Clustering with KDEs: Art Historical and Archaeological Applications

C.C. Beardah[1], S. Porcinai[2] and M.J. Baxter[1]

[1] School of Science, The Nottingham Trent University, Clifton,
Nottingham NG11 8NS, United Kingdom.
christian.beardah@ntu.ac.uk and michael.baxter@ntu.ac.uk.
[2] Instituto di Ricerca sulle Onde Elettromagnetiche "Nello Carrara,"
Consiglio Nazionale delle Ricerche, Florence, Italy.
porcinai@iroe.fi.cnr.it.

Abstract. At previous CAA conferences we have reported on the use of Kernel Density Estimates (KDEs) for data display in up to three dimensions. Issues of sample size and the associated "curse of dimensionality" have influenced much of our previous work. However, in this paper we investigate the possibilities of the methodology in cases where there is no shortage of data. In particular, we present a method of cluster identification based upon contouring two- or three-dimensional KDEs. If desired, this approach can be used as the starting point for more formal methods of cluster analysis. Our main application involves the analysis of data generated using image spectroscopy on works of art, though the utility of the technique for archaeological applications should be clear. Our aim is to identify non-homogeneity, represented by clusters, which could possibly be explained by forgery or earlier restoration.
Keywords: Kernel Density Estimation; multivariate analysis; cluster analysis; image spectroscopy.

1 Introduction

When analysing multivariate statistical data of the kind that often arises in archaeology, archaeometry and related areas, it is almost always useful to apply a battery of techniques, rather than one method in isolation. At previous CAA conferences we have reported on the use of Kernel Density Estimates (KDEs) for data display in up to three dimensions. (See Baxter et al., 1997, Beardah and Baxter, 1996, 1999 for examples.) KDEs provide a simple method of looking for interesting structure within a data set, often characterised by non-normality and very often represented by groupings within the data that may have some archaeological or other significance. Due to the "curse of dimensionality" and the usual problems presented by displaying high dimensional data, KDEs are of most utility when the data is of dimension three or less. However, higher dimensional data are often analysed by subjecting them to some dimension reduction technique such as principal component analysis (PCA). KDE methods can then be applied to the first one, two or three components of the PCA scores in order to identify structure. Provided that the first few PCs explain a high percentage of the variation in the original data, there is every reason to suppose that the original data also exhibits this structure. Of course, this combination of PCA and KDE techniques is rather informal and exploratory in nature. On the other hand, there is no reason why the potential clusters exhibited by such methodology should not be used as the starting point for more formal clustering techniques.

The general idea of this suggested methodology is outlined below.

1. Data collection. For the purposes of this paper we assume that the data are high dimensional.
2. Application of a dimension reducing technique such as PCA.
3. Visual analysis of the first one, two or three PCs using contouring (based upon KDEs) in an attempt to identify clusters of similar objects.
4. If desirable, use of these clusters as the starting point for more formal clustering techniques.

In the main application discussed here, the use of contouring in point 3 above is vital. The reason is that data sets generated using image spectroscopy on works of art tend to consist of a very large number of observations. This usually results in traditional PCA plots that are far too dense to interpret, even if there is clear structure in the data.

Of course, the really interesting questions are often post-analysis when we ask whether the clusters mean anything in the original context.

1.1 KDEs for two- and three-dimensional data

The interested reader is referred to Silverman, 1986 and Wand and Jones, 1995 for excellent general introductions to the subject of KDEs. For a less mathematical, more archaeological treatment, see Baxter et al., 1997 and Beardah and Baxter, 1996. Archaeological applications of two- and three-dimensional KDEs are also the subject of Beardah, 1998 and Beardah and Baxter, 1999.

Figures 1 and 2 show examples of KDEs based upon bivariate and trivariate data respectively. Note that for bivariate data, the KDE is three-dimensional, while contours are two-dimensional. Analogously, for trivariate data, the KDE is four-dimensional, while contours are three-dimensional. In both cases, contouring is based upon an idea by Bowman and Foster, 1993 whereby contours enclose the p% most dense of the data.

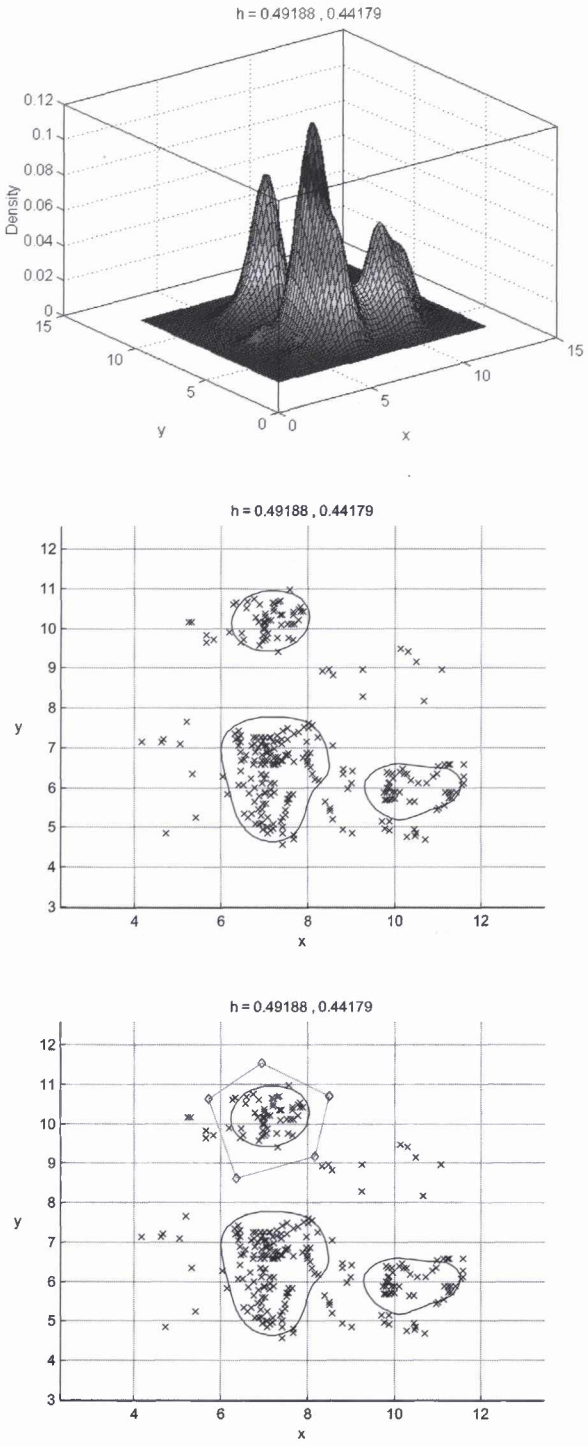

Fig. 1. For data representing the spatial locations of n=276 bone splinters (see Binford, 1978 and Blankholm, 1991), from the top: (a) a bivariate KDE; (b) a scatter plot overlaid with a 75% contour, (c) a polygon enclosing one section of the 75% contour

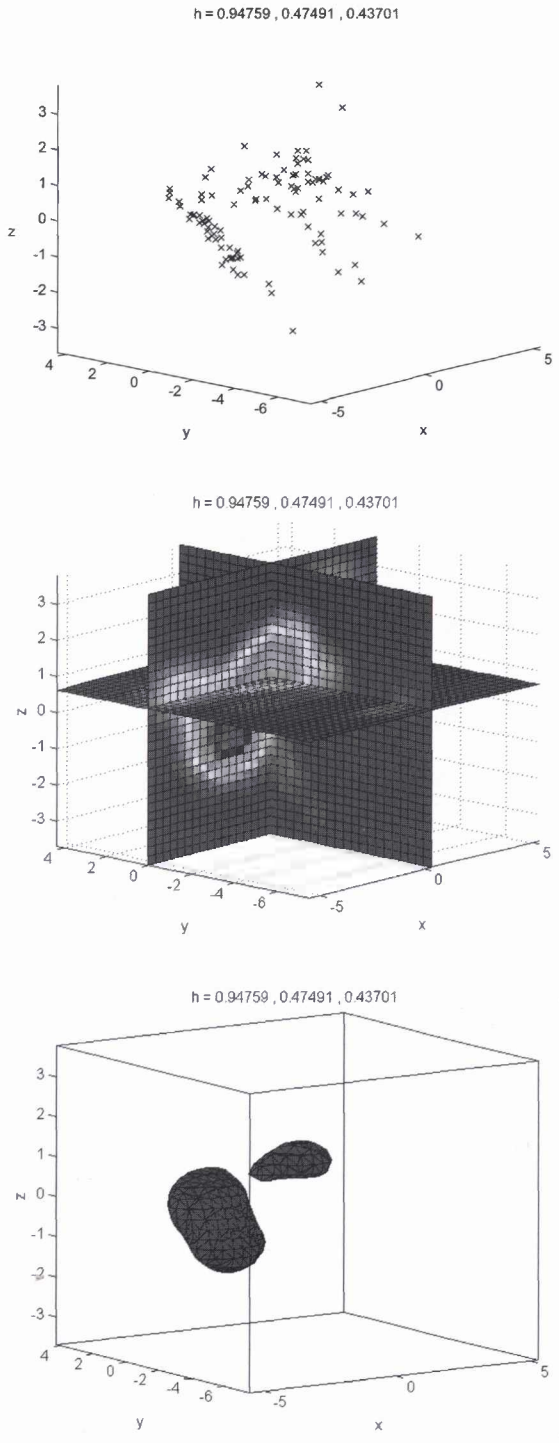

Fig. 2. For the first three PCs of data representing the chemical composition of n=105 specimens of Romano-British waste glass, from the top: (a) a scatter plot of the data; (b) three-dimensional slices through the trivariate KDE and (c) a 40% contour shell. (These data are extensively analysed in Baxter, 1994.)

It should be clear from Figures 1 and 2 how simple it is to use contours based upon KDEs as an informal clustering technique. In order to do so, we need to identify the data points that lie within a certain contour. Mathematically, this is not quite as simple as it may appear. For example, consider Fig. 1b. The 75% contour is split into three distinct parts enclosing separate subsets of the data. By identifying the data points with density higher than that defining the 75% contour level, we can easily identify *all* data points lying inside the *union* of these three contours. However, if we wish to separately identify the data points lying inside the topmost contour of Fig. 1b, for example, it is less obvious how this could be done, especially for large data sets where identifying each data point visually is not an option. The answer is to define a polygon surrounding the contour of interest and then to find data points that both lie inside the polygon *and* have density higher than that defining the 75% contour level. Fig. 1c shows this process. This technique, though more difficult to implement, applies equally to the case of three-dimensional data.

1.2 Clustering using Mahalanobis distance

In the previous section we have seen that KDEs and contouring can be used to identify possible groupings within a data set. This technique could be applied to either raw data of low dimensionality, or to the first two or three PCs of data with high dimensionality. If the clusters are clear and well defined, then the preliminary analysis may end at this stage. However, if the clustering is less clear, then it may be worthwhile to use the informal clustering given by the method of the section 1.1 as the starting point for a more formal technique of cluster analysis. Such a technique is outlined in this section.

The clustering method we have used is largely based upon the Mahalanobis distance between a point, $\underline{x}$, in p-dimensional space, and a group with mean $\underline{y}$ and p by p covariance matrix Σ. This is given by

$$d^2_{xy} = (\underline{x} - \underline{y})^T \Sigma (\underline{x} - \underline{y})$$

where $\underline{x}$ and $\underline{y}$ are represented by p by 1 vectors.

In simple terms, new objects are added to an existing cluster if they are "close" to the centre of the cluster. Clusters are iteratively "grown" from an initial clustering which can be as small as just one object. (Contouring with KDEs will provide this initial clustering.) It should be noted that, in order for the Mahalanobis distance to be used, the cluster size, n_g, needs to be greater than the number of variables, p and preferably $n_g >> p$. As we shall see in the next section, this restriction is not a problem for our main application, though it may be for archaeological applications where sample sizes are typically "small". Beier and Mommsen, 1994, extend the basic idea discussed above to include uncertainties in data measurement and "dilution" effects.

2 Analysis of works of art using image spectroscopy

2.1 Background and introduction

Our main application of the methodology described above will be in the analysis of data generated using image spectroscopy on works of art. Data representing the reflectance spectrum for a fine mesh of points covering the surface of a painting can be subjected to PCA in an effort to identify non-homogeneity, represented by clusters. The first three PCA components typically explain 95-99% of the variation in data of this type, and it is therefore natural to explore the structure exhibited by the PCs using KDE techniques. If desired, potential clusters so identified can be used as the starting point for more formal clustering techniques such as that outlined in section 1.2. Post-analysis, elements within identified clusters must be mapped back to spatial locations on the original image.

One of the most important objectives in the scientific investigations of paintings (or more general works of art) is the characterisation of the materials constituting the object under examination. This kind of information is useful, for example, in building knowledge of the painting technique used by the artist and for monitoring the status of conservation of the work of art.

Non-destructive and non-invasive techniques should be the first step in the investigation of works of art. Such methods provide analytical information that is often less precise than so-called invasive analytical techniques, which on the other hand need sampling operations. The application of non-invasive techniques is recommended in order to obtain information which could be used to guide and limit the micro sampling required by invasive methods. Image Spectroscopy (IS) techniques (see Baronti et al., 1998, for further details) are particularly suitable to this purpose not only because they extend the measurement domain to a wide area of the painting, but also because the technique provides data that retain the visual representation of the object.

For this paper, data associated with two paintings has been used, see Fig. 3. The first of these is a test tablet painted with four known pigments. For each pigment three rectangular strips were homogeneously painted using (i) the pure pigment, (ii) a mixture of the pure pigment and 5% carbon black by weight, (iii) a mixture of the pure pigment and 10% carbon black by weight. Twelve zones of the same size were thus created differentiated by their dominant wavelength or their brightness resulting from the percentage of carbon black, see Fig. 3a.

The second painting used is a scene from the *Holy Trinity* Predella (on exhibit at the Uffizi gallery), painted by Luca Signorelli in the early sixteenth century. It is an oil-painted panel of 32 cm by 204 cm, divided into three scenes of Jesus's Passion. This study concentrates upon the flagellation scene, see Fig. 3b.

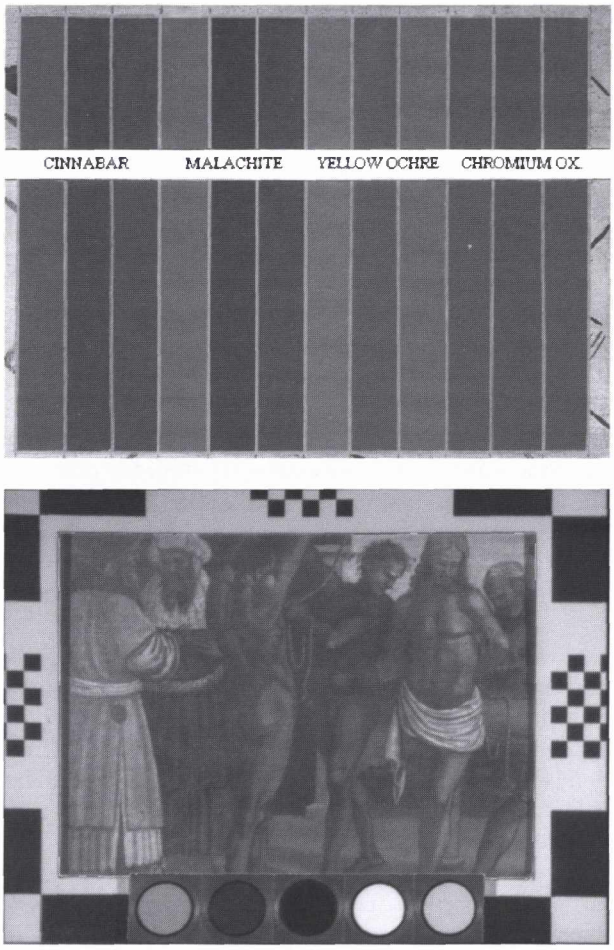

Fig. 3. (a) test tablet, (b) Signorelli's Predella, flagellation scene.

2.2 The data

For each painting, the data were collected using a vidicon camera (with a PbO-PbS detector). The surface of the painting is illuminated by two projectors (150W quartz tungsten halogen lamps) oriented symmetrically at 45 degrees with respect to the painting's plane. Wavelength selection is achieved by means of a set of interferential filters with bandwidth of 10nm. The filters are lodged in eight-slot filter wheels that are placed one at a time, on a wheel drive activated and timed by a personal computer. The acquisition of each image is the result of a real-time average on 100 frames. Again, see Baronti et al., 1998, for full details.

The data set is a sequence of 29 images acquired at the following wavelengths (nm): 420 through 700 in steps of 20; 750 through 1200 in steps of 50; 1300, 1400, 1450, 1550. Each image is characterised by a column of values in the range 0 to 255, one entry per pixel, which represent the grey scale values at each pixel (0 - black, to 255 - white). The grey scale levels of a pixel, across the whole sequence of the image, are the reflectance spectrum of that pixel. So following the usual statistical notation the objects are the pixels (around 86297 in the case of the test tablet and 237300 in the case of Signorelli's Predella) and the variables the wavelengths investigated. In the case of the Predella, this results in a data matrix of 237300 rows and 29 columns, where each entry of the matrix is a whole number in the range 0 to 255.

Due to the high dimensionality of the data set, it is natural to apply PCA to compress this information (usually two wavelengths close to each other are highly correlated) into fewer variables. Tables 1 and 2 show the amount of variation explained by the first 4 PCs in the case of the standardised test tablet data and the standardised Predella data sets respectively.

Table 1. PCA based on standardised test tablet data (86297 by 29)

	Variance	Cumulative Variance
PC1	64.6%	64.6%
PC2	20.0%	84.6%
PC3	9.7%	94.3%
PC4	4.0%	98.3%

Table 2. PCA based on standardised Predella data (237300 by 29)

	Variance	Cumulative Variance
PC1	76.1%	76.1%
PC2	14.7%	90.8%
PC3	4.8%	95.6%
PC4	1.9%	97.5%

For each principal component an image can be reconstructed, as in the case of each wavelength of the raw data. Clearly, it is easier to visually analyse four or five PC images instead of the whole sequence of 29 original images. In addition, the compression of information makes attractive the application of KDE techniques, which can be applied to the first one, two or three PCs for an informal selection of apparent clusters within the data.

A further advantage of the combination of methods outlined here is in terms of computing time required. The application of clustering methods such as that based upon the Mahalanobis distance using *all* the measured variables would need very extensive computer resources (recall that the data matrix has dimensions 237300 by 29 in the case of Signorelli's Predella). Instead, we have applied the clustering technique of section 1.2 to the first four PC scores. In addition, due to the enormous size of the data set, it is often useful to use a randomly selected subset of pixels in order to further reduce computing time. Moreover, KDEs play a very important role as a guide for Mahalanobis distance clustering. The selection of initial clusters by KDE techniques makes the segmentation of the image more reliable and again reduces the computing time.

Once a clustering has been made, pixels belonging to a particular cluster are "mapped back" by giving them a specified colour in the grey scale image, thus enabling different clusters to be examined in their origi-

nal context. Of course, the number of distinguishable colours limits the number of clusters that it is wise to map back simultaneously.

2.3 Analysis of the test tablet

If this methodology is to have any potential, then it must be expected to perform well on the test tablet, where, visually at least, we have twelve clearly separate areas of the painting.

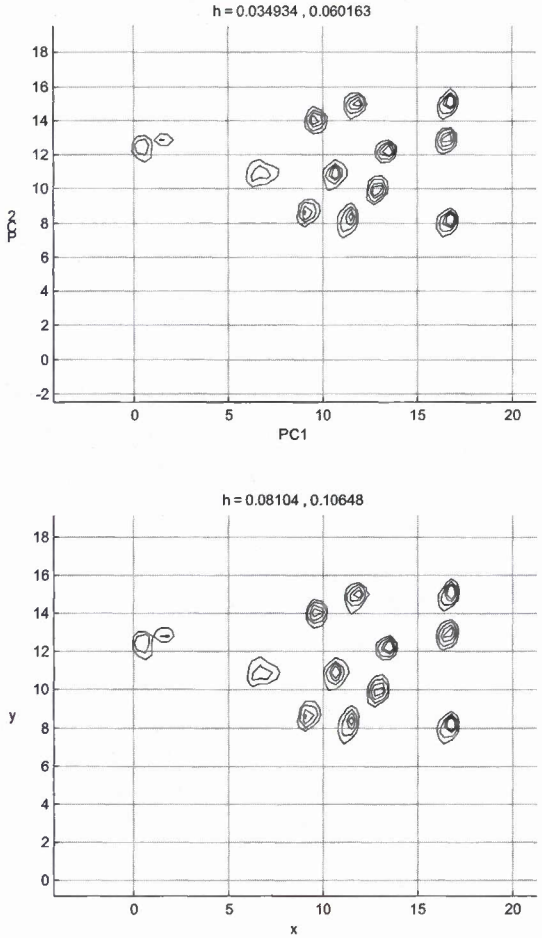

Fig. 4. 25, 50, 75 and 90% contours for KDEs based upon (a) the whole data set (86297 pixels) and (b), a random sample of 10050 pixels for the first two PCs of the test tablet data

For the first two PCs of the test tablet data set, Fig. 4 shows contouring for a KDE based upon (a) the whole data set (86297 pixels) and (b), a random sample of 10050 pixels. There are clearly twelve distinct clusters here.

As expected, the twelve clusters visible correspond to the twelve strips of the original image. (Note that as the clusters for the test tablet are so clear, no Mahalanobis distance clustering was undertaken.) For example, when mapped back to the original image, the cluster in the bottom right of the Fig. 4 corresponds to the fourth strip of the test tablet. The only question raised by this analysis is regarding the presence of the slightly split cluster on the extreme left of Fig. 4. The two parts of this split cluster are the upper and lower parts of the first strip in the test tablet. This is probably due to a very small non-homogeneity of the painted surface of this strip or to a non-homogeneity of lighting. When the first three, rather than two PCs are used, the results (not illustrated here) are just as clear, with twelve distinct contour shells corresponding to the twelve areas of the test tablet.

2.4 Analysis of Signorelli's Predella

Analysis of Signorelli's Predella has resulted in one of the first instances of data obtained using the Image Spectroscopy system developed at IROE.

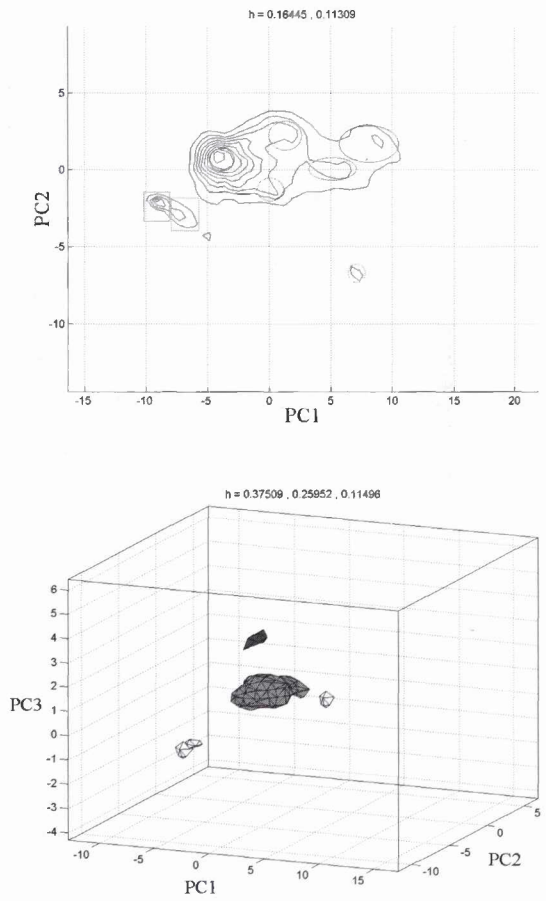

Fig. 5. (a) bivariate contours using (for illustration) varying levels of inclusion, and (b) trivariate 30% contour shells for a KDE based upon a random sample of about 10000 pixels for the first two and three PCs respectively of the Predella data

As a result, this painting was not investigated to clarify particular issues regarding the history of the painting itself, rather to establish the feasibility of the methodology.

Fig. 5 shows examples of bivariate and trivariate contouring based upon the first two and three PCs respectively, using a random sample of about 10000 pixels. The regions of the original painting represented by the indicated clusters of Fig. 5a, b, after "growing" by means of the Mahalanobis distance algorithm, are

shown in Fig. 6a, b respectively. It can be noted that part of the dress of the second person from the left is mapped with the same colour (light blue) as part of the bodies in the middle of the scene. In fact, such pixels belong to a very high-density area in the plot based upon the first two PCs (Fig. 5a). The trivariate 30% contour based upon a KDE of the first three PCs shows how this area in fact consists of two clusters (shown as red and green) which appear well split in Fig. 5b. These two clusters are mapped in Fig. 6b.

3 Summary and conclusions

Recall that clusters can be used for either or both of the following, in the context of the original painting.

- Identification of inhomogeneities of spectral response and therefore of physical-chemical properties of the surface. The non-homogeneities may be due to under-drawings, retouching or restoration. Hidden aspects could be highlighted especially when the data are acquired in the near-infrared region (NIR) since NIR radiation has a relatively high penetration depth (the penetration depth depends, amongst other factors, on the absorbivity of the painted layer, the wavelength etc.).
- Segmentation of the imaged scene into areas of similar spectral behaviour which help in investigating the pigment distribution of paintings and in facilitating successive analysis for the identification of the materials.

We feel that the combination of PCA, KDEs with contouring and a cluster analysis technique such as that based upon Mahalanobis distance is a powerful method that can be used to obtain a segmentation of the image into areas with homogeneous spectral behaviour. The segmentation is very good and clear using the test tablet. Here there are four groups of three strips painted using the same pigment with a different carbon black content (so the strips are *not* completely different). The obtained segmentation for the Signorelli painting is reasonable and confirms the power of the methodology even in a (very) much more complex case.

3.1 Software

The techniques discussed in this paper have been implemented in the MATLAB package by the first named author and are freely available (email: christian.beardah@ntu.ac.uk). The routines include the facility to import and analyse the user's own data. All the illustrations were generated using this software.

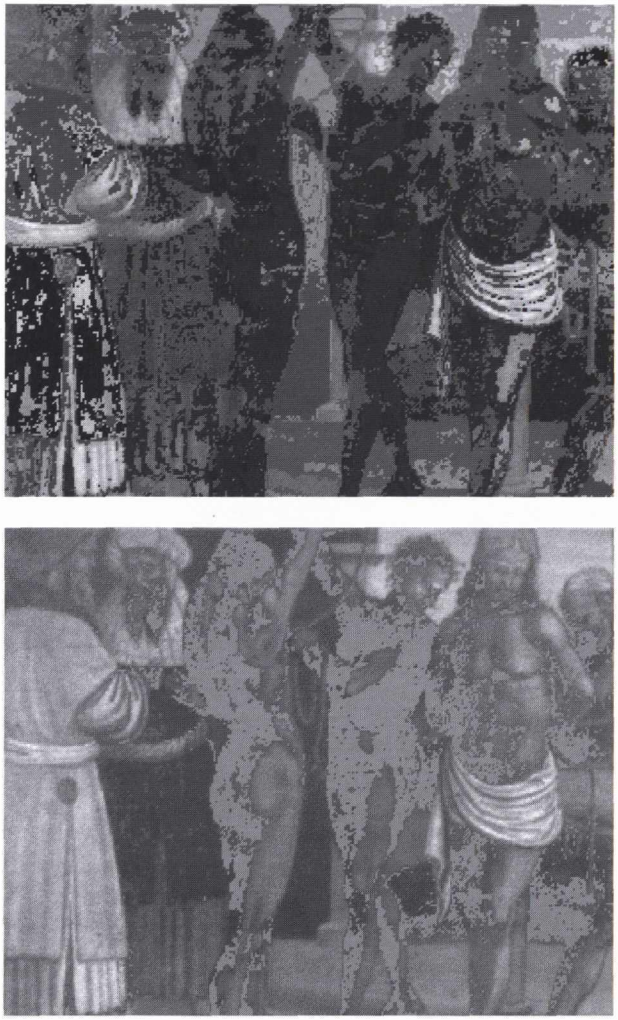

Fig. 6. The regions of the original painting represented by clusters "grown" from initial clusterings shown in Fig. 5a, b respectively (Mahalanobis distance clustering)

References

Baronti, S., Casini, A., Lotti, F. and Porcinai, S., 1998. Multispectral imaging system for the mapping of pigments in works of art by use of principal-component analysis, *Applied Optics* 37 (8), 1299-1309.

Baxter, M.J., 1994. *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh University Press.

Baxter, M.J., Beardah, C.C. and Wright, R.V.S., 1997. Some Archaeological Applications of Kernel Density Estimates, *Journal of Archaeological Science* 24, 347-354.

Beardah, C.C., 1998. Uses of Multivariate Kernel Density Estimates in Archaeology. In: Dingwall, L., Exon, S., Gaffney, V., Laflin, S. and van Leusen, M. (eds), *Computer Applications and Quantitative Methods in Archaeology 1997* (British Archaeological Reports, International Series S750).

Beardah, C.C. and Baxter, M.J., 1996. The Archaeological use of Kernel Density Estimates, *Internet Archaeology* 1. (http://intarch.ac.uk/journal/issue1/beardah_index.html)

Beardah, C.C. and Baxter, M.J., 1999. Three-dimensional data display using Kernel Density Estimates. In: Barcelo, J.A., Briz, I. and Vila, A. (eds), *Computer Applications and Quantitative Methods in Archaeology 1998* (British Archaeological Reports, International Series 757), 163-9.

Beier, T. and Mommsen, H., 1994. Modified Mahalanobis filters for grouping pottery by chemical composition, *Archaeometry* 36, 287-306.

Binford, L.R., 1978. Dimensional analysis of behaviour and site structure: learning from an Eskimo hunting stand, *American Antiquity* 34, 330-361.

Blankholm, H.P., 1991. *Intrasite Spatial Analysis in Theory and Practice*. Aarhus: Aarhus University Press.

Bowman, A. and Foster, P., 1993. Density Based Exploration of Bivariate Data, *Statistics and Computing* 3, 171-7.

Silverman, B., 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Wand, M.P. and Jones, M.C., 1995. *Kernel Smoothing*. Chapman and Hall, London.