

Chapter 4. Multispectral Clustering Technique

The multispectral clustering technique is based on the algorithm proposed by Goldberg and Shlien was described in Section 2.7.1 [GOL78]. Like most clustering algorithms, it is based on the assumption that peaks in the data set correspond to cluster centers. In the original form of the algorithm, clusters are interactively split and merged in order to reduce the p value (Section 2.7.1). A study in the performance of conventional clustering algorithms on histograms (of wood images) with the aim of segmenting the images has never been done before. Hence, the first goal was to examine the usefulness of this algorithm in segmenting and locating the desired wood features.

This algorithm also has a few problems in its original form. The method is interactive. The user has to visually examine the results of the classification and select the clusters that need to be split or merged. While this gives it flexibility which is good, the segmentation algorithm cannot be interactive in its final form, critically limiting its use in real time defect detection systems. The second goal was thus to modify this algorithm such that it can run without user intervention.

The original algorithm is extremely computationally complex considering all the mathematical manipulations that have to be performed on a 262,144 element histogram. Further, the process is iterative which adds to the complexity. Therefore, the third goal was to make the iterations converge fast.

To achieve all the three goals, the effect of the algorithm on images of wood had to be extensively studied, along with the color characteristics of wood. The histograms were thresholded to obtain clusters as described in the original algorithm. Once the initial clusters are obtained, the results were visually examined to see the effect of splitting and merging the clusters. This effect was tested so as to give an insight to the performance of the algorithm on the histograms obtained from images of wooden boards.

To simplify the problem, 4-connectivity was used instead of 8 connectivity. The concept of 4-connectivity is explained in Figure 4.1. Another reason is that 4 - connectivity should give a more detailed break up of the clusters and help understand better the effect of the algorithm on the histograms. The algorithm was applied to three different species of wood described in Section 3.5. The exact procedure which was followed is given in Section 4.1.

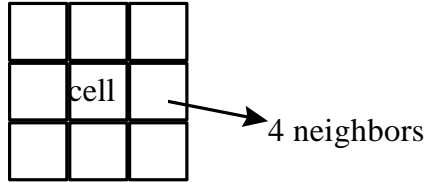


Figure 4.1: Demonstration of 4-connectivity in 2 dimensions

4.1 Modified Algorithm

Let $\hat{H}(i, j, k)$ be the 3 dimensional histogram generated from the r, g, b color image (Section 2.7.1). A threshold T_i is first selected but is not necessarily the mean of the histogram as described in Section 2.7.1. Initial testing found setting T_i equal to the mean gave very poor results. Various experiments were performed to study the effect of varying T_i . These experiments will be described in the Section 4.2.

1. Threshold the histogram \hat{H} , and divide the elements of the histogram into two sets, $S1$ and $S2$ such that the elements in $S1$ all have values greater than or equal to T_i . The elements in $S2$ have values less than T_i .
2. Find the connected components contained in the set $S1$, using 4 connectivity. Assume this gives n clusters.
3. Assign each element of $S2$ to a cluster among the n clusters, whose mean is closest to the element in $S2$. Now the histogram is divided into n clusters. This gives the required map described in Section 3.3. This step represents the completion of one iteration of the algorithm. Steps 2 – 4 are illustrated for a 1 dimensional data set in Figures 2.3, in Chapter 2.
4. The clusters were interactively split or merged after looking at the results of the clustering process. The criteria used to split or merge the cluster will be described in Section 4.5

4.2 Effect of varying the Initial Threshold T_i

Studying the results of segmentation after one iteration forms a basis for understanding the effect of the algorithm on the histogram. The result of the first iteration is most important since it will impact all the subsequent iterations. If the clustering is done right in the first iteration, it will help the algorithm converge very fast. On the other hand, if there is a large amount of false clustering (defect regions in clearwood clusters and vice versa) after the first iteration, it will take many more iterations to rectify clusters by splitting and merging them. As mentioned earlier in real time applications that are critical in time, having a large number of iterations is highly undesirable, especially when the algorithm is as complex as the multispectral clustering algorithm. The first goal is thus to optimize the first iteration of the algorithm.

The result of segmentation after the first iteration is solely dependent on the value of the initial threshold T_i . Good clustering can be obtained after the first iteration if T_i is selected near the optimum value. Also, an optimum T_i will make the algorithm converge faster. It is thus very important to be able to identify the optimum value of T_i for any board.

To form a good basis for selecting T_i , its effect had to be studied in detail. Further, since there are no documented results of how the clustering algorithm will perform on the histograms of wood, a study of the effect of T_i forms the first step in developing a better understanding of the algorithm. The initial threshold T_i was varied over a wide range of values, and the results were studied on boards of different species. Once the results of the first iteration are obtained, they can be used to set the thresholds for subsequent iterations.

The effect of T_i was studied on a number of boards (around 10 boards from each of the 3 species mentioned in Section 3.5). The details of the results for two such boards each belonging to a different species of wood, pine, and yellow poplar, are tabulated in Tables 4.1 and 4.2. The results of these two boards are representative of the general results obtained from these species.

These tables summarize the effect T_i has on the algorithm when T_i is varied from the mean value to a much larger value.

In order to perform an accurate analysis, T_i was set initially equal to the mean, and varied in steps of 20. The final segmentation remained the same over a range of values and changed noticeably only at certain T_i values. The values of T_i for which there was a noticeable change from the previous segmentation are recorded in the tables. Such experiments were carried out on several boards and the results obtained were consistent. It will thus suffice to describe in detail one board specimen from a species.

Figure 4.2(a) shows the image of a pine board, that has features like, distinct grain patterns, knots, blue stain, and a small region of pith. The pith region is located in the center of the board. The blue stains are scattered along the left edges and a few areas in the center. These are the regions which need to be identified by the segmentation process. Figures 4.2(b - e) show the effect of varying the initial threshold T_i on the final classification.

Table 4.1 Effect of varying the threshold for p8.dat (Southern Pine)

Cluster	Centers	1 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	2 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	3 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	4 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	5 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	6 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	7 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	8 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	9 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	10 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$
Threshold	Mean = 6	37.7 28.12 14.52	0 0 3	13 11 11	13 12 6	43 22 6	43 35 3	60 49 26	62 51 31	62 53 36	63 52 32
	100	37.84 28.47 14.42	59 49 31								
	300	38.745 28.3809 15.0532	2.5 0 0	25.875 13.625 0							
	400	38.75 29.39 15.027	2 0 0	25 13 0	27 15 0						
	420	38.7892 29.4193 15.0597	2 0 0	22.5 14 2	25 13 0						
	435	38.818 29.443 15.056	2 0 0	22.5 14 2							
	450	38.8552 29.451 15.058	2 0 0	22.5 14 2	53 44 28						
	500	38.93 29.485 15.04	2 0 0	23 14 2	53 44 28						
	600	39.07 29.63 15.18	1 0 0								
	800	39.25 29.79 15.30	0.5 0 0								

Table 4.2: Effect of varying the threshold for y3a.dat (Yellow Poplar)

Cluster	Centers	1 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	2 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	3 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	4 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	5 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	6 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	7 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$	8 $\begin{pmatrix} r \\ g \\ b \end{pmatrix}$
		→							
Threshold	Mean 6.87	40.65 32.797 20.539	6 3 2	16 13 8	35.5 18 4	47 31 17	53 43 22	56 43 22	63 47 33
	50	40.5221 32.8965 20.852	42 29.5 12						
	100	40.6118 33.0286 21.0263							
	200	41.2289 33.6949 21.7428							
	500	48.096 40.4094 27.8101	14.3387 8.83871 1.62903	25.5 19.25 8.75	27 21 10	30.333 23 11.33			
	800	49.395 41.7253 29.061	9.762 4.89 0.017	17 11 1.5	18.33 12 2.33	19 13 3.5	20 14 4	33.75 25.5 12.25	
	1000	49.4866 41.8705 29.2612	9.264 4.5 0	17 11 2	18 12 2				

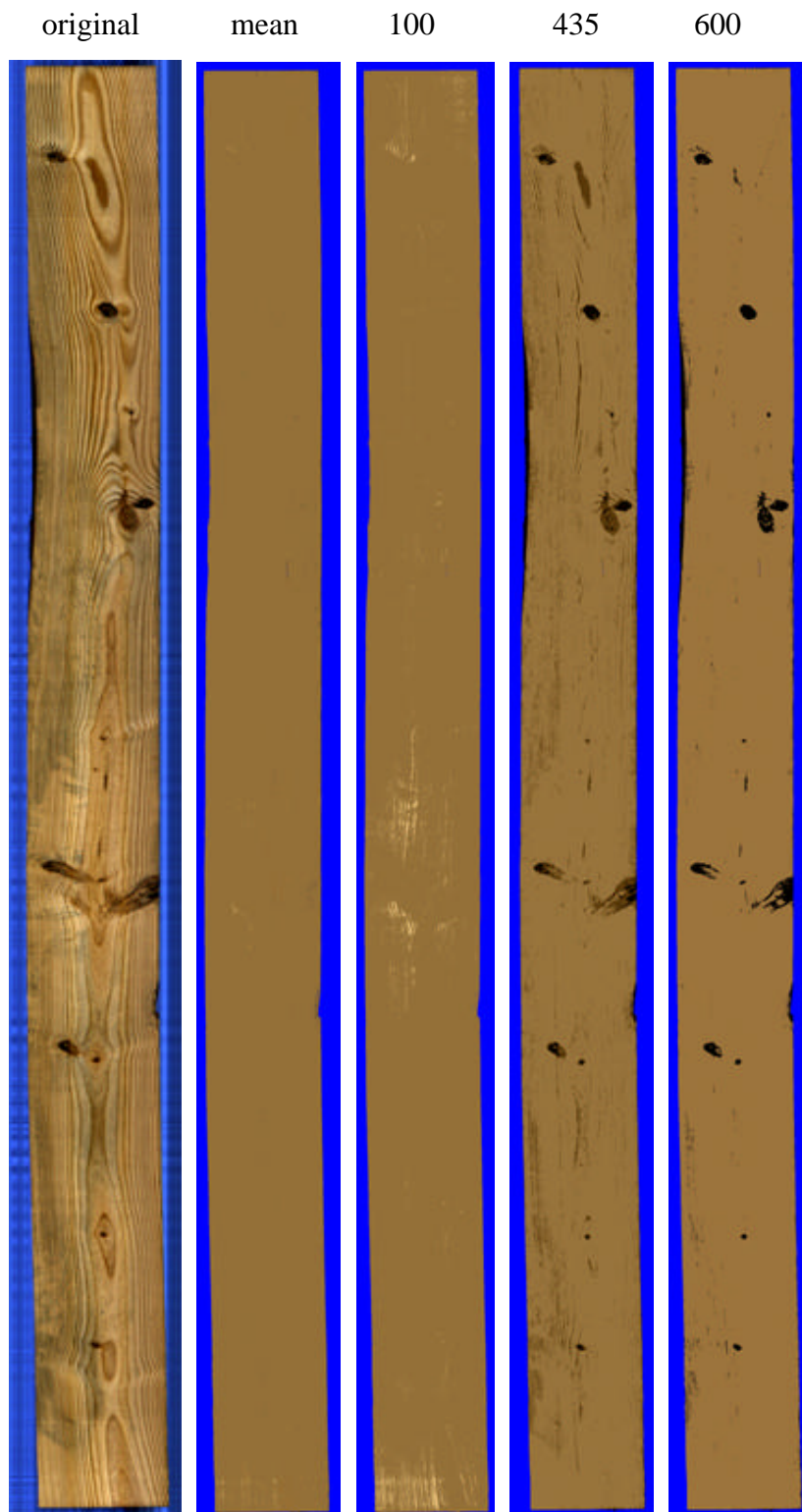


Figure 4.2: (a) Image of a Pine board p8.dat, (b - e) Effect of setting $T_i = \text{mean}, 100, 435,$
and 600

The mean of the histogram for the board image of Figure 4(a) is found to be 6.87 using the procedure described in Section 2.7.1. If T_i is set to this value, ten clusters are formed. The (r, g, b) values of the cluster centers of each of the ten clusters are given on the first row of Table 4.1. Unfortunately, Figure 4.2(b) shows that none of these clusters represent a true classification of the different features on the board. In other words, none of the features on the board like the knots, pith etc. are identified. When the threshold is set this low, the algorithm actually picks up only noise. This is demonstrated in Figure 4.3. It was found that the total number of non zero elements in the histogram were 11798, which is only 4.5% of the total number of elements in the histogram. Further, as is expected 6208 (2.37% of the total number of elements in the histogram) elements cells have values above the mean value of the histogram. Since the histogram is quite sparse, it is very common to have *holes* and sudden dips or spikes in the histogram that represents manifestations of the noise in the image. When the threshold is set near the mean, these regions are separated into different clusters as shown in Figure 4.3.

As T_i is increased, the clustering of stray noise is reduced and at $T_i = 100$, only two clusters are formed. The cluster centers are given on the second row of Table 4.1. While the first cluster includes most of the board, the second cluster does not segment the image into regions that correspond to any feature on the board which needs to be segmented. At $T_i = 300$, the dark regions start to form separate clusters. The really dark knots fall into cluster 2, (Table 4.1), while the reddish knots fall in a cluster 3. Cluster 1 can be considered as the clearwood cluster, and clusters 2 and 3 as defect clusters. The reddish grain pattern on the top of the board is also in cluster 3. This is very undesirable. While the knots are generally considered as defects, grain patterns are not. They should ideally fall into the same cluster as clearwood. The blue stain is also part of cluster 3. Actually only the combination of blue stain and grain is detected. The reason for this is, the blue stain on the earlywood regions are very difficult to distinguish from some of the other parts of the board. There are also some clearwood pixels that are falsely classified into a defect cluster.

An examination of Figure 4.2 (d) suggests that $T_i = 435$ is a good threshold values since it produces the best segmentation for the first iteration. All the major features of the board are

being detected by the algorithm. However, there is one major drawback. It is not possible to separate the reddish grain pattern from the knots where there is a very subtle change in color, and also, some of the blue stain is not being identified. From Table 4.1, cluster 1 (38.818, 29.443, 15.056) is closest to the clearwood regions of the image. As T_i is increased beyond 435, the clear wood cluster is broken into two clusters (Table 4.1), with the lighter regions of the board forming a separate cluster. Examination of the image (image not shown) showed that some of the reddish grain pattern was still put in the defect cluster. Finally, some of the prominent blue stains were put into one of the clearwood clusters.

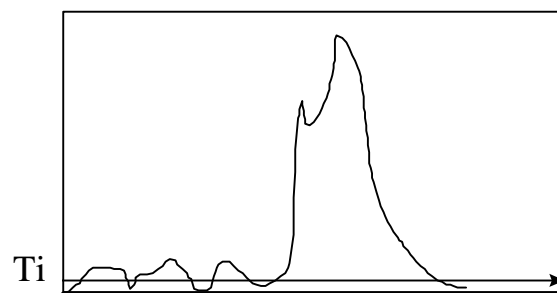


Figure 4.3: Separation of Clusters

When $T_i = 600$, only the knots and the pith region comprise the defect cluster. The grain patterns are no longer a part of the defect cluster which is desirable, but the blue stain is completely missed. Also, parts of the knots that are lighter in color are now classified as clearwood.

The experiments that were performed on the pine boards to determine the effects of T_i were repeated on yellow poplar boards. The results of one yellow poplar board that is representative of the species is described. The Yellow Poplar board shown in Figure 4.4(a) has some mineral streak which appears as a dark colored defect, worm holes, heartwood and sapwood regions and some saw marks on the top of the board which is actually a planing defect. The final segmentation needs to identify these regions. The results of using different values of T_i are tabulated in Table 4.2.

As in the previous case, initially T_i is set equal to the mean of the histogram of the image. Many clusters are formed none of which represent a true segmentation as can be seen in Figure 4.4(b). As T_i is increased above this average value of stray noise, there is only one cluster being formed in the thresholded histogram. Only one cluster is formed if all the elements of the histogram that are above T_i form a single connected component. The (r, g, b) values of the center of this connected component is given in Table 4.2, in the rows corresponding to $T_i = 100$ and 200.

The dark mineral streak areas and the planing defect begin to form a separate cluster when the threshold is increased to 500 as seen in Figure 4.4(c). As T_i is increased to 800, there is no significant change in the segmentation (Figure 4.4 (d)), except traces of the heartwood region in the center become visible. When T_i is set to 1000, again only the dark mineral streak areas and the planing defect are detected as defects as shown in Figure 4.4(e). The algorithm picks up only the dark colored defects and the planing defects though four clusters are formed. It would have been very useful if the defects were in different clusters based on the nature of the defects. It is interesting to note that the planing defect is contained in the same cluster as the areas of mineral streak.

It was shown that it was not possible to obtain heartwood - sapwood separation for any value of T_i . Thus, in order to obtain the heartwood - sapwood separation it is evident that more than one iteration is necessary. Further, a careful examination of the images in Figures 4.4 (c), (d) and (e) showed that parts of the region with dark colored mineral streak and the planing defect fell into the same cluster. Thus the results of the clustering after one iteration are not useful and more iterations are needed to obtain a useful segmentation of the image.

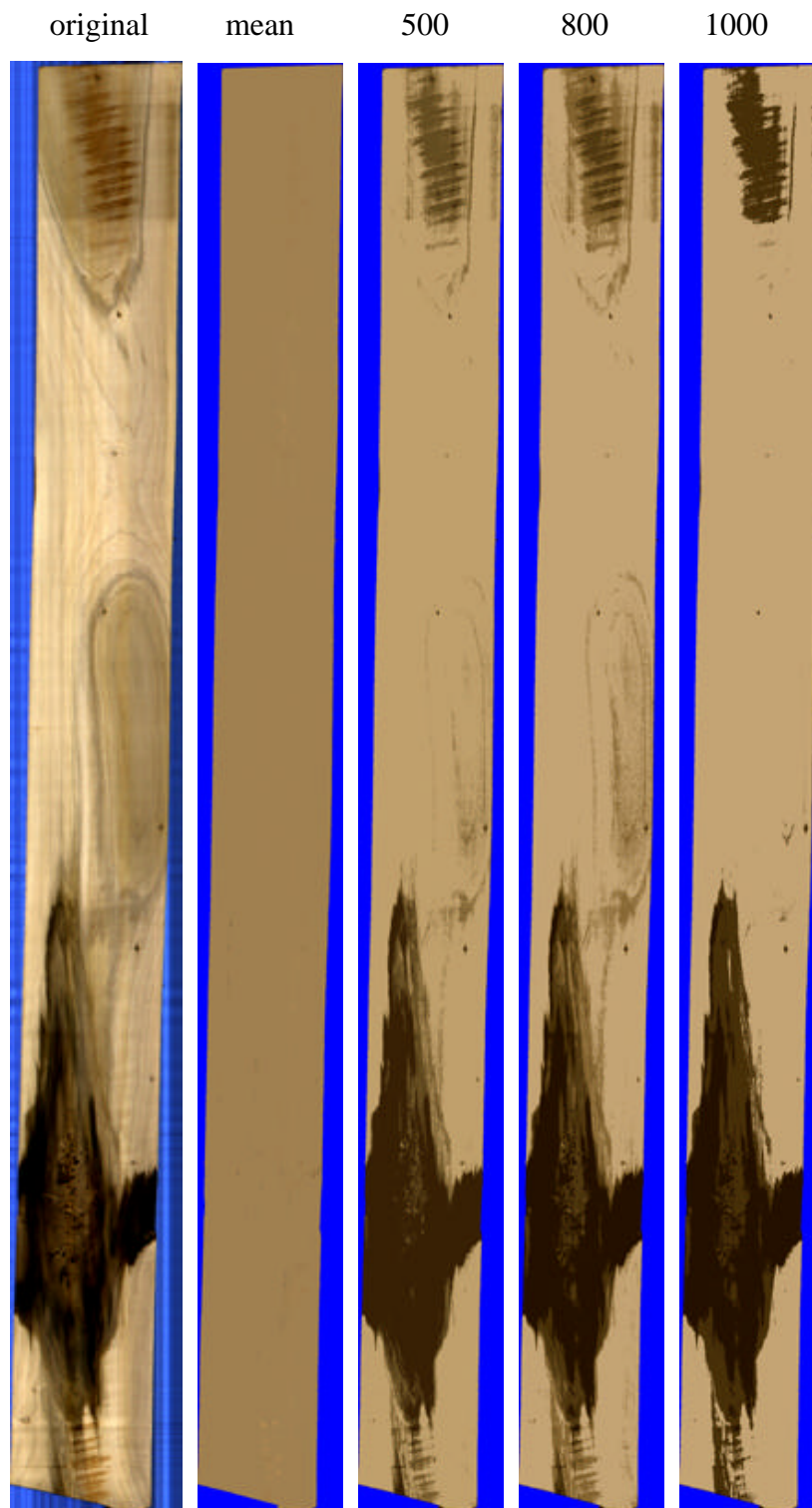


Figure 4.4: (a) Image of Yellow Poplar board y3a.dat, (b - e) Effect of setting T_i at mean, 500, 800, 1000

4.3 Analysis of the Effect of the Initial Threshold T_i

A few conclusions can be drawn about the effect of T_i on the clustering process based on the results of the above study. The best segmentation that can be obtained with the optimum value of T_i can generally detect only the dark colored defects. This result was found to be true for all the three species of wood that were considered, including the oak boards. It was thus possible to obtain good results with oak boards that had predominantly dark colored defects. One such board is shown in Figure 4.5(a). Figure 4.5(b) show the typical best case results obtained with such boards. The board shown in Figure 4.5a is an oak board with a uniform color. The dark regions on the board are the areas that have to be identified by the segmentation algorithm. Figure 4.5b shows the result of using the initial threshold T_i of 500.

The experiments described in the previous section show that if a board has defects that are not dark in color, it is not possible to identify all the defects in one iteration. Suppose for some threshold T_i , a total of n clusters are formed. It would be useful if it were possible to divide these n clusters into n_1 clearwood clusters and n_2 defect clusters. It would then be possible to merge the desired clearwood and defect clusters in the next iteration to obtain a satisfactory segmentation. Thus the segmentation can be completed in two iterations. This however cannot be done for the pine or the yellow poplar boards that were considered in the study. In the pine boards, it was difficult to exclude areas of grain pattern from the defect clusters, while at the same time identifying all of the defect regions. In the yellow poplar boards, it was difficult to obtain good heartwood-sapwood separation. All this leads to the conclusion that more than one iteration is required to identify all the important features on the boards. The use of multiple iterations will now be examined.

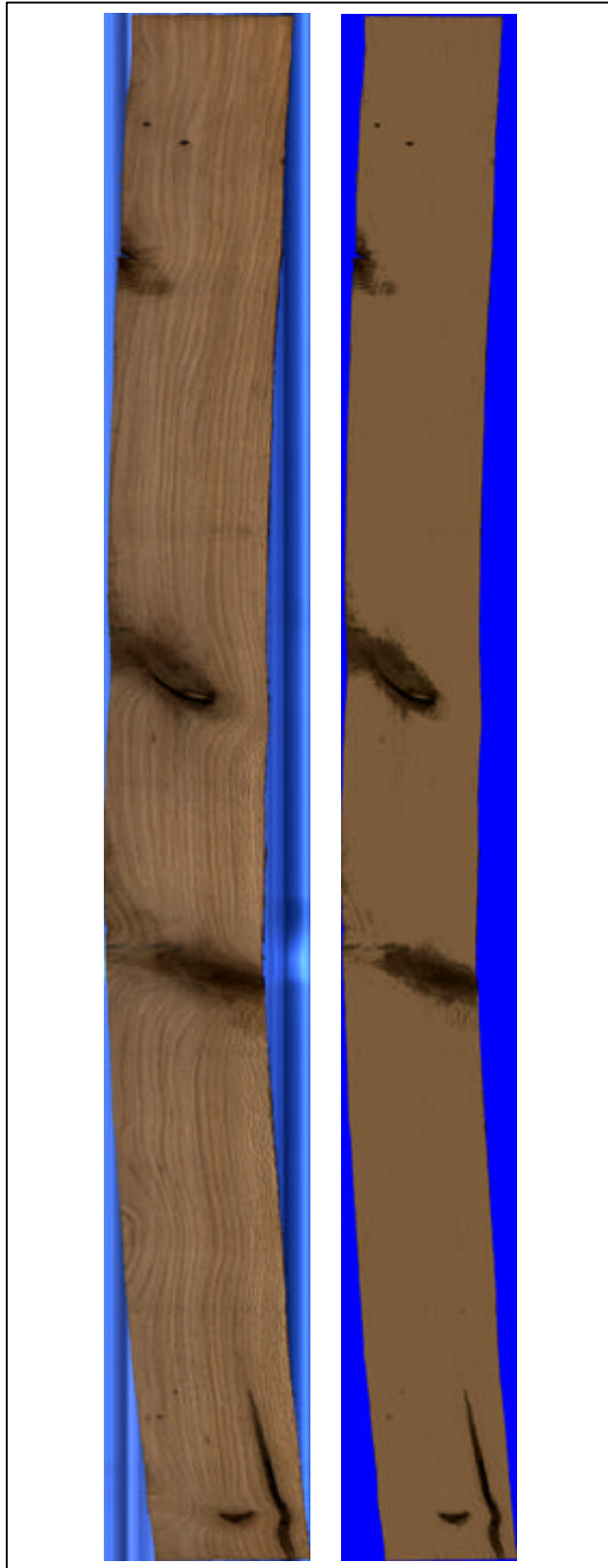


Figure 4.5: (a) Image of Oak board c104a.dat, (b) Segmentation result with $T_i = 500$

4.4 Multiple Iterations

When a single iteration is not sufficient to find all the features on a board, some of the identified clusters have to be split or merged to obtain a better segmentation (Section 4.1). The goal here was to test the feasibility of obtaining a good segmentation using multiple iterations of this algorithm. If good segmentation can be achieved, then the next step is to determine whether the number of iterations required to generate this segmentation is acceptable. Note that these first two studies are based on manually examining the data to determine if an identified cluster should be split or merged. If both findings are positive the final step would be to create algorithms for automatically making this split/merge decision.

To proceed in examining the utility of multiple iterations, the best results of the first iterations were used. After examining the clusters visually, the cluster having the maximum area of erroneous classification was selected to be split in the next iteration. Similarly, all the defect clusters were merged together when they seemed to belong to the same category. Figure 4.6 shows the image of a yellow poplar board. The results of segmentation after three iterations, are shown in Figure 4.7. In the second iteration, the main cluster which had both heartwood and sapwood was split, and in the third iteration, all the defect clusters were merged together. However, the heartwood region is still not clearly segmented out.

The other boards used in this study gave similar results. The initial threshold T_i was set to the optimum value, and the right clusters were manually split/merged. Even after using the best parameters, it was not possible to get a reasonable segmentation within 3-4 iterations. Every additional iteration increases the complexity of the algorithm and the time taken and it is not possible to implement a large number of iterations in a real time system.

Experiments with multiple iterations also revealed another drawback of the algorithm. The manual clustering results represent the best case segmentation results. If the split/merge procedure is made automatic, it is probable that the algorithm may take a few more iterations to

converge, compared to the optimum manual procedure. All these factors make the algorithm unsuitable for use in real time defect detection systems.

At this point it is worthwhile to step aside, and examine the reason for obtaining the poor results using the multispectral clustering technique. By looking at the histograms and their color characteristics, good insight can be developed into how the value of the threshold T_i affects the formation of the clusters. For ease of visualization, most of the analysis will be explained in 2-dimensions. These ideas will then be extended to 3-dimensions, without the aid of graphics.



Figure 4.6: Image of a Yellow Poplar board y2a.dat

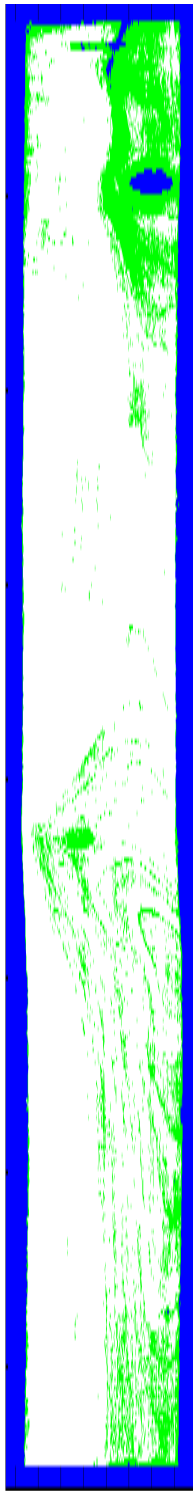


Figure 4.7: Result of segmentation of Figure 4.6 after 3 iterations

4.5 Histogram Characteristics

The histograms of boards have some unique features that will be described in this section. At first some general observations will be made about 2-dimensional histograms that are easier to visualize than 3-dimensional histograms. After describing these characteristics, an attempt will be made to explain the above described performance based on these characteristics.

It was found that each 3-dimensional r, g, b histogram of a board occupies only a relatively small volume of the full color space. This was true for all the boards used in the study. To demonstrate this, an oak board shown in Figure 4.8 will be used. The board has a few dark knots and the remaining area is *clearwood*. This is a very simple board in the sense that the clearwood region does not have a wide range of color, and the defects on the board are only knots. The total number of non-zero colors or elements in the histogram are 4587, which represents only about 1.75% of the total volume of full color space. A careful examination of the histogram data shows the presence of *holes* in the data, i.e., an element with value zero surrounded by elements with non zero value. These holes result because of the basic nature of histograms and, at times, the presence of noise. The subject of noise in the histograms, will be dealt with in the next section. Note noise was found to have a significant affect on the clustering process.

Some other general remarks can be made about the nature and shape of the board histograms, when observed in two dimensions. The two dimensional red-blue (r-b) and red-green (r-g) histograms of the oak board in Figure 4.8 are shown in Figures 4.9(a) and 4.9(b). The observations illustrated using this histogram is very typical for wood histograms, in general. There is generally one main peak that corresponds to the clear wood and some other small peaks that correspond to defects. However, these defect peaks are not prominent and, in most cases appear only as a tapering tail of the prominent clearwood peak. This, observation of course, results directly from the fact that defects almost always make up only a small percentage of a board surface.

It was experimentally found that the three components of color image of a board, the red, green and blue channels, generally have different variances. The height of the histogram in Figure 4.6b is lower than the height of the histogram in Figure 4.6c since the histogram is more spread out in the first case. The range of the three channels are shown below.

Channels	Minimum value	Maximum Value
Red	0	53
Green	0	44
Blue	0	63



Figure 4.8: Oak Board c116b.dat