Block-segmentation and Classification of Grayscale Postal Images

Kush R. Varshney
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853

Abstract

Features and algorithms for the block-segmentation and classification of documents are evaluated for applicability to scanned grayscale postal images. Classification is limited to three categories: background, picture, and text. Features derived from transform coefficient distributions prove ill-suited to postal images, whereas features based on spatial-domain statistics, edge information, and gradient direction prove effective. It is shown that a multiscale, context-based classification algorithm succeeds when used in conjunction with a well-suited, multidimensional feature space for first pass classification. It is also shown that execution time can be greatly reduced by the use of features that are not computationally intensive.

1 Introduction

The general problem of the segmentation and classification of document images into categories such as background, text, and picture, has numerous applications, such as printing processes that separately render different local image types [1], image compression for teleconferencing [2] and pre-processing prior to Optical Character Recognition (OCR) [3]. In the postal industry, address fields on envelopes must be located prior to being sent to OCR readers, and stamps must be located for evaluation by appropriate systems to ensure revenue protection [4].

Desktop publishing and teleconferencing applications tend to deal with data sets comprised of computer-generated images: images produced digitally, and processed digitally without the introduction of noise [5]. Postal images, on the other hand, are acquired by scanning envelopes and magazines in non-ideal conditions. There are many sources of noise, such as the physical imprinting of text onto envelopes, the handling of envelopes, and the scanning process. Thus, techniques suited to classifying computer-generated images cannot suitably classify scanned images. Classification techniques developed for computer-generated images include those based on wavelet transform coefficient distributions [1] and discrete cosine transform (DCT) coefficient distributions and energies [2]. Classification features that perform better for scanned documents include those based on edge information [2, 3, 6, 7], and pixel intensity distributions and statistics [2, 3, 5, 8].

There are, in general, three approaches to image segmentation: object-based segmentation, layer-based segmentation, and block-based segmentation [9]. I will focus specifically on block-based segmentation because it is simple and fast, and thus can be integrated into postal systems. The general procedure for block-based segmentation and classification partitions an image into non-overlapping n×n blocks. A transformation of the form $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ returns a scalar value which is used to classify the image block. I will refer to these scalar values as *features*. Block-segmentation and classification algorithms can be based on the information provided by a

multidimensional feature space. Algorithms can be improved by context-based classification supplementing first pass classification from feature information [1].

I will first discuss a feature that works well for computer-generated images, goodness of fit of wavelet transform coefficient distribution to the Laplacian distribution [1]. Then I will discuss features based directly on pixel intensity values in the spatial domain. After considering the effectiveness of a context-based classification component in a segmentation algorithm, I will give experimental results that show that first pass classification features matched to the postal data set yield better results than wavelet based classification. Additionally, I will show that these features improve execution time and thus can be integrated into postal systems.

2 Wavelet Transform Coefficient Distribution

One feature that can be used for the classification of image blocks is based on wavelet coefficient distributions [1]. As shown in Figure 1, the wavelet coefficients in high frequency bands of the Haar Transform tend to follow a Laplacian distribution for pictures, whereas for text, they are concentrated on a few values. Picture blocks have many varieties of transitions between adjacent pixels and thus produce a variety of wavelet coefficients, specifically a Laplacian distribution. Contrarily, text blocks have very few different transitions between adjacent pixels due to the bi-value nature of text. Therefore, only those wavelet coefficients corresponding to the pixel intensity transitions are produced. The classification feature that emerges from these observations is the goodness of fit of the wavelet coefficients to the Laplacian distribution. The goodness of fit can be quantified by the χ^2 test. It is interesting to note that the wavelet transform maintains spatial and frequency information, but making a histogram of the coefficients discards the spatial content. Also, histograms of the difference between adjacent pixels show the same trends: picture blocks have Laplacian distribution, and text blocks have distributions concentrated at certain values.
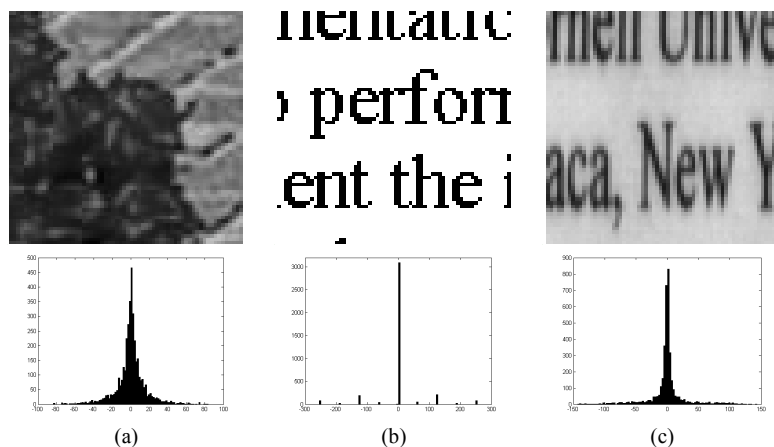

Figure 1. Histograms of Haar wavelet coefficients: (a) picture block, (b) computer-generated text block, and (c) postal text block. First row: original images. Second row: histograms.

The Laplacian χ^2 feature works well when applied to computer-generated images, but has relatively poor performance when applied to postal images [4]. As shown in Figure 1c, this is

because the text of postal images does not abruptly fall off from black to white, but descends smoothly, passing through many shades of gray. Consequently, a mixture of wavelet coefficients is produced, not just a few concentrated values. This is known as the *dirty text* problem [4]. Thus, the aptitude of the Laplacian χ^2 feature at distinguishing picture from text is hindered by the dirty text problem, to the point that classification error is around 25% for well-performing, pre-processed images with multiscale, context-based classification [4].

3 Classification Features Suited to Scanned Images

3.1 Histogram of Pixel Intensity

One feature based on the histogram of pixel intensity in a block is a count of the number of modes, where a mode is defined as a value that is a local maximum and the cumulative probability around it is above a pre-selected threshold [5]. Background blocks have one mode, text blocks have two modes, and picture blocks have many modes. This feature was not investigated because, as shown in Figure 2, text block histograms in postal images do not show the bi-valued property necessary for two modes. This is again a consequence of the dirty text problem. The histogram differences between pictures and text have also been used for classification by neural-network methods [8].
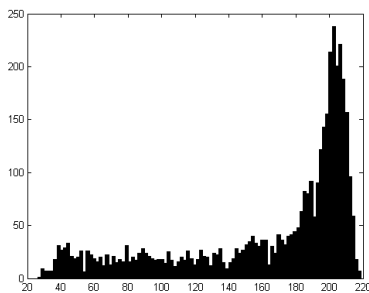
Figure 2. Pixel Intensity Histogram for Postal Text

3.2 Mean and Standard Deviation

The mean of pixel intensities (μ) in an image block is a useful feature for classification [3]. As Table 1 shows, the three classes, background, picture, and text, can be separated into three bands based on μ. However, non-white backgrounds and light pictures may cause misclassification. The feature μ is a good supplemental tool in classification, but should not be taken as primary because of the variety of envelope background colors.

The standard deviation of pixel intensities (σ) in an image block is a feasible feature for classification of scanned images [3]. It can be taken as a primary feature for identifying background blocks. As indicated in Table 1, σ is very small for background images. As can also be seen in the table, σ cannot be used to effectively discriminate pictures from text in postal images.

3

The related features variance and absolute deviation of pixel intensities have also been mentioned as classification features [2], and should perform similarly to σ.

Type	μ	σ
Background	230.97	3.47
Picture	98.04	68.09
Text	162.88	52.98

Table 1. Mean (μ) and Standard Deviation (σ) for trial postal image blocks.

3.3 Active Pixels

A count of active pixels (α) is defined as the number of pixels whose intensity falls below the threshold $\mu - k \cdot \sigma$, where μ and σ are as above, and k is a constant to be selected [3]. This feature follows the method of adaptive thresholding in the sense that a unique threshold is applied for each block of the image. Due to the preponderance of white pixels and presence of dark pixels in text blocks, the feature α can be used to segregate picture blocks and text blocks. The small σ value associated with background blocks limits the effectiveness of identifying background using α. The performance of this feature is not heavily reliant on the choice of k. Figure 3 shows α for different values of k in three postal image blocks.
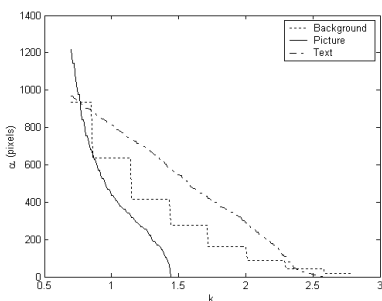
Figure 3. Active pixels (α) as a function of k for trial postal image blocks.

3.4 Sum of Second Derivatives of Average Intensity

Text blocks tend to have many edges, which implies that the average intensity tends to change greatly when traversing the block. A sum of average intensity second derivative magnitudes ($|I_{av,x}''|$) evaluated at all pixels along the block width gives a quantification of edges in the horizontal direction. A summation along the block height similarly provides a quantitative idea about edges in the vertical direction [3]. The features D_x and D_y are defined below.

$$D_x = \sum_{blockwidth} |I_{av,x}''| ;$$

$$D_y = \sum_{blockheight} |I_{av,y}''| ,$$

where $I_{av,x}''$ and $I_{av,y}''$ are second derivatives from the Savitzky-Golay filter. The Savitzky-Golay filter performs noise filtering while giving the second derivative. It is quick to calculate, requiring a single matrix multiplication. The Savitzky-Golay filter is subject to two parameters: degree and frame length. These can be manipulated for optimal classification performance by the features D_x and D_y. Smaller frame lengths preserve the fine structure of text characters and

4

thus, tend to perform better. The two features exhibit three bands of values for the three categories of image blocks, even when different frame lengths are used. D_y as a function of frame length is shown for three postal image blocks in Figure 4. These features are viable for block-classification.
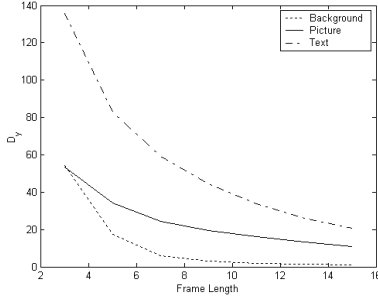

Figure 4. D_y as a function of frame length for trial postal image blocks.

3.5 Gradient Vector Direction

Neural-network methods have taken advantage of gradient vector direction histograms [8], but heretofore no scalar feature has been extracted. The gradient vector is easily defined in terms of its horizontal and vertical components:

$$\Delta h_{ij} = I_{(i+1)j} - I_{ij} \, ;$$
$$\Delta v_{ij} = I_{i(j+1)} - I_{ij} \, .$$

The vector gradient direction θ is then given by:

$$\theta_{ij} = \tan^{-1}(\Delta v_{ij}/\Delta h_{ij}) \, ,$$

and the magnitude $|\mathbf{g}|$ is defined by:

$$|\mathbf{g}|^2 = (\Delta h_{ij})^2 + (\Delta v_{ij})^2 .$$

As in Figure 5a, a histogram of θ having 24 15°-wide bins shows that for text blocks, many gradient vectors point in the four cardinal directions due to the alignment of text edges. Picture blocks and background blocks do not produce such peaks. Gradient vectors whose direction is in one of the four cardinal directions can be labeled as a *cardinal gradient vector*. A feature that can be extracted is the number of cardinal gradient vectors in the image block (γ_c).

$$\gamma_c = \sum_{i,j} (-\tfrac{13\pi}{24} \le \theta_{ij} < -\tfrac{11\pi}{24}) + (-\tfrac{\pi}{24} \le \theta_{ij} < \tfrac{\pi}{24}) + (\tfrac{11\pi}{24} \le \theta_{ij} < \tfrac{13\pi}{24}) + (\tfrac{23\pi}{24} \le \theta_{ij} < \pi) + (-\pi \le \theta_{ij} < -\tfrac{23\pi}{24})$$

However, γ_c discards magnitude information. Smooth regions, having little gradient magnitude and almost arbitrary gradient direction, influence the feature as much as regions with steep gradients. Therefore, a better feature sums the magnitude of the cardinal gradient vectors. Let this feature, the sum of cardinal gradient vector magnitudes, be denoted g_c. The histogram weighted by magnitude accentuates the peaks, as shown in Figure 5b, yielding a more effective classification feature. If the requisite condition that the document image be square aligned is met, both features can be used in block-segmentation and classification algorithms.

5
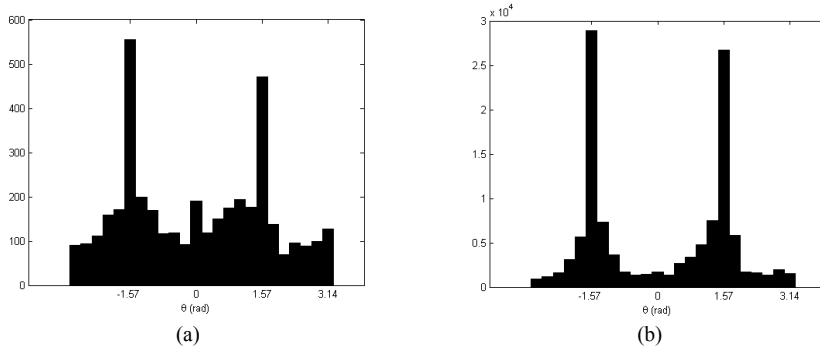
Figure 5. Histogram of text block gradient vector direction: (a) unweighted and (b) weighted.

4 Block-segmentation and Classification Algorithms

4.1 Simple Algorithm

The simplest block-segmentation and classification algorithms use a single block size. They break up the complete image into a grid of blocks and then use a single feature to classify individual blocks. In these algorithms, the choice of block size is critical. A large block enclosing a contiguous region of the same class will tend to have conspicuous features. However, the likelihood of containing multiple classes in a single block increases with increased block size. On the other hand, there is a higher probability that small blocks will contain a single class. Features, however, are less conclusive for classification.

4.2 Multidimensional Feature Space

When many features are used in conjunction to classify image blocks, classification error is reduced. The fusion of information provided by multiple features to produce a single decision or classification is, however, a very complicated problem. Thus, the development of the optimal decision rule to enable classification based on many features is difficult to achieve. The combination of information improves classification even if the best method to combine data is not employed.

4.3 Multiscale Architecture and Context-based Classification

Multiscale algorithms can take advantage of both the prominent features of large blocks and the likelihood of a single class in small blocks [1, 3]. This type of algorithm begins segmentation and classification with large blocks. If a large block cannot be classified, it is subdivided into four sub-blocks for further evaluation. This multiscale approach resolves an image starting with coarse classification and progressing to fine classification.

The multiscale technique is further enhanced by using context information [1]. Performance is improved when information garnered during the first pass classification of surrounding blocks is used to help classify unclassified blocks. Rules have been developed to implement such a context-based algorithm [1]. It has been shown that classification error is reduced when a context-based approach supplements first pass classification [1, 4] and thus is a useful component of a block-segmentation and classification algorithm.

5 Experiment

5.1 Development of Decision Tree

I selected the 6 features μ, σ, α, D_x, D_y, and g_c to use as the basis for classification. All are computationally fast and provide different types of information. I randomly selected 10 64×64 image blocks of each category: picture, text, and background, from the database of images produced by the postal systems of Lockheed Martin Systems Integration, Owego. After calculation of the six features for the image blocks, I developed a decision tree, shown in Figure 6, by inspection. More rigorous and elegant techniques may be employed in the future for better classification.
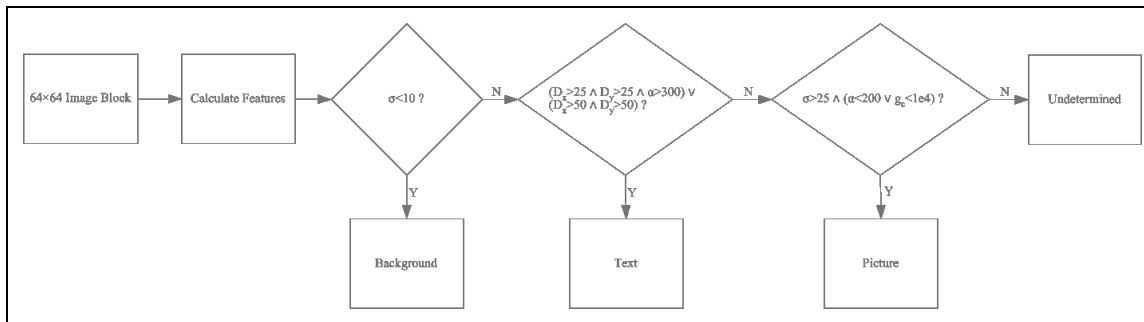
Figure 6. Decision tree for first pass classification.

5.2 Experimental Block-segmentation and Classification

After the development of the decision tree, 8 postal images were segmented using the front end provided by the decision tree and the back end provided by context-based classification [4]. The segmentation results are shown in Appendix A. Segmentation results are also shown for an image that failed with the wavelet-based approach, but performed well with the six-dimensional feature space approach.

5.3 Quantitative Analysis of Classification Error

Hand segmentation was performed on each of the postal images at 32×32 block resolution to obtain a truth segmentation and percent error was then calculated. The results are tabulated in Table 2. The table also shows results obtained from the wavelet-based approach [4]. Many misclassified image blocks were at the envelope boundary, which the wavelet-based approach classified as text and the 6-dimensional features space approach classified as picture, instead of background. Other classification error is a matter of precision rather than accuracy. The remaining error has its source in true misclassification.

5.4 Execution Time

Execution time is an important consideration for segmentation algorithms, especially for postal systems. The first pass classification based on the six-dimensional feature space runs much faster than the wavelet-based classification, as shown in Table 2.

Image ID	Wavelet	6-Dimensional	Wavelet	6-Dimensional
	Classification Error (%)		First Pass Execution Time (s)	
1	34.67	23.06	157.76	5.81
2	39.17	28.50	148.37	5.64
3	29.78	21.94	183.01	6.65
4	40.11	31.53	176.66	5.26
5	43.86	36.69	237.90	5.94
6	35.08	33.33	211.75	4.62
7	26.72	25.67	225.68	4.42
8	14.44	9.00	174.21	3.54

Table 2. Experimental results.

6 Conclusion

Spatial domain features enable good block-segmentation and classification of grayscale postal images with fast execution time. A multidimensional feature space based approach presents an area that can be optimized using sophisticated techniques, but yields good results even if a decision tree is developed by inspection. Context-based classification, when combined with a first pass classification scheme suited to the particular data set, decreases classification error.

7 References

[1] J. Li and R.M. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," *IEEE Trans. Image Processing*, vol. 9, pp. 1604-1616, Sept. 2000.

[2] I. Keslassy, M. Kalman, D. Wang, and B. Girod, "Classification of compound images based on transform coefficient likelihood," *Proc. 2001 Int. Conf. on Image Processing*, vol. 1, pp. 750-753.

[3] A. Suvichakorn, S. Watcharabusaracum, and W. Sinthupinyo, "Simple Layout Segmentation of Gray-Scale Document Images," in *Proc. Fifth IAPR Int. Workshop on Document Analysis Systems*.

[4] B. Lau, J. Chen, K. L. Ong, and K. Koo, "Segmentation of Postal Images Using Wavelet Techniques," Honors Project, School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, May 2002.

[5] X. Li and S. Lei, "Block-based segmentation and adaptive coding for visually lossless compression of scanned documents," *Proc. 2001 Int. Conf. on Image Processing*, vol. 3, pp. 450-453.

[6] C.-J. Park, J.-H. Jeon, T.-M. Koo, and H.-M. Choi, "An edge-based block segmentation and classification for document analysis with automatic character string extraction," *1996 IEEE Int. Conf. on Systems, Man, and Cybernetics*, vol. 1, pp. 707-712.

[7] Q. Yuan and C.L. Tan, "Text extraction from gray scale document images using edge information," *Proc. Sixth Int. Conf. on Document Analysis and Recognition*, pp. 302-306.

[8] S. Imade, S. Tatsuta, and T. Wada, "Segmentation and classification for mixed text/image documents using neural network," *Proc. Second Int. Conf. on Document Analysis and Recognition*, pp. 930-934.

[9] A. Said and A. Drukarev, "Simplified segmentation for compound image compression," *Proc. 1999 Int. Conf. on Image Processing*, vol. 1, pp. 229-233.
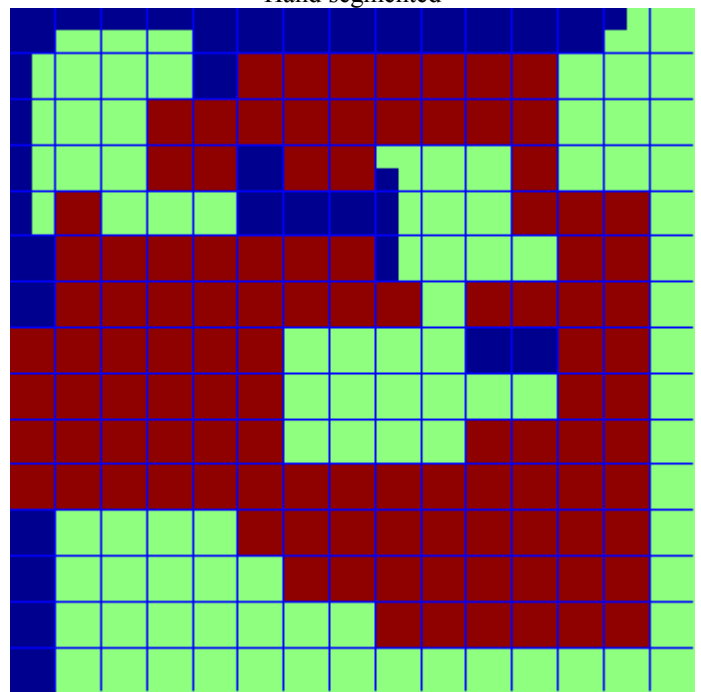
Appendix A

Image ID 1
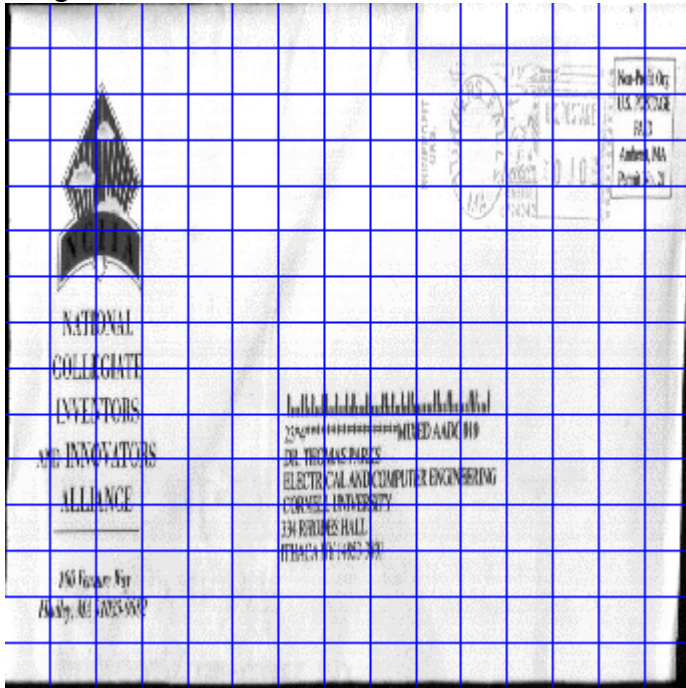
Image
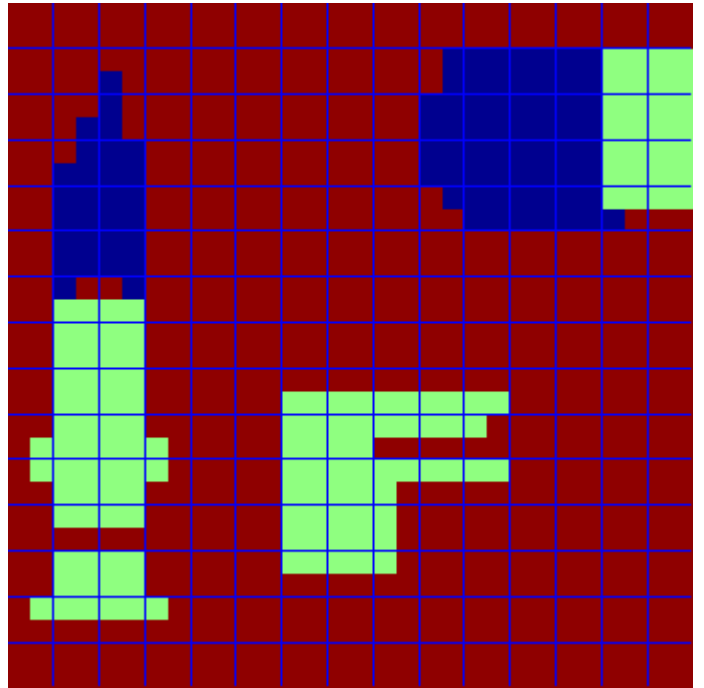
Hand segmented
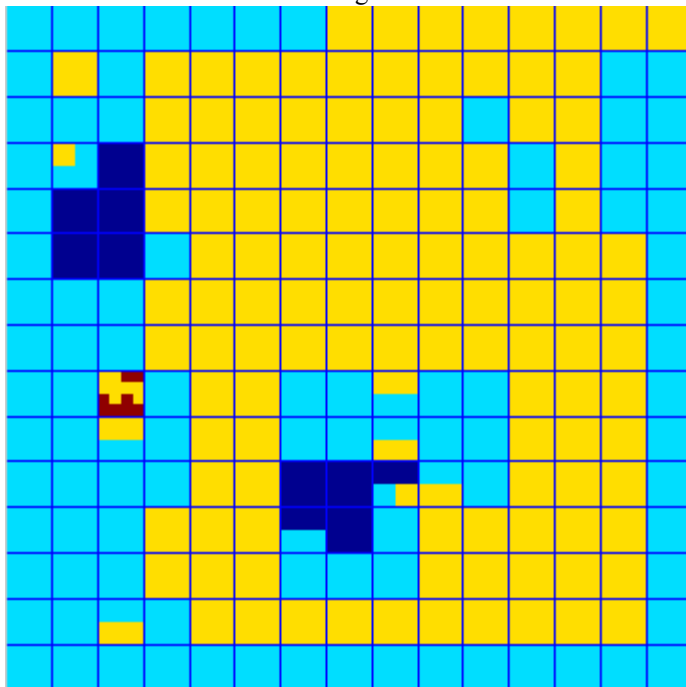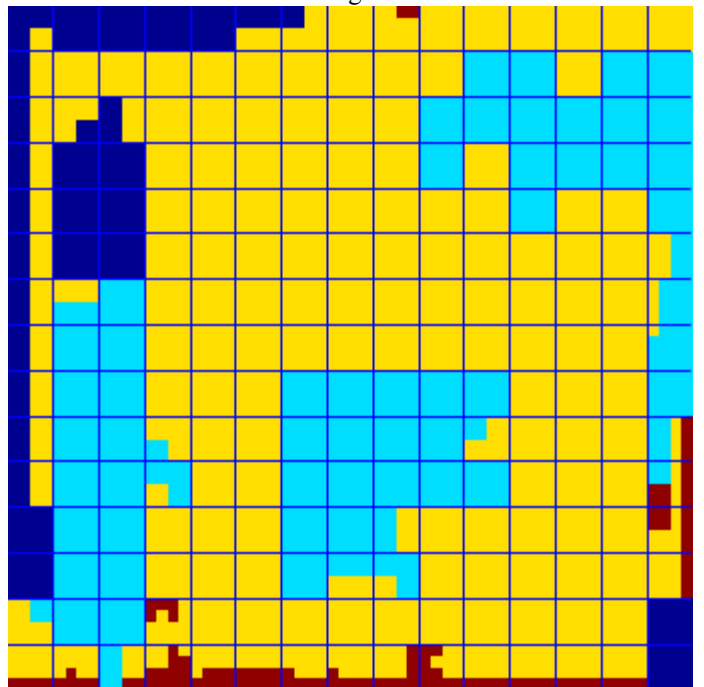
Wavelet Segmented

6-Dimensional Feature Space Segmented

Image ID 2
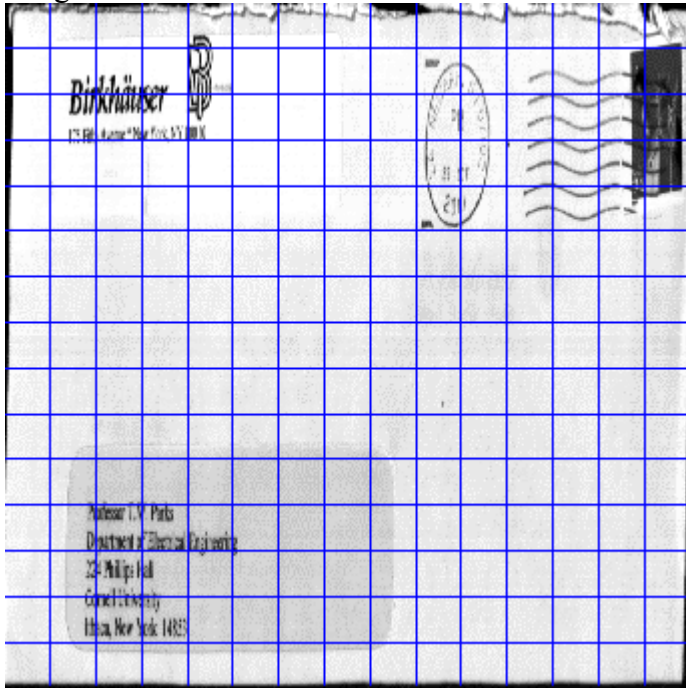
Image
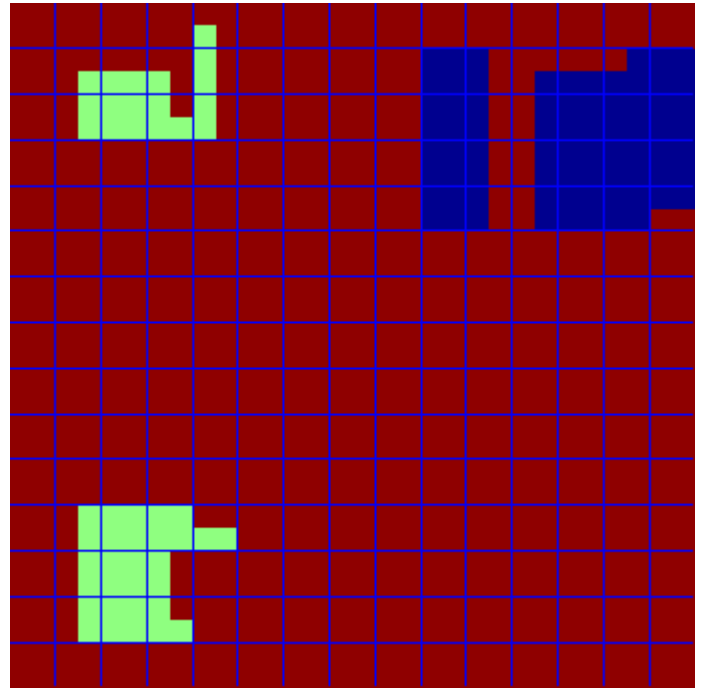
Hand segmented
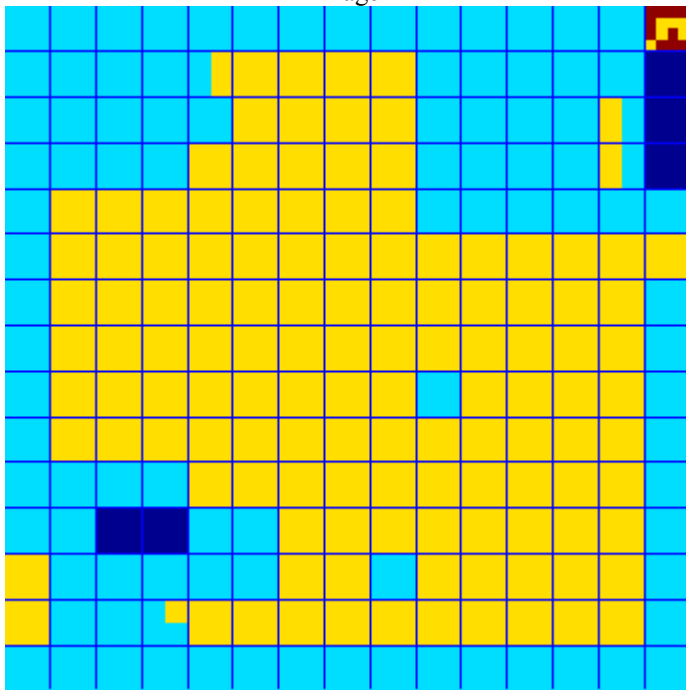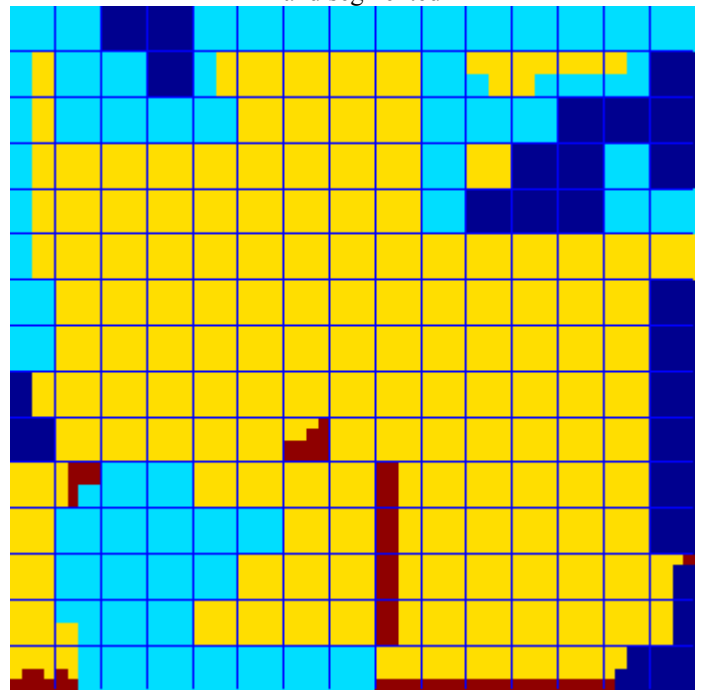
Wavelet Segmented

6-Dimensional Feature Space Segmented

Image ID 3
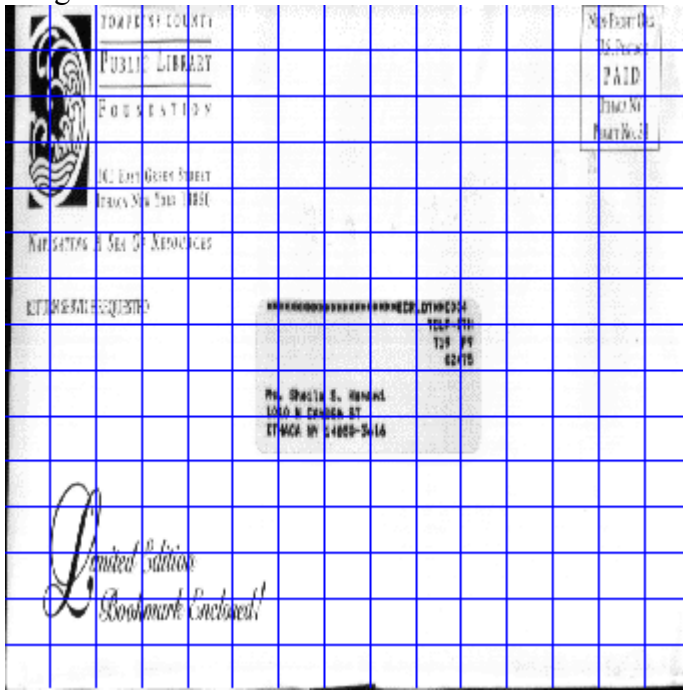
Image
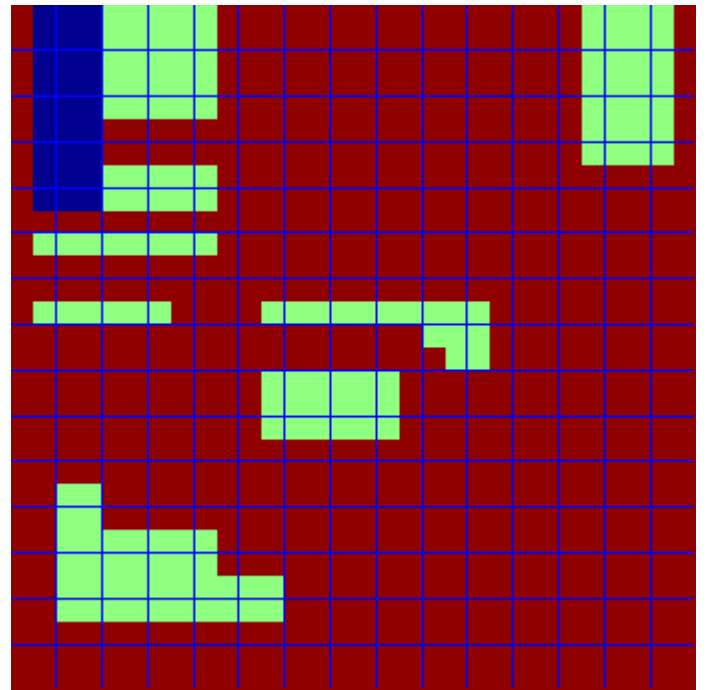
Hand segmented
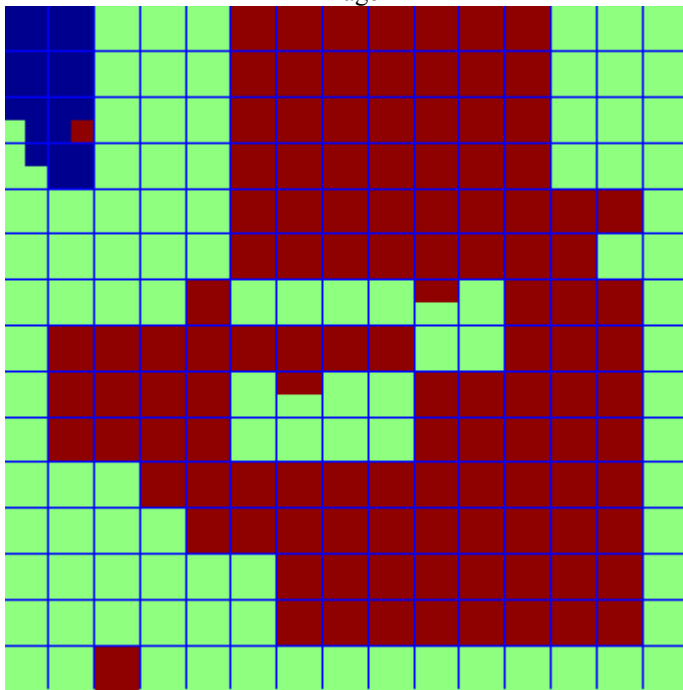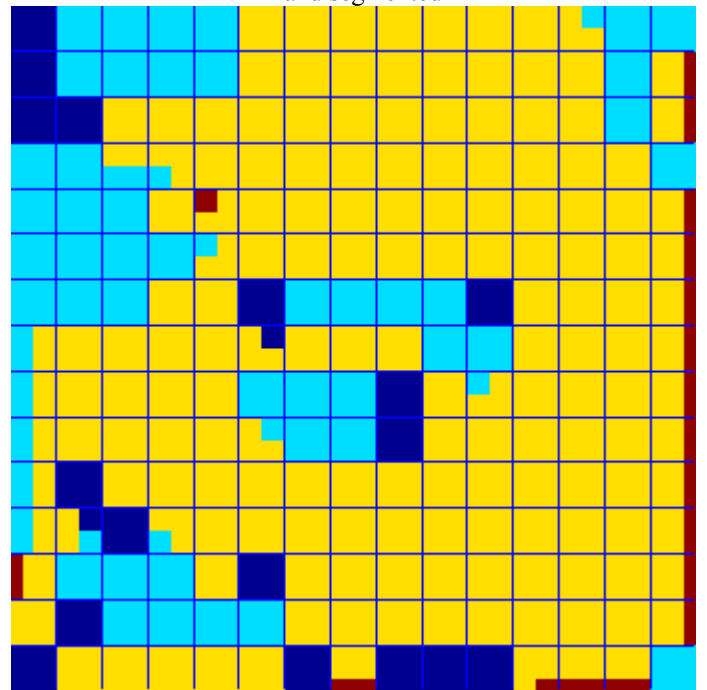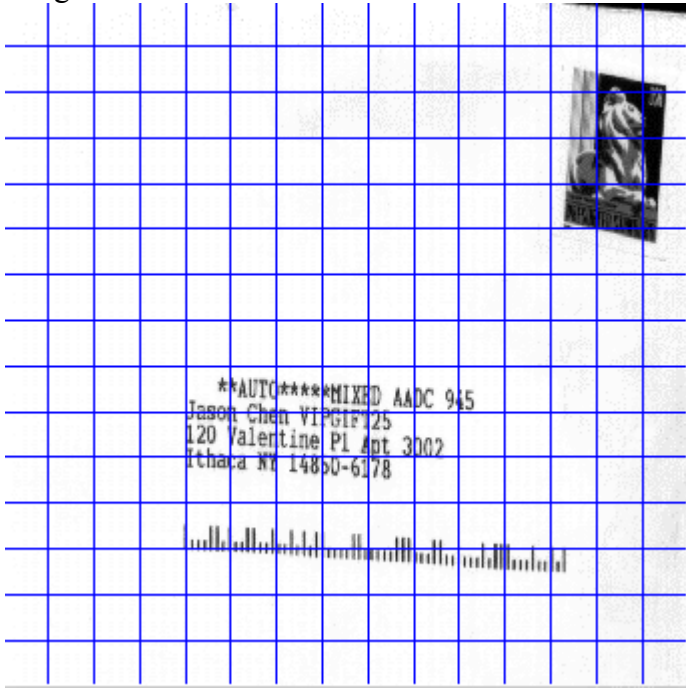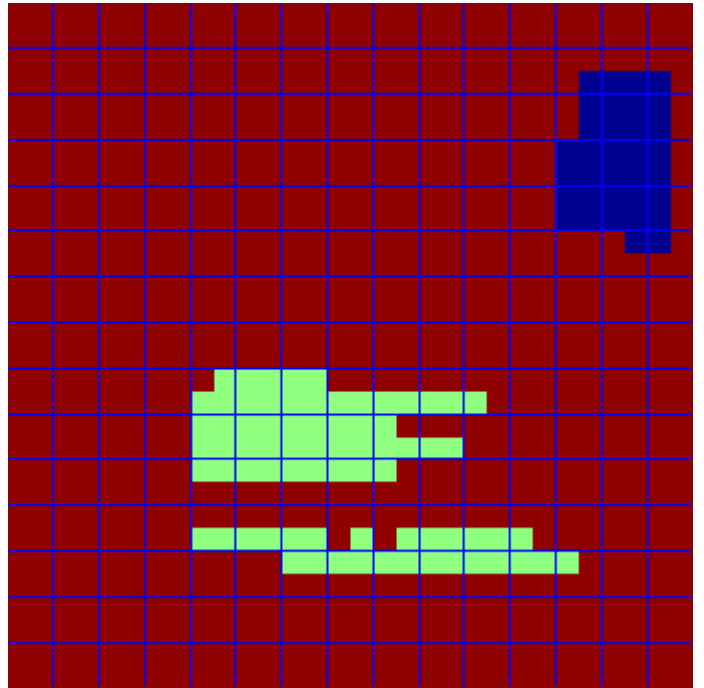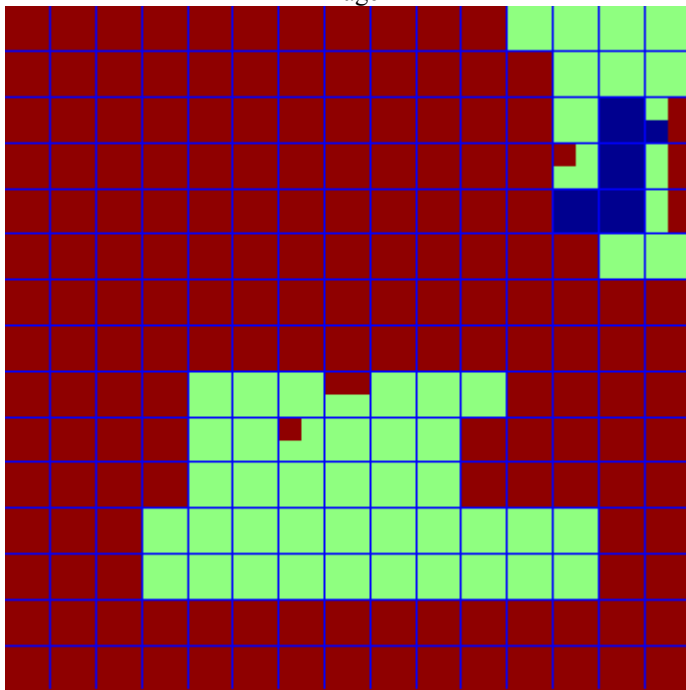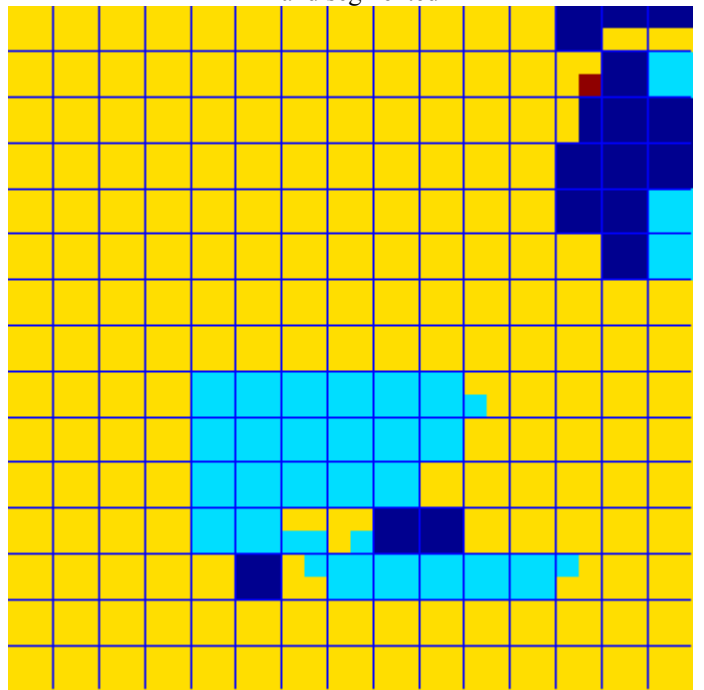

Wavelet Segmented
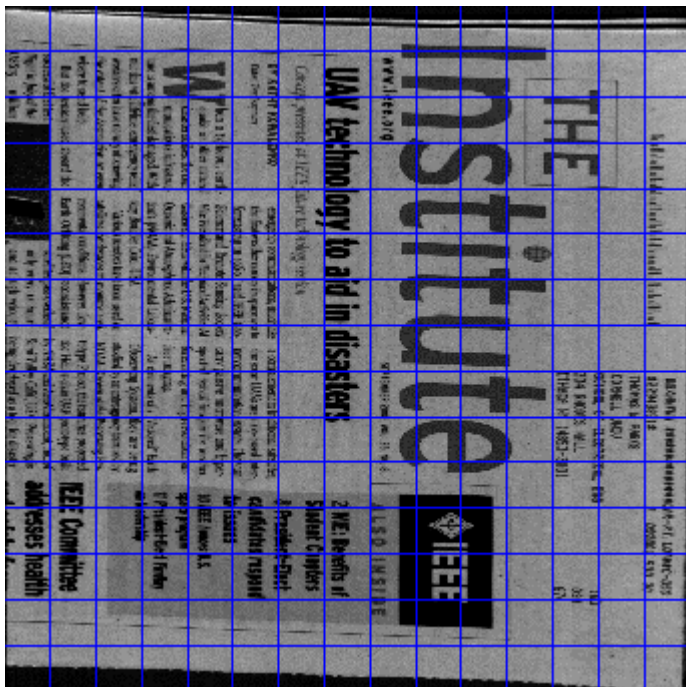

6-Dimensional Feature Space Segmented

Image ID 4
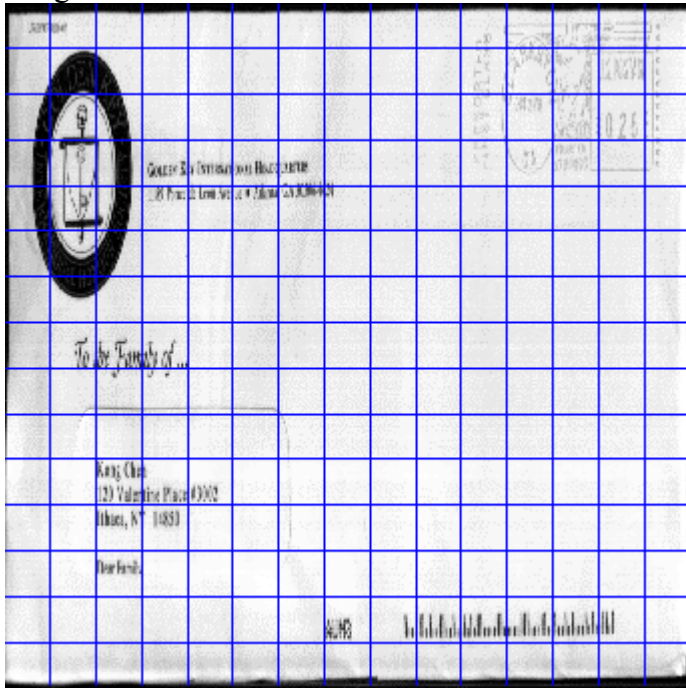
Image
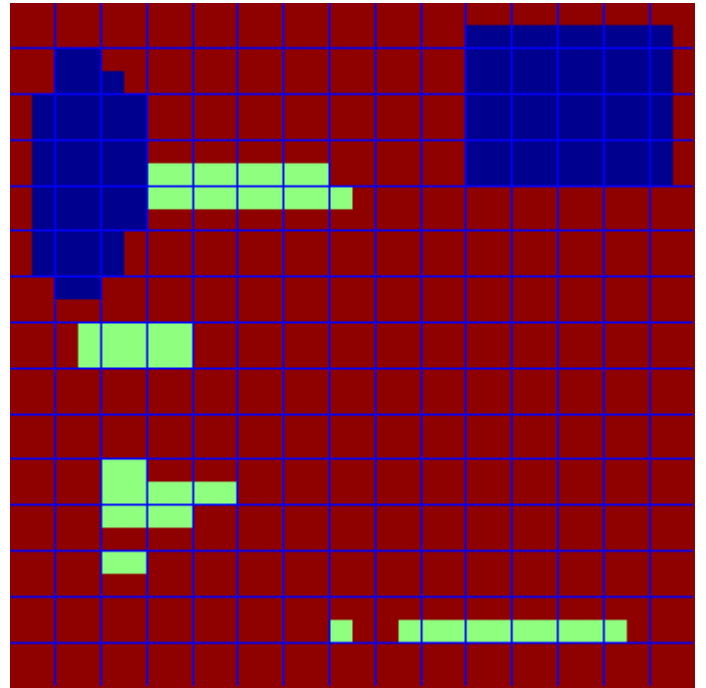
Hand segmented
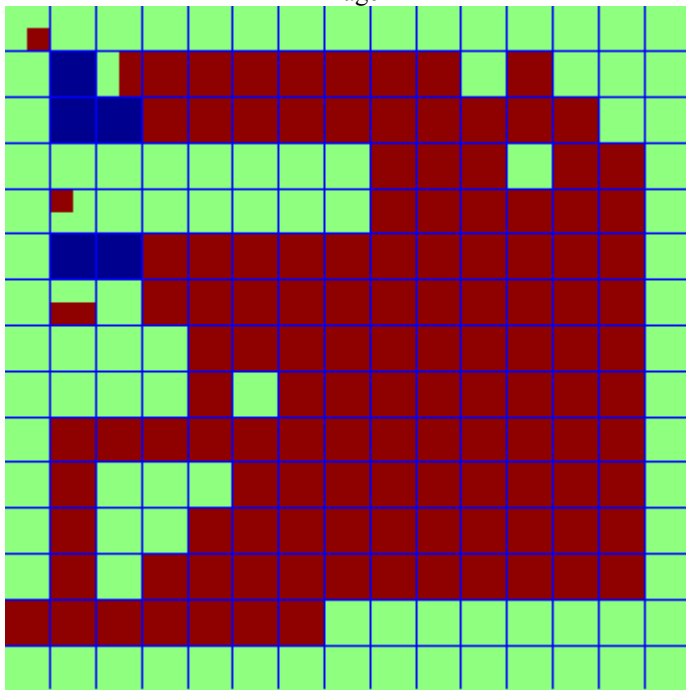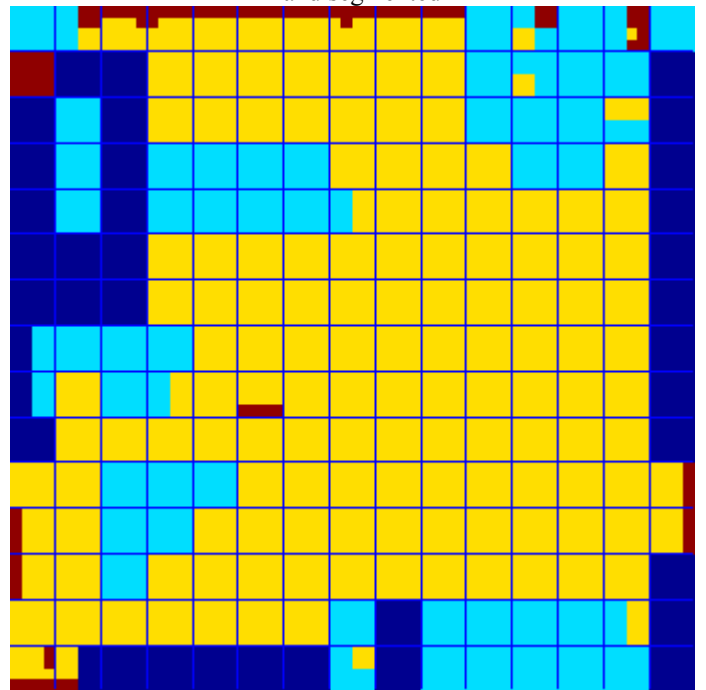
Wavelet Segmented

6-Dimensional Feature Space Segmented

Image ID 5
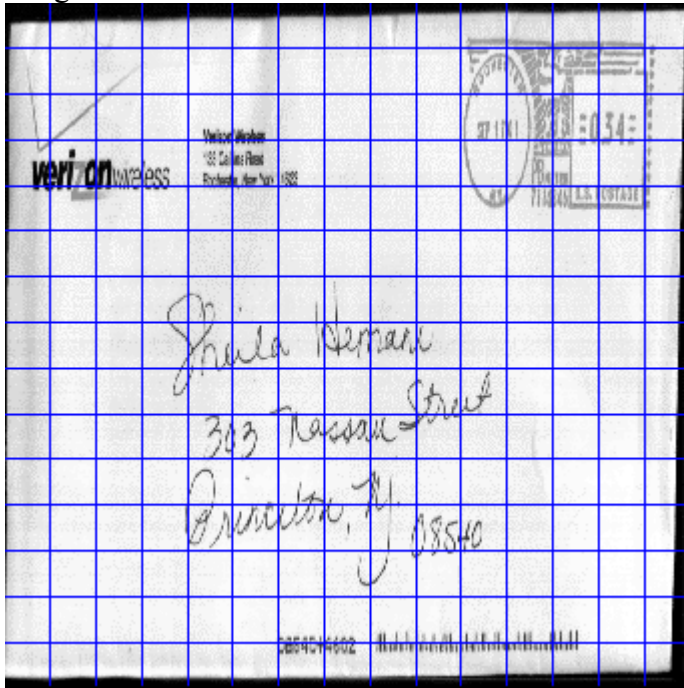
Image
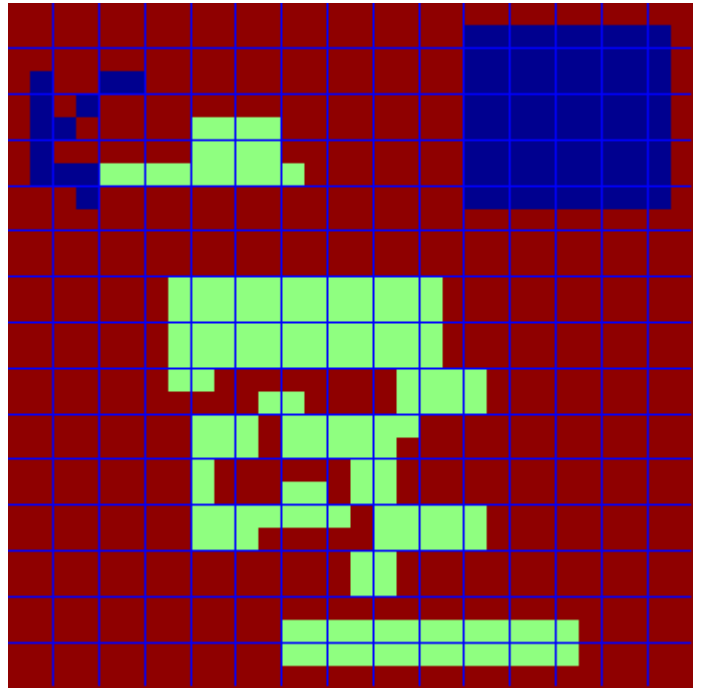
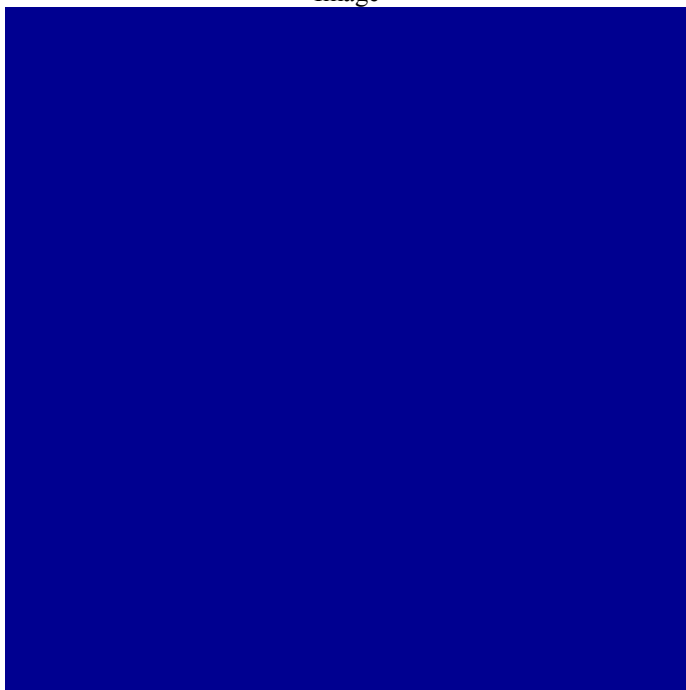

Hand segmented
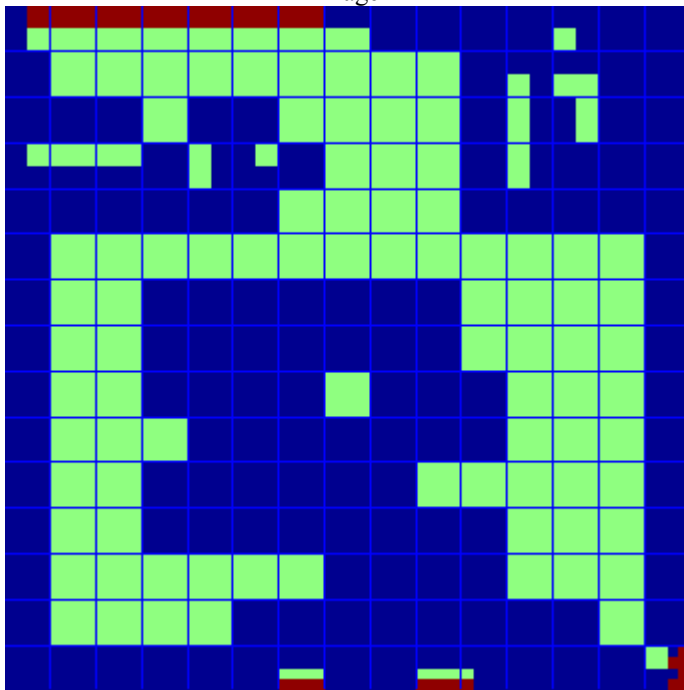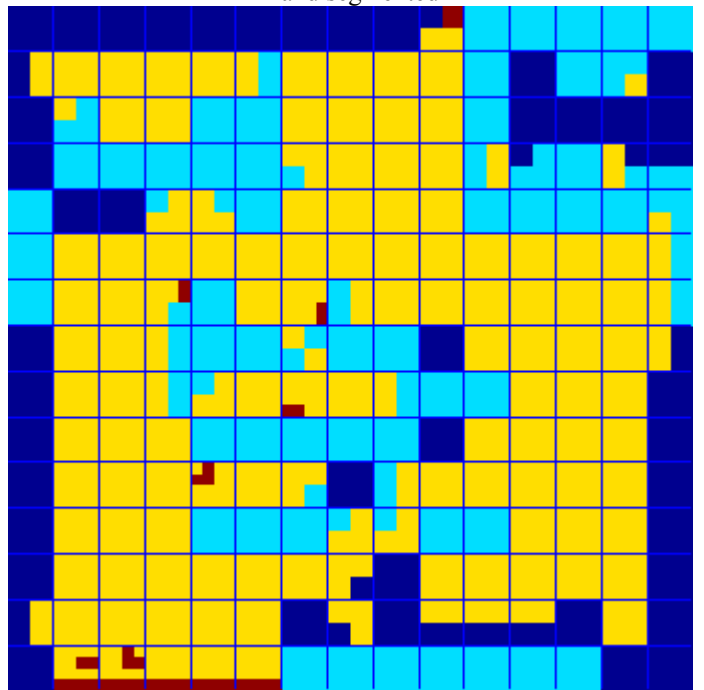
Wavelet Segmented

6-Dimensional Feature Space Segmented

Image ID 6
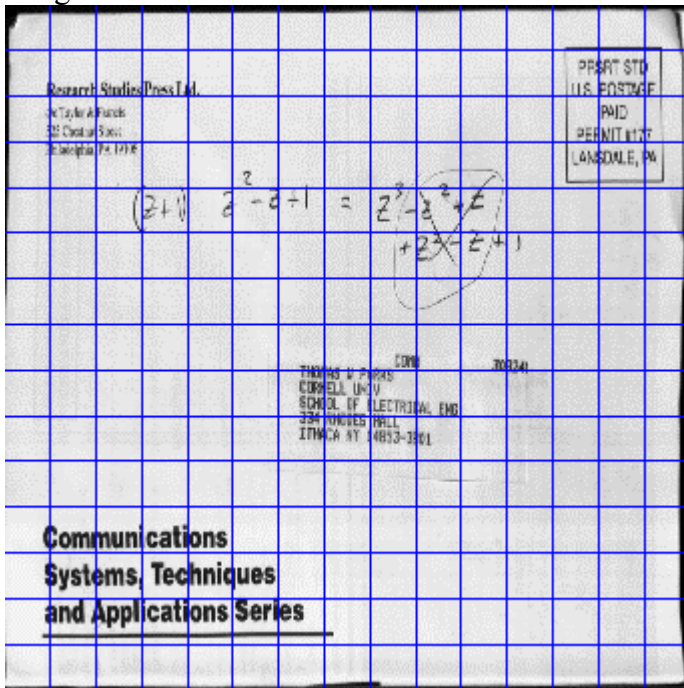
Image
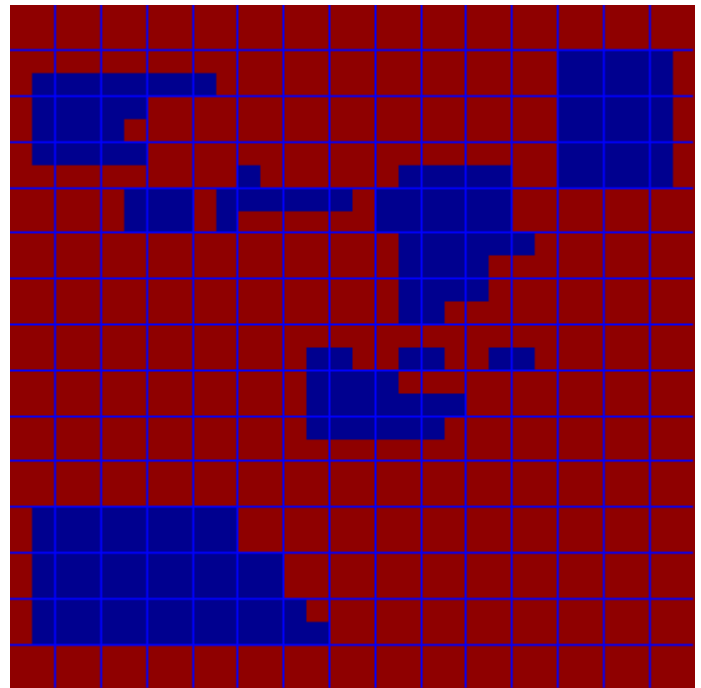
Hand segmented

Wavelet Segmented

6-Dimensional Feature Space Segmented

Image ID 7

Image

Hand segmented

Wavelet Segmented

6-Dimensional Feature Space Segmented

16
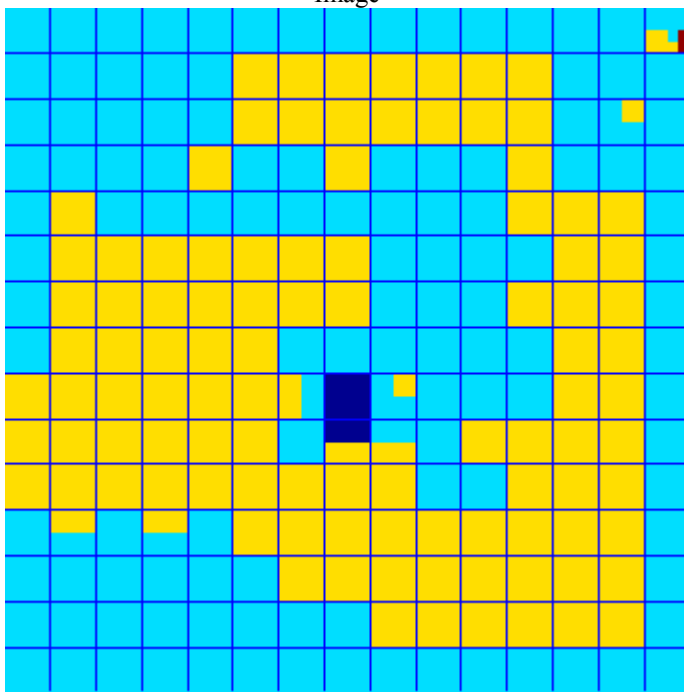
Image ID 8

Image

Hand segmented
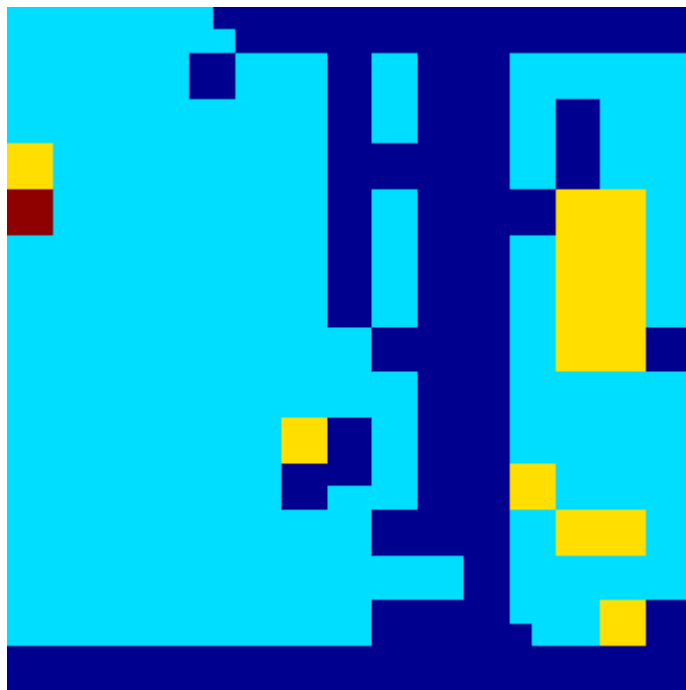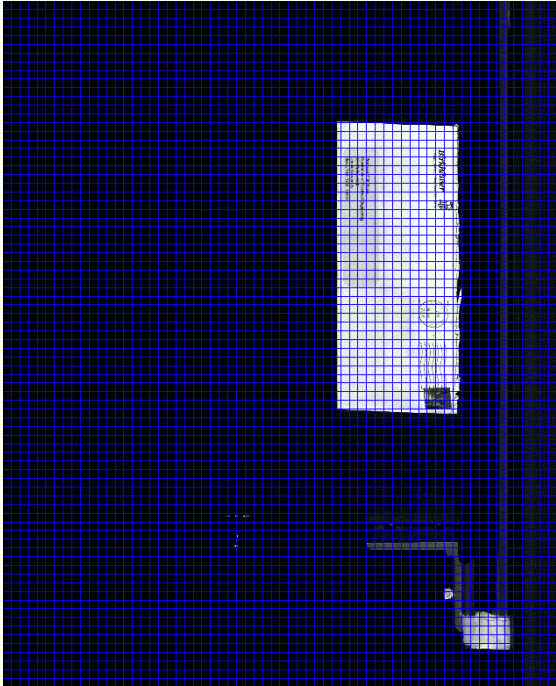

Wavelet Segmented

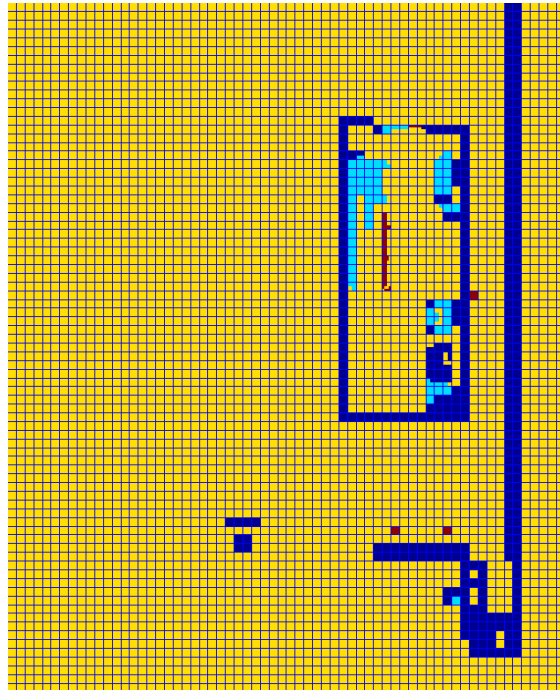6-Dimensional Feature Space Segmented

Image

Wavelet Segmented

6-Dimensional Feature Space Segmented

Image

6-Dimensional Feature Space Segmented

Appendix B

Guide to Matlab Code

The block-segmentation and classification algorithm can be executed as follows:
```
C = postalseg('filename');
```

More information can be found by the command:
```
help postalseg
```

A list of Matlab functions used by the algorithm is separated into the functions taken unmodified from [4] and the functions that are new or modified in Table B.1.

Unmodified	New or Modified
bivalue2	activepixels
checkAdjNeighbor	classify64
createList	DxDy
handsegment*	features
L_classifier	firstpass
meanSD	graddir
microclassifier	postalseg
step12	splitCI

Table B.1. Matlab functions.
* Utility for experimentation, but not used by algorithm.