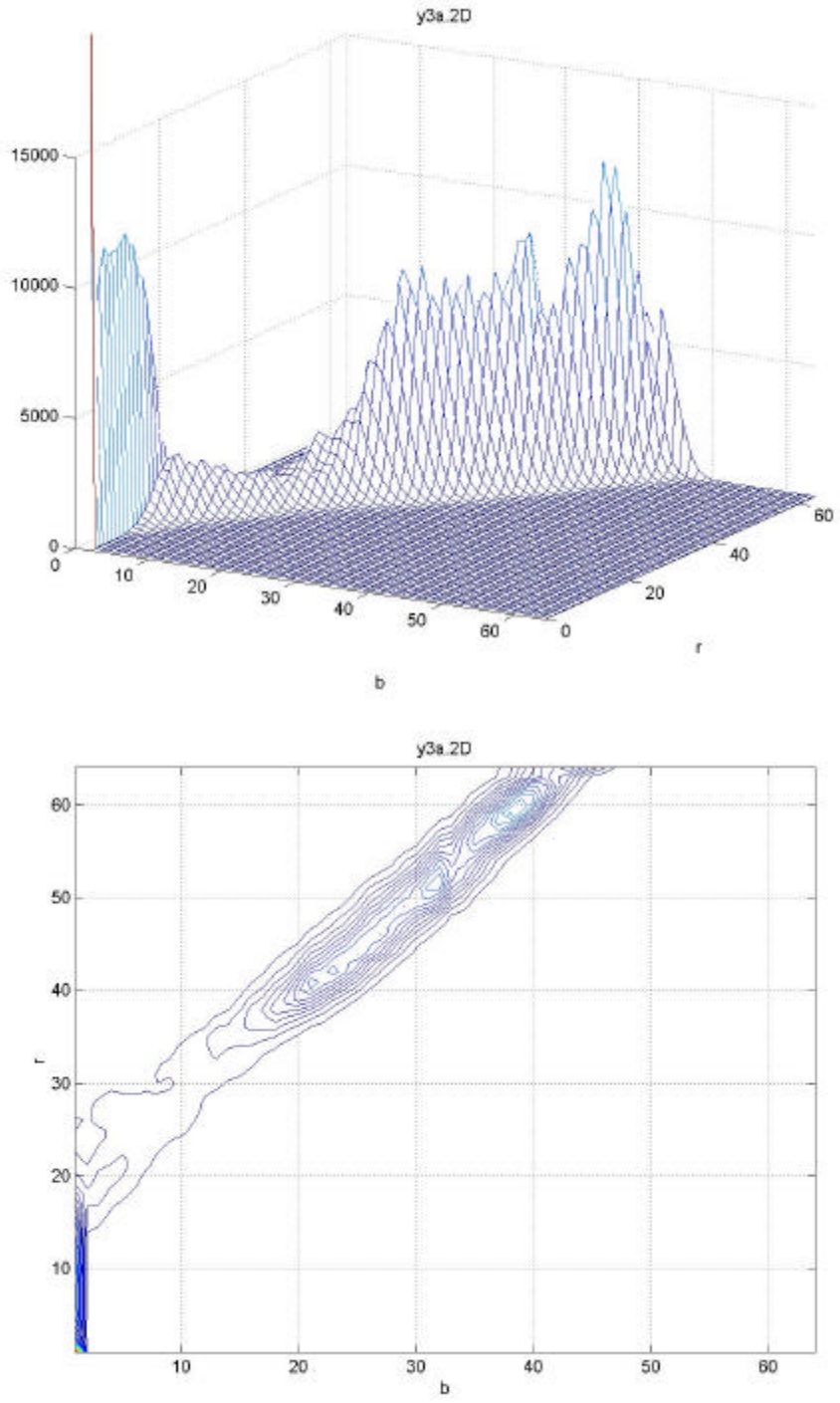


Figure 4.13: (a) Two dimensional r-b histogram of the yellow poplar board y3a.dat of Figure 4.4(a), (b) Contour plot of the r-b histogram

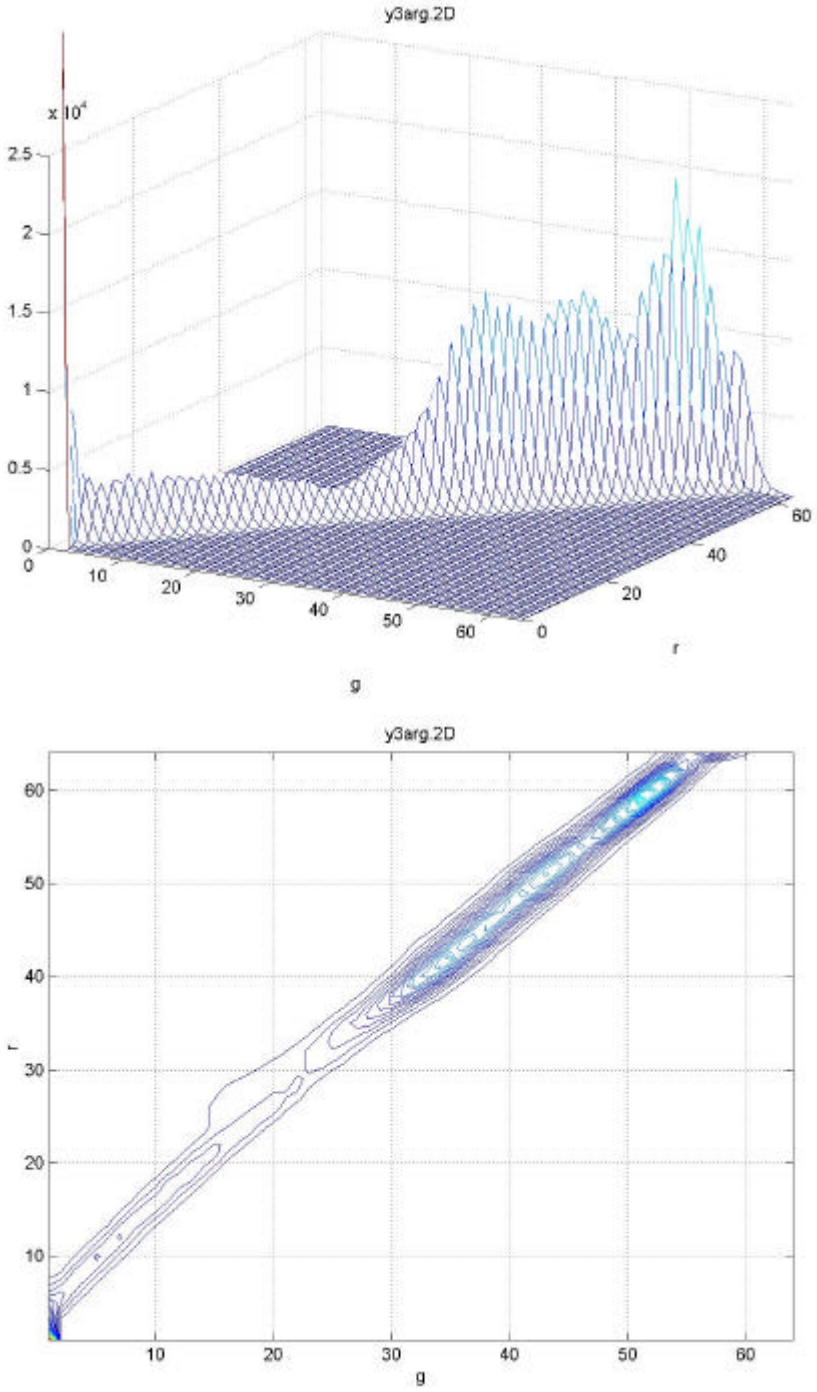


Figure 4.13: (c) Two dimensional r-g histogram of the yellow poplar board y3a.dat of Figure 4.4(a), (d) Contour plot of the r-g histogram

4.6 Filtering the Histogram

The goal of filtering is to remove or reduce the extent of small false peaks that occur in the histogram, while, at the same time maintaining the original shape of the histogram. Doing so should help reduce the number of false clusters that are created, and consequently should reduce the number of iterations needed to obtain a good segmentation.

The filtering process has to be done in 3-dimensions since the multispectral clustering algorithm described in this chapter operates on 3-dimensional histograms. The 2-dimensional histograms were used only to graphically depict the histograms. Spatial filtering or convolution in 1 and 2 dimensions are very common. In 1-dimension the filter is a window or a linear array which is convolved with the data. In 2-dimensions, the filter is represented as a 2-dimensional spatial mask which is convolved with the two dimensional data (Section 2.8). The same concept can be extended to filtering in higher dimensions. A 3-dimensional filter is represented as a cube which is convolved with 3-dimensional data [GON92].

For example, if F represents a three dimensional filter of size 3x3x3, and H(a, b, c) is any point in the 3 dimensional histogram, the value of the filtered histogram H_f at {a, b, c) is computed as

$$H_f(a,b,c) = \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{2} F(i,j,k)\,xH(a-1+i, b-1+j, c-1+k)$$

Note that the histogram is padded with zeros around all the boundaries.

Averaging and Gaussian filters described in Section 2.8, of sizes ranging from 3x3x3 to 11x11x11 were applied on the histograms of the boards (Section 3.1.1). A Gaussian filter is generated by a Gaussian distribution with a mean at the center of the filter and variance defined by the covariance matrix. The extent of smoothing due to the filter is decided by the covariance matrix. For a filter of a given size, the extent of smoothing can be increased by increasing the values of the covariance matrix. Values for the filter were generated for each element of the filter, at the center of each element of the filter. In other words the Gaussian distribution was

evaluated at the center of each filter element, using the mean and the covariance matrix as described above. The filter coefficients are then normalized to 1.

Median filtering was also applied in an attempt to remove the stray peaks in the histograms. Further, median filtering followed by Gaussian smoothing was also tried. The technique of scale space filtering, which is an iterative form of filtering, is discussed in the next section. This technique was examined since it is supposed to give superior results compared to simple filtering. All these filters, however had some common drawbacks, and had similar effects on the histograms. The summary of using all these filters and their effect on the histograms will be discussed in Section 4.8.

4.7 Scale Space Filtering Approach to Clustering

Scale space filtering technique uses repeated filtering operations to smooth out the irregularities in the signal as described in Section 2.8. The concept of scale-space filtering [WIL90] is to repeatedly filter the signal to smooth out spurious peaks, while using the original signal to obtain the accurate location of the peaks. Smoothing the signal helps eliminate noise and spurious peaks. However, with repeated filtering, the location of the peaks are shifted. So, once the peaks are identified in the filtered signal, the original signal has to be used to determine the exact location of the peak.

A variation of this basic procedure is developed in order to simplify the filtering operation. Instead of using a spatial mask to filter the histograms, the filtered histogram is obtained as a sum of normal distributions located at each cell. This procedure reduces the number of computations required to obtain the filtered histogram. The filtered histogram *H_new*, for a two dimensional case, is obtained at every cell (*k*,*l*), *k* = 0, 1, …63; *l* = 0,1, … 63; as,

$$H\_new(k,l) = \frac{\sum_{i=0}^{i=n}\sum_{j=0}^{j=m} N(i,j,s)H(k,l)}{\sum_{k=0}^{k=63}\sum_{l=0}^{l=63} H(k,l)}$$

where,

$N(i,j,\sigma)$ = normal distribution with mean = (i,j), and the standard deviation is given by σ

The mean of a normal distribution is defined in Section 2.8.2.

H = original histogram

σ = scale of the filtering

The filtered histogram is evaluated as a weighted sum of normal distributions centered at each point in the histogram. The peaks are then identified in *H_new*.

The function *P* is evaluated to obtain the final classification [STP85].

$$p(x_i|C_k) = \frac{1}{2}(1 + \langle \nabla \hat{p}(x_i) \rangle \cdot \langle d_{i,k} \rangle) e^{-|d_{i,k}|^2/2(s)^2}$$

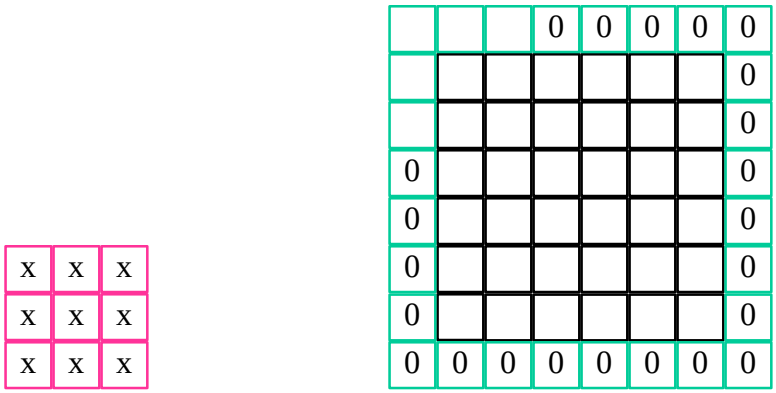
where $\nabla \hat{p}(x_i)$ is the gradient operation on $\nabla \hat{p}\, s$ at $d_{i,k} = (\mu_k - x_i)$ and $\langle a \rangle \cdot \langle b \rangle$ represents the inner product between two normalized vectors.

Here the number of clusters is equal to the number of peaks in the filtered histogram. This method has some of the drawbacks of filtering. The peaks corresponding to the defect areas are weak and difficult to locate. Further, the classification produces linear boundaries do not correspond to features in wood, since the natural clusters formed in wood are rarely linear.

4.8 Drawbacks of Filtering

While filtering reduces irregularities in the histogram, it also destroys data. A large value of σ can smooth out most of the irregularities, but also could destroy a true peak. This is a serious problem. Since the defect regions are very small, they are almost completely eliminated by any filtering operation. In a 1 dimensional signal, filtering causes loss of data at the edges which is different from the smoothing effect on the entire signal. If a filter of length M is used on data of length L, the length of the filtered signal is (L - M/2). This is so, because the signal is assumed to be padded with zeros at both the edges.

In a two dimensional signal, data is lost at the four edges. This is demonstrated in Figures 4.12. Figure 4.12b shows a 2 dimensional signal padded with zeros around the edges. The filtering is performed using a 3x3 mask, shown on the top, left corner of the 2 dimensional signal in Figure 4.12b. The shaded region around the periphery represent the cells which are affected by the zero padding around the edges.



3x3 Filter

2D Signal padded with zeros being convolved

Figure 4.12(a): 3x3 filter (b): Demonstration of 2D filtering

In a 3 dimensional signal, data is lost over the entire surface of the signal, unlike in a 2-dimensional signal where data is lost along the boundaries, and a 1-dimensional signal where data is lost only at the edges. Since the histograms of wood are quite sparse, data tends to be destroyed easily when smoothing filters are applied to it. Thus small regions of the histogram is can be completely annihilated by the filtering process especially if the filter is large.

The histograms have very deep valleys around the peaks due to the presence of noise. A very high order filter is necessary to smooth out the deep valleys in the histogram, but using a higher order filter also destroys the signal in regions where the clusters are not very pronounced as described above. Using a large filter will significantly reduce the false peaks in the histograms. However, it also destroys data making the filtered data quite useless. A smaller filter will retain the data in the original histogram but does not minimize the false peaks.

4.9 Evaluation of the Algorithm

In the multispectral clustering algorithm the choice of the initial threshold T_i plays an important role in obtaining a good segmentation. Using the average of the histogram as T_i on board images, generally resulted in only one cluster being formed even though L defects were present. In the event that more than one cluster was found using this value for T_i, the resulting clusters did not correspond to a true segmentation of the defects. Since the color histogram occupies a very small volume in the full 3-dimensional color space, the average histogram element value generally comes out to be between 1 - 10. Increasing the value of T_i still led to a lot of false clusters being formed. This phenomena has been explained based on the observed characteristics of these histograms.

Generally dark colored features are relatively easy to separate using this method. A good segmentation is possible even after one iteration only if these features are targeted as defects. There are two reasons for this. First, the dark regions are physically well separated from the main clear wood cluster in 3-dimensional color space. Second, the shape of the histogram is such that it is possible to set the T_i threshold value so that the dark regions fall in one cluster and the rest of the histogram falls in another cluster.

As the threshold T_i is increased, more clusters are formed, but many false clusters are also formed. The segmentation algorithm can be used if only the dark colored features need to be located. Interestingly, such defects can typically be found using simple techniques in 1-dimensional black/white imaging. The multispectral clustering technique is unable to locate, in a reasonable number of iterations, even upon using the best set of parameters, blue stain on oak boards, and the heartwood and sapwood regions in yellow poplar boards. In pine boards, the grain pattern is often classified as a defect.

The algorithm assumes that concentrations or clusters in the data vectors always form distinct peaks in the histogram data [GOL78]. This is one of the major drawbacks of the algorithm. The algorithm fails when if no distinct peaks and valleys exist. Unfortunately, distinct cluster peaks do not occur in most board images, even ones that contain numerous

defects. There is often more than one peak corresponding to a single region on the board. For instance the clearwood region may have one main peak in the histogram with many smaller peaks surrounding it. On the other hand, sometimes distinct peaks may not exist for two different regions on the board. For example, in yellow poplar board, the heartwood region can merge into the sapwood region without producing a distinct peak. Thus the existence of peaks do not always indicate a separate cluster.

Another important drawback of this technique, a drawback that renders it very hard to implement in an automatic fashion, is that the thresholds in the first and the subsequent iterations are heavily dependent on the nature of the board surface. For instance, if a board has a large area of blue stain, a high value of T_i is needed to obtain good segmentation. On the other hand, if the blue stain content is relatively low, a high value for T_i results in the stain being merged with another cluster, usually the clearwood cluster. Thus the threshold works only if it is set at the optimum value and this value is dependent on the very features we are trying to locate using the algorithm. The above holds for other features as well. Hence in order to get a good segmentation, a different value for T_i must be used on every board.

Selection of a good value for T_i is the key to reducing the number of iterations, and the importance of this selection can never be overstressed. The parameters (thresholds and the selected clusters) have to be set so that the algorithm converges quickly in order for the algorithm to be practically useful. Further, if it is not possible to select a good threshold in the first iteration, several iterations are required before a reasonable clustering can be obtained.

Broadly it can be concluded that this method works well on color histograms only if the defect and clearwood areas are well separated in 3-dimensional space and the noise levels are within reasonable limits. For images of wood this means only the dark colored defects can be identified easily.

4.10 Conclusions

The techniques used in the two clustering approaches described in this chapter form the framework for most conventional clustering algorithms. The assumption that peaks in a data set correspond to cluster centers is one of the basic assumptions made in most of the conventional clustering techniques. The inherent nature of the wood histograms along with noise, violates this assumption.

Traditionally filtering is almost always used to reduce the level of noise in a signal. Filtering has been extensively studied in 1 and 2 dimensions, but not in higher dimensions. The histograms of wood images occupy a very small volume in the 3 dimensional space, and this makes them very susceptible to the destruction of data. While a very high order filter is necessary to smooth out the deep valleys in the histogram, it also destroys the signal in regions where the defect clusters are not very pronounced.

The studies and experiments described in this chapter conclusively prove that the conventional approaches to clustering are not capable of locating all the features (defects) on the wood surface in an efficient manner. All this indicates the need for an entirely new approach to segmentation. The algorithm should either be robust in the presence of the noise, or the filtering process should be such that the signal is not destroyed where the signal is weak. This led to the development of a new model based approach, which will be described in the next chapter.