mean shift step is large for low density regions corresponding to valleys, and descreased as x approaches a mode.

3. DISTRIBUTION FREE CLUSTERING

Associated with the two estimates (the density and its gradient) there are two basic algorithms of nonparametric clustering. For a given window radius h, both algorithms automatically detect the number of existing clusters and their corresponding boundaries.

Using the density estimate (1), a hierarchical structure of the data can be obtained as follows. For each point x_i, search its neighbourhood for a parent X_j, for which the quantity $[\hat{f}(X_j) - \hat{f}(X_i)] \cdot \|X_j - X_i\|^{-1}$ is positive and maximum, i.e. X_j is the steepest uphill from X_i. If the above quantity is negative for all X_j in the neighbourhood, X_i is declared to be a root node of the tree structure. Root nodes are assumed to be close to a mode of the underlying distribution. Clustering is performed in a natural way by following the branches of the structure. The algorithm, called graph theoretical clustering, is described in detail in Fukunaga [1, p. 538]. The hierarchical structure can also be obtained through iterative thresholding [12] or through splitting [13] of the density estimate.

The second algorithm uses the density gradient estimate to define an iterative, hill climbing technique which detects the modes and valleys in the underlying distribution. The *mean shift* procedure is an adaptive steepest ascent technique which computes the mean shift vector [7] for each data point, translates the kernel by that amount, and repeats the computations until a mode is reached. Generalisations of the mean shift procedure and applications to clustering are discussed in Cheng [6]. A variant of the mean shift, called maximum entropy clustering, is presented by Rose et al. [14], and face tracking based on mean shift is described by Bradski [15].

However, clustering through applying the mean shift procedure to each data point cannot be satisfactory in practical applications, since the convergence over the low density regions is poor, while high density regions can present plateaus without a clear local maximum.

When the data set is large (over 10,000 points), the most important drawback of the two algorithms discussed above is their computational complexity. They require the density or density gradient estimation at each data point which has a complexity of $O(n^2)$ for a set of n data points. The complexity problem is induced by the lack of efficiency of the multidimensional range searching. The performance of the d-dimensional trees used in range searching is quite difficult to predict for random data [7, p. 385].

In section 5 we present a probabilistic mean shift type algorithm, which takes in account the difficulties mentioned so far, and whose complexity is $O(mn)$, with $m \ll n$.

4. MEAN SHIFT CONVERGENCE

In this section, we prove that the mean shift procedure applied to discrete data is guaranteed to converge. Let $\{Y_k\}_{k=1,2,\dots}$ denote the sequence of successive locations of the mean shift procedure. By definition, we have for each $k =$ 1, 2, …

$$Y_{k+1} = \frac{1}{n_k} \sum_{X_i \in S_h(Y_k)} X_i \tag{10}$$

where Y_1 is the centre of the initial window and n_k is the number of points falling in the window $S_h(Y_k)$ centred on Y_k.

The convergence of the mean shift has been justified as a consequence of relation (9), (see Cheng [6]). However, while it is true that the mean shift vector $M_h(x)$ has the direction of the gradient density estimate at x, it is not apparent that the density estimate at locations $\{Y_k\}_{k=1,2\dots}$ is a monotonic increasing sequence. Moving in the direction of the gradient guarantees hill climbing only for infinitesimal steps. The following theorem, however, asserts the convergence.

Theorem 1. Let $\hat{f}_E = \{\hat{f}_k(Y_k, K_E)\}_{k=1,2,\dots}$ be the sequence of density estimates obtained using Epanechnikov kernel and computed in the points $\{Y_k\}_{k=1,2\dots}$ defined by the successive locations of the mean shift procedure with uniform kernel. The sequence is convergent.

Proof: Since the data set $\{X_i\}_{i=1\dots n}$ has finite cardinality n, the sequence $\hat{f}_E$ is bounded. It is shown in the appendix that $\hat{f}_E$ is strictly monotonic increasing, i.e. if $Y_k \neq Y_{k+1}$ then $\hat{f}_E(k) < \hat{f}_E(k+1)$, for all $k = 1, 2, \dots$. Being bounded and strictly monotonic increasing, the sequence $\hat{f}_E$ is convergent. Note that if $Y_k = Y_{k+1}$, then Y_k is the limit of $\hat{f}_E$, i.e. Y_k is the fixed point of the mean shift procedure.

5. CLUSTERING ALGORITHM

The steps of the algorithm are described below.

1. *Define a random tessellation of the space with $m \ll n$ spheres $S_h(x)$*. To reduce the computational load, a set of m points $X_1 \dots X_m$ called the *sample set* is randomly selected from the data. Two constraints are imposed on the points retained in the sample set. The distance between any two neighbours should not be smaller than h, the radius of the sphere $S_h(x)$, and the sample points should not lie in sparsely populated regions. The latter condition is required to avoid low density clusters. A region is sparsely populated whenever the number of points inside the sphere is below a threshold T_1. The distance and density constraints automatically determine the size m of the sample set. The spheres centred on the sample set cover most of the data points. When the processing time is not critical, the distance constraint can be relaxed, thus increasing the tessellation resolution.

2. *Apply the mean shift procedure to the sample set.* A set containing *m cluster centre candidates* is defined by the points of convergence of the m mean shift procedures. Note the decrease in computational complexity which is now $O(mn)$, with $m \ll n$, and that the computation of the mean shift vectors is based almost on the entire data set. Therefore, the quality of the density gradient estimate is not diminished by the use of sampling.

3. *Perturb the cluster centre candidates and reapply the mean shift procedure*. Since a local plateau can prematurely stop the iterations, each cluster centre candidate is perturbed by a random vector of small norm, and the mean shift procedure is let to converge again.

4. *Derive the cluster centres $Y_1 \dots Y_p$ from the cluster centre candidates*. Any subset of cluster centre candidates which

are sufficiently close to each other (for any given point in the subset there is at least another point in the subset such that their distance is less than h), defines a *cluster centre*. The cluster centre is the mean of the cluster centre candidates in the subset. Note that $p \leqslant m$.

5. *Validate the cluster centres*. Between any two cluster centres Y_i and Y_j a significant valley should occur in the underlying density. The existence of the valley is tested for each pair (Y_i, Y_j). The sphere $S_h(x)$ is moved with step h along the line defined by (Y_i, Y_j), and the weighted number the data points lying in the sphere is counted at each position, i.e. the density is estimated with Epanechnikov kernel K_E along the line. Whenever the ratio between min[$\hat{f}(X_i)$, $\hat{f}(X_j)$] and the minimum density along the line is larger than a threshold T_2, a valley is assumed between Y_i and Y_j. If no valley was found between Y_i and Y_j, the cluster centre of lower density (Y_i or Y_j) is removed from the set of cluster centres.

6. *Delineate the clusters*. At this stage, each sample point is associated with a cluster centre. To allocate the data points a k-nearest neighbour technique is employed, i.e. each data point belongs to the cluster defined by the majority of its k-nearest sample points.

6. PERFORMANCE EVALUATION

The clustering algorithm makes use of three parameters: the searching sphere radius h which controls the sensitivity of the decomposition, the threshold T_1, which imposes the density constraint, and the threshold T_2, corresponding to the minimum acceptable peak-valley ratio. The parameters T_1 and T_2 generally have a weak influence on the result. All the experimental results described here were obtained with $T_1 = 50$ and $T_2 = 1.2$. Unless it is specified otherwise, we used $k = 1$, i.e. clusters were delineated using the nearest neighbour for the last step of the algorithm. Since the experimental data sets had different scales, the sphere radius h has been changed accordingly. Note also that the Improved Absolute Error Inequality [16] was employed to efficiently compute Euclidean distances.

Experiment 1

The first example is shown in Fig. 1. The data set contained 32,640 points with dimension $d = 3$, grouped into three non-linearly separable clusters (Fig. 1(a)). A standard unsupervised procedure such as ISODATA [17] would fail on this data. Using a radius $h = 0.2$, the sample set obtained

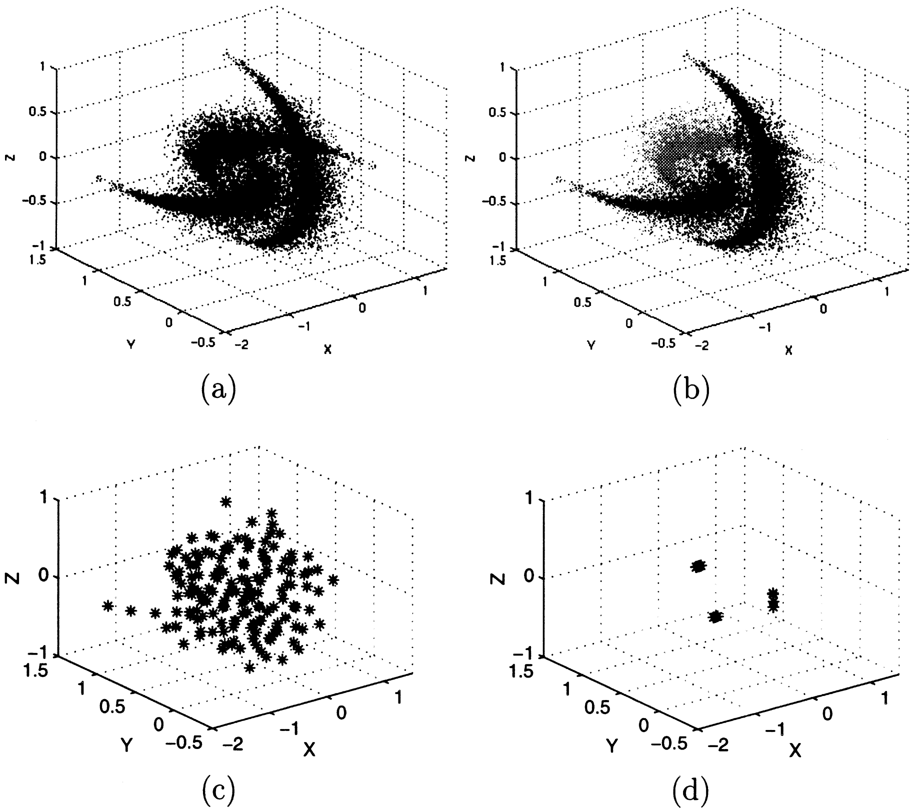

Fig. 1. First experiment. (a) Original data set (32,640 points); (b) cluster delineation (three clusters represented with different grey levels); (c) sample set (167 points); (d) cluster centre candidates.

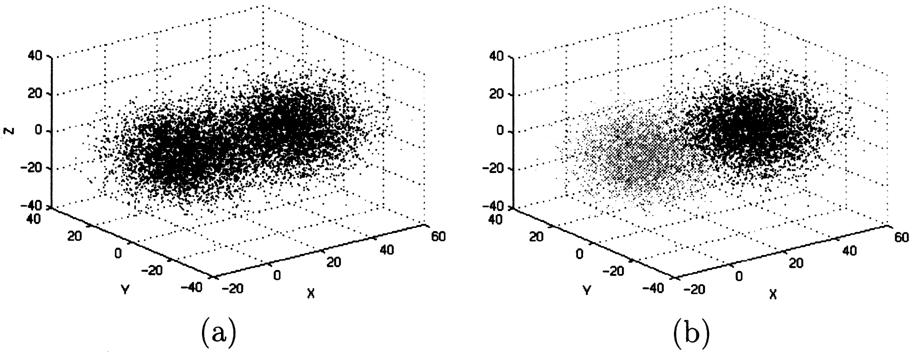

Fig. 2. Second experiment. (a) Original data set (10,000 points); (b) cluster delineation (two clusters).

had 167 points (Fig. 1(c)) and converged to three cluster centres. In Fig. 1(b), the three extracted clusters are shown, having 11,050, 10,874 and 10,716 points, respectively. The algorithm running time was less than 10 seconds on a standard workstation.

Experiment 2

A simpler clustering example is shown in Fig. 2. The purpose of this experiment was to compare the performance of the nonparametric algorithm with the performance of the classical Bayes classifier. The data set contained 10,000 points with dimension $d = 3$ coming from two normal distributions with covariance 10^2I and mean vector $(0, 0, 0)^T$ and $(40, 0, 0)^T$, respectively.

Figure 2(b) shows the delineated clusters corresponding to a radius $h = 10$. Using the Bayes classifier the error rate is 2.34%, due to 234 points that overlap. Figure 3 shows the error rate resulted from our algorithm for sphere radii between 8 and 20. The allocation of data points to the modes used k-nearest neighbours, where k was taken as 1 and 3, respectively. The error rate increases with radius h due to the increase in the boundary delineation error. The straight line in the graph represents the Bayes error rate. The performance of the nonparametric algorithm is very close to that of the Bayes classifier, in spite of a no *a priori* knowledge being used in the nonparametric case.

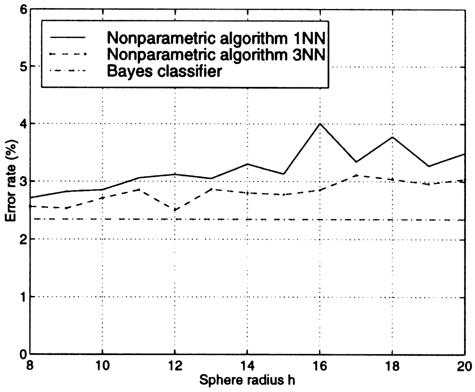

Fig. 3. The error rate of the proposed algorithm for different values of the sphere radius and different number of nearest neighbours. The dash-dotted line represents the error rate of the Bayes classifier.

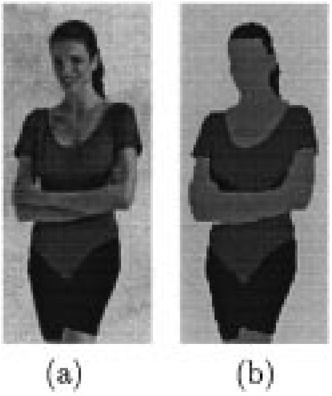

Fig. 4. Third experiment. (a) Original image; (b) segmented image using nonparametric clustering.

Experiment 3

The third experiment shows the application of the new algorithm to the segmentation of the colour image in Fig. 4(a). Clustering is performed in the perceptually uniform $L^*u^*v^*$ colour space, each delineated cluster corresponding to homogeneous regions in the image. The colour space (Fig. 5(a)) contained 14,826 points, and four clusters have been extracted by using a radius of $h = 10$. Note the irregular boundaries of the clusters in Fig. 5(b). The clustering quality can be assessed by observing the segmented image in Fig. 4(b), where spatial constraints have been used to remove small regions containing less than 25 pixels [10].

We tested the stability of the algorithm by using different sets of sample points, each set resulting in a distinct tessellation of the input space. Four values of the window radius h have been considered: 4, 7, 16 and 22. Ten trials have been performed for each window radius. The algorithm proved to be very stable producing similar mode locations and cluster delineations for a given radius value. Table 1 shows the number of detected clusters corresponding to each radius class.

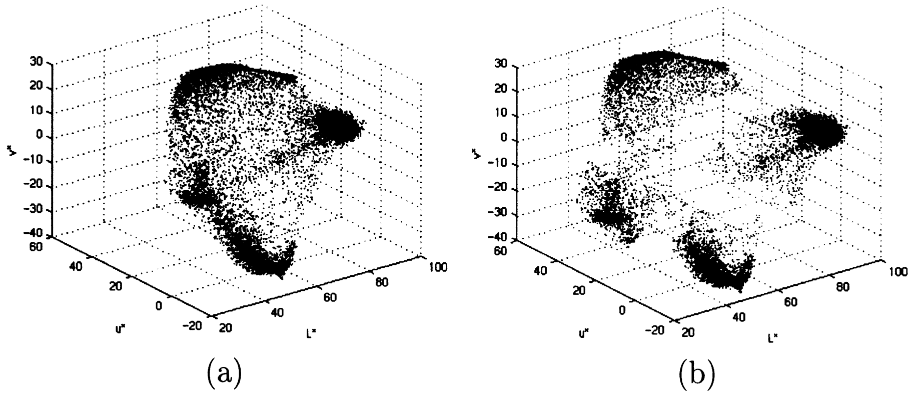

Fig. 5. Third experiment. (a) Original data set (14,826 colour points); (b) cluster delineation (four clusters). The position of each cluster has been shifted to show the delineation.

Table 1. Number of detected clusters versus the sphere radius for Experiment 3

Detected clusters	5	4	4	3
Sphere radius	4	7	16	22

Experiment 4

A second colour segumentation example is presented in Fig. 6. Using the same radius of $h = 10$, the algorithm extracted three colour clusters.

Experiment 5

A difficult data set is shown in Fig. 7(a). it contains 17,748 points and represents the first two components of the $L^*u^*v^*$ space of the colour image in Fig. 8(a). We used only this subspace to be able to visualise the behaviour of the algorithm. Large amounts of background noise, asymmetric clusters, narrow peaks and large plateaus are present. Real data often have such a complex structure.

Using a radius of $h = 5$, the proposed algorithm detected seven clusters (Fig. 7(b)). The 47 sample points are shown in Fig. 7(c), together with the Epanechnikov estimate of the density. The estimate was computed with a resolution of one on both axes and using the same window radius of $h = 5$. The sample points converged to the cluster centre candidates (Fig. 7(d)) located at the local maxima of the density estimate. The valley test further removed some of the cluster centre candidates located on plateaus, allowing the correct cluster delineation.

The decomposition in Fig. 7b does not have a physical meaning, since it is based only on two dimensions. However, when the colour image is processed in the full (3-dimensional) colour space the obtained segmentation is satisfactory (see Fig. 8(b)).

7. DISCUSSION

Under general conditions, the use of the algorithm has to be preceded by an application-dependent preprocessing stage to normalise the data. When no *a priori* information is available, the optimal window radius can be obtained as the centre of the largest operating range, which yields the same number of clusters for a given data [1, p. 541]. In practice, however, the final objective of the decomposition is task dependent, therefore, top-down information controls the window radius.

The new algorithm has been applied with excellent results to the task of real-time segmentation of medical images

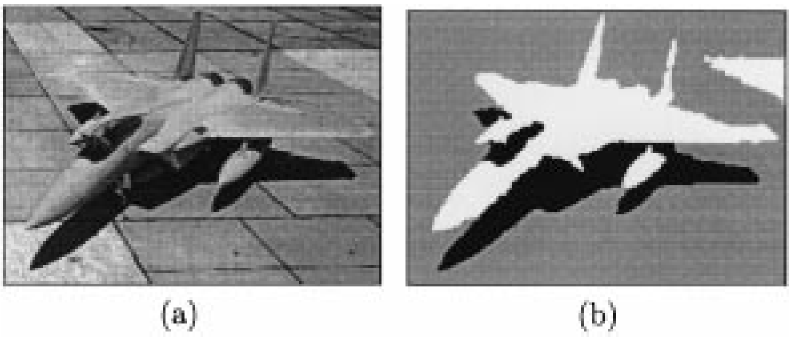

Fig. 6. Fourth experiment. (a) Original image; (b) segmented image using nonparametric clustering.

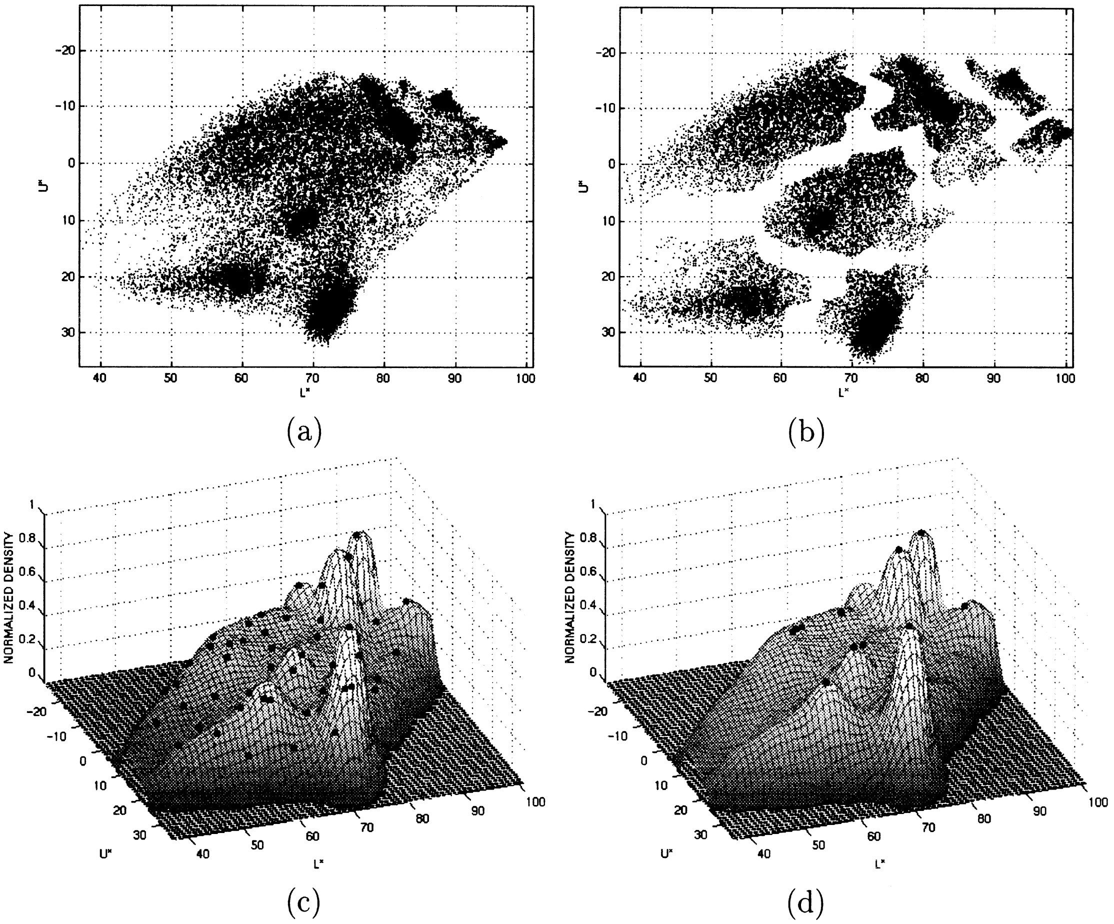

Fig. 7. Fifth experiment. (a) Original data set (17,748 2D points); (b) data decomposition (seven clusters). The position of each cluster has been shifted to show the delineation; (c) sample set (47 points) and the Epanechnikov density estimate; (d) cluster centre candidates.

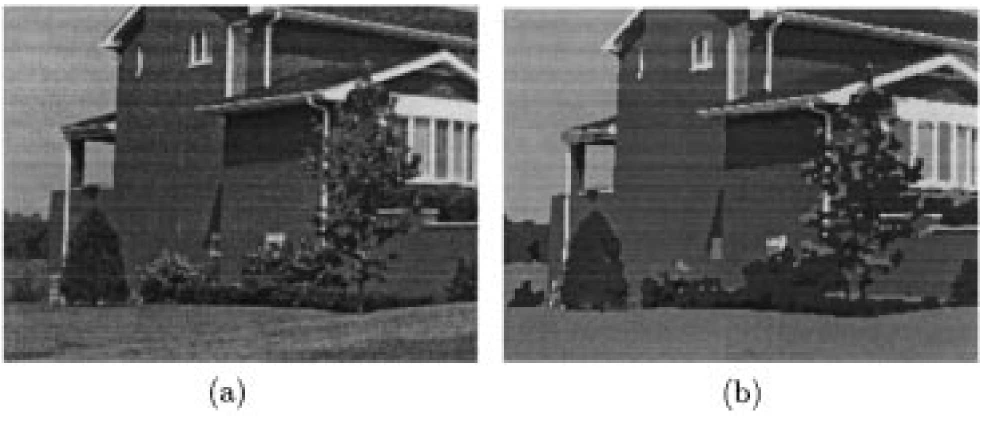

Fig. 8. The colour image used in the fifth experiment. (a) Original; (b) segmented.

in a retrieval system for diagnostic pathology [18]. The nonparametric nature of the algorithm and its robustness to noise allowed the use of a fixed radius for the processing of hundreds of digital specimens captured under different conditions.

Despite the expected difficulty given the reduced number of data points, the proposed algorithm showed good performance for the standard IRIS data, which contains 150 points. Using a window radius of 4, letting all the individual points seek the closest mode, and delineating the clusters based on the three nearest neighbours, the correct number of three clusters were detected. Only seven points were misclassified, a result which is superior to that of seven other clustering algorithms compared by Bajcsy and Ahuja [19]. When the cluster delineation was based on the nearest neighbour, the number of misclassified points increased to eight.

Acknowledgements

The authors were supported by the NSF under the grant IRI–9530546.

References

1. Fukunaga K. Introduction to Statistical Pattern Recognition, Academic Press, 1990
2. Scott DW. Multivariate Density Estimation, Wiley, 1992
3. Silverman BW. Density Estimation for Statistics and Data Analysis, Chapman & Hall, 1986
4. Terrell GR, Scott DW. Variable density estimation. Ann Statistics 1992;20:1236–1265
5. Fukunaga K, Hostetler LD. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Info Theory 1975;21:32–40
6. Cheng Y. Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Machine Intell 1995;17:790–799
7. Sedgewick R. Algorithms in C++, Addision-Wesley, 1992
8. Friedman JH, Bentley JL,Finkel RA. An algorithm for finding best matches in logarithmic expected time. ACM Trans Mathematical Software 1977;3:209–226
9. Nene SA, Nayar SK. A simple algorithm for nearest neighbor search in high dimensions. IEEE Trans Pattern Anal Machine Intell 1997;19:989–1003
10. Comaniciu D, Meer P. Robust analysis of feature spaces: color image segmentation. Proc IEEE Conf on Comp Vis and Pattern Recognition, Puerto Rico, 1997, pp 750–755
11. Popat K, Picard RW. Cluster-based probability model and its application to image and texture processing. IEEE Trans Image Process 1997;6(2):268–284
12. Herbin M, Bonnet N, Vautrot P. A clustering method based on the estimation of the probability density function and on the skeleton by influence zones. Pattern Recognition Letters 1996; 17: 1141–1150
13. Garica JA, Valdivia JF, Cortijo FJ, Molina R. A dynamic approach for clustering data. Signal Processing 1995;44:181–196
14. Rose K, Gurewitz E, Fox GC. Constrained clustering as an optimization method. IEEE Trans Pattern Anal Machine Intell 1993;15:785–794
15. Bradski GR. Computer vision face tracking as a component of a perceptual user interface. Proc IEEE Workshop on Applications of Computer Vision. Princeton, 1998, pp 214–219
16. Pan JS, McInnes FR, Jack MA. Fast clustering algorithms for vector quantization. Pattern Recognition 1996;29:511–518
17. Jain AK, Dubes RC. Algorithms for Clustering Data, Prentice Hall, 1988
18. Comaniciu D, Meer P, Foran D, Medl A. Bimodal system for interactive indexing and retrieval of pathology images. Proc IEEE Workshop on Applications of Computer Vision, Princeton, 1998, pp 76–81
19. Bajcsy P, Ahuja N. Location- and density-based hierarchical clustering using similarity analysis. IEEE Trans Pattern Anal Machine Intell 1998;20:1011–1015

Dorin Comaniciu received the Dipl. Engn. and D.Sc. degrees in electrical engineering from the Polytechnic University of Bucharest, Romania, in 1988 and 1995, respectively. From 1988 to 1990 he was with ICE Felix Computers, Romania. Between 1991 and 1995 he was a teaching assistant at the Polytechnic University of Bucharest. He held research appointments in Germany (multimedia processing) and France (image compression). He is currently completing the Ph.D. degree in the Department of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. His research interests include nonparametric robust methods for computer vision, content-based image/video retrieval, and data compression.

Peter Meer received the Dipl. Engn. degree from the Bucharest Polytechnic Institute, Bucharest, Romania in 1971, and the D.Sc. degree from the Technion, Israel Institute of Technology, Haifa, Israel, in 1986, both in electrical engineering. From 1971 to 1979 he was with the Computer Research Institute, Cluj, Romania, working on R&D of digital hardware. Between 1986 and 1990 he was Assistant Research Scientist at the Center for Automation Research, University of Maryland at College Park. in 1991 he joined the Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ and is currently an Associate Professor. He held visiting appointments in Japan, Korea, Sweden and Israel and was on the organizing committees of several international workshops and conferences. His research interest is in application of modern statistical methods to image understanding problems.

Correspondence and offprint requests to: P. Meer, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08855, USA. Email: meer@caip.rutgers.edu

APPENDIX

In this section, we will show that the sequence $\hat{f}_E = \{\hat{f}_k(\mathrm{Y}_k, K_E)_{k=1,2\ldots}$ is strictly monotonic increasing, i.e. if $\mathrm{Y}_k \neq \mathrm{Y}_{k+1}$, then $\hat{f}_E(k) < \hat{f}_E(k+1)$, for all $k = 1, 2, \ldots$.

Let n_k, n'_k and n''_k with $n_k = n'_k + n''_k$ be the number of data points falling in the d-dimensional windows (see Fig. 9).

$$S_h(\mathrm{Y}_k),\ S'_h(\mathrm{Y}_k) = S_h(\mathrm{Y}_k) - S''_h(\mathrm{Y}_k)$$
$$\text{and } S''_h(\mathrm{Y}_k) = S_h(\mathrm{Y}_k) \cap S_h(\mathrm{Y}_{k+1})$$

Without loss of generality, we can assume the origin located at Y_k. Using the definition of the density estimate (1) with the Epanechnikov kernel (4), and noting that $\|\mathrm{Y}_k - \mathrm{X}_i\|^2 = \|\mathrm{X}_i\|^2$, we have

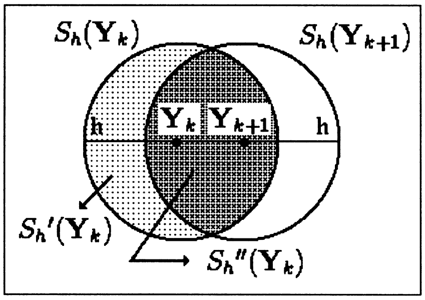

Fig. 9. d-dimensional windows used in the proof of convergence: $S_h(\mathrm{Y}_k)$, $S'_h(\mathrm{Y}_k) = S_h(\mathrm{Y}_k) - S''_h(\mathrm{Y}_k)$ and $S''_h(\mathrm{Y}_k) = \mathrm{S}_h(\mathrm{Y}_k) \cap \mathrm{S}_h(\mathrm{Y}_{k+1})$. The point Y_{k+1} is the mean of the data points falling in $S_h(\mathrm{Y}_k)$.

$$\hat{f}_E(k) = \hat{f}_k(\mathrm{Y}_k, K_E) = \frac{1}{nh^d} \sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} K_E \left(\frac{\mathrm{Y}_k - \mathrm{X}_i}{h} \right)$$

$$= \frac{d+2}{2n(h^d c_d)} \sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} \left(1 - \frac{\|\mathrm{X}_i\|^2}{h^2} \right) \qquad \text{(A1)}$$

Since the kernel K_E is nonnegative, we also have

$$\hat{f}_E(k+1) = \hat{f}_{k+1}\,(\mathrm{Y}_{k+1}, K_E)$$

$$\geq \frac{1}{nh^d} \sum_{\mathrm{X}_i \in S''_h(\mathrm{Y}_k)} K_E \left(\frac{\mathrm{Y}_{k+1} - \mathrm{X}_i}{h} \right) \qquad \text{(A2)}$$

$$= \frac{d+2}{2n(h^d c_d)} \sum_{\mathrm{X}_i \in S''_h(\mathrm{Y}_k)} \left(1 - \frac{\|\mathrm{Y}_{k+1} - \mathrm{X}_i\|^2}{h^2} \right)$$

Hence, knowing that $n'_k = n_k - n''_k$, we obtain

$$\hat{f}_E(k+1) - \hat{f}_E(k) \geq \frac{d+2}{2n(h^d c_d)h^2}$$

$$\left[\sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} \|\mathrm{X}_i\|^2 - \sum_{\mathrm{X}_i \in S''_h(\mathrm{Y}_k)} \|\mathrm{Y}_{k+1} - \mathrm{X}_i\|^2 - n'_k h^2 \right] \qquad \text{(A3)}$$

where the last term appears due to the different summation boundaries.

Also, by definition $\|\mathrm{Y}_{k+1} - \mathrm{X}_i\|^2 \geq \mathrm{h}^2$ for all $\mathrm{X}_i \in S'_h(Y_k)$, which implies that

$$\sum_{\mathrm{X}_i \in S'_h(\mathrm{Y}_k)} \|\mathrm{Y}_{k+1} - \mathrm{X}_i\|^2 \geq n'_k h^2 \qquad \text{(A4)}$$

Finally, employing Eq. (A4) in Eq. (A3), and using Eq. (10), we obtain

$$\hat{f}_E(k+1) - \hat{f}_E(k)$$

$$\geq \frac{d+2}{2n(h^d c_d)h^2} \left[\sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} \|\mathrm{X}_k\|^2 - \sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} \|\mathrm{Y}_{k+1} - \mathrm{X}_i\|^2 \right]$$

$$= \frac{d+2}{2n(h^d c_d)h^2} \left[2\mathrm{Y}_{k+1}^T \sum_{\mathrm{X}_i \in S_h(\mathrm{Y}_k)} \mathrm{X}_i - n_k \|\mathrm{Y}_{k+1}\|^2 \right] \qquad \text{(A5)}$$

$$= \frac{d+2}{2n(h^d c_d)h^2}\, n_k \|\mathrm{Y}_{k+1}\|^2$$

The last item of the relation (A5) is strictly positive, except when $\mathrm{Y}_k = \mathrm{Y}_{k+1} = 0$.