Chapter 2

Development of Document Image Preprocessor using Hybrid Feature Extraction Technique

2.1 Introduction

The aim of automatic document processing system is to convert a scanned document image into symbolic form such as ASCII symbols for easy computer processing (e.g., modification, storage, retrieval). A scanned document image is a pattern. In the document recognition system, a pattern may be a whole document, a line of text, a printed or handwritten word or even a single isolated character [21], [22], [23]. Sample document image patterns are as shown in figure 2.1.

Figure 2.1: Sample document images: (a) printed document, (b) handwritten Kannada document, (c) printed Telugu document, (d) a single word, (e) a single character.

These patterns need to be properly represented for identification by pattern recognition algorithms. Pattern recognition algorithms essentially consist of two modules: feature extractor and classifier [24]. The feature extractor represents the pattern to the

Some parts of the material in this chapter appear in the following paper:

classifier. The classifier assigns symbols of class to the pattern presented to it. The aim of the feature extractor is to describe a given pattern by means of the minimum number of features or attributes that are effective in discriminating pattern classes.

In this chapter, the concept of feature extraction is introduced. Details of development of preprocessors for recognition systems, i.e. the feature extraction modules are discussed. Thus this chapter forms the basis for subsequent chapters. The organization of the chapter is as follows. The meaning of the word 'feature' is discussed in section 2.2. Existing methods available for feature extraction of document images are explained in section 2.3. This section concentrates on the methods which are used more often along with neural networks. The developed method of feature extraction is described in section 2.4. Experiments and results are reported in section 2.5. Detailed discussions and comparisons are presented in section 2.6. Finally conclusions are made in section 2.7.

2.2 Feature

A feature is a characteristic attribute of a pattern or an object. A single pattern may be described or represented by a set of such features. These should be distinct and discriminative to identify or classify the pattern [25], [26]. For example consider the following illustrations.

Illustration 1: For identification of a child, a grown up person and an elderly person, the feature age can be very effective.

Illustration 2: For identification of a cat or tiger, height, colour, and length are very discriminative features.

The process of finding these features in a given pattern or an object is feature extraction. A set of features are required to represent a document image for a recognition system. Features can be real values. These values can be obtained for each document to be represented by various methods as discussed in the following section.

2.3 Existing methods

A feature extraction method is used to reduce the bitmap image of a sample pattern. The primary purpose of feature extraction is reduction of the dimensionality. There are two main reasons to keep the dimensionality of the pattern representation (i.e., the number of
features) as small as possible: measurement cost and classification accuracy [27]. A limited yet salient feature set simplifies both pattern representation and the classifiers that are built on the selected representation. Consequently, the resulting classifier is faster and uses less memory [28].

Feature extraction is largely dependent on the task, input and recognition algorithm used. If statistical pattern recognizers are used, the standard features tend to be based on contours of pattern, with features such as Fourier descriptors, moment invariants and other boundary features. If structural recognizers are used, features that represent the structure of the pattern are preferred [29].

Practical applications are often demanding as regards throughput, and this limits the acceptable complexity of the feature extractor, as well as any other module in a document reading system [30]. Feature extractors which make use of sub-functions as contour extraction or skeletonization are time-consuming algorithms, besides being hazardous when applied to damaged, incomplete, or noisy bitmaps. Moreover neural network classifiers impose the condition that the same number (and kind) of features has to be produced for every sample; this is not true in the case of most structural feature extractors [30]. The existing simple methods of feature extraction for use with neural network recognizers are as follows.

2.3.1 Simple bar mask encoder

When the application of artificial neural networks is confined to recognition in particular, it is advantageous to concentrate more on the black pixel distribution of the pattern. This idea has been incorporated in the bar mask encoding technique of feature extraction.

A bar mask encoder is similar to the seven segment alphanumeric display. In this method the pattern bitmap is divided into several regions as in the display. These regions are called capture regions. In each region the number of '1's or 'ON' pixels is counted. A feature of the region is obtained by dividing this value by the total number of pixels (including both '1's and '0's).

Illustration: The following image of digit '9' has been divided into four horizontal regions and three vertical regions, which overlap one another (figure 2.2). Again it is divided into nine non-overlapping regions also. In each region the numbers of 'ON'
pixels are counted and this count is divided by the total number of features, resulting to a real value. This is the value of the feature of that region. In this case 16 (4+3+9) such feature values can be obtained from the input image.

Various alterations of this method of feature extraction are used in the recent literature. The different forms of a simple bar mask encoder can be found in [31], [32]. The development trend seems to be focused mainly on changing the shapes of the regions to identify particular shapes in patterns. The algorithm 2.1 presents the generalized steps involved in such a method of feature extraction.

Algorithm 2.1: Feature extraction by simple bar mask encoder
1. Receive the input document image.
2. Threshold the input image into a two-tone image. Now only '1's or '0's are present in the image.
3. Divide the original image into $h$ horizontal regions.
4. Divide the original image into $v$ vertical regions. These regions may overlap with the above horizontal regions.
5. For first region (out of $h + v$ regions) count the number of 'ON' (1's) pixels. Count the total number of pixels in the region. Calculate feature_1 = number of 1's ('ON' pixels) / number of total pixels.

6. Repeat step 5 for all the regions and find feature_2, feature_3...feature $(h + v)$ (e.g. If $h = 4$, $v = 3$, total features = 7).

2.3.2 Moment invariant feature extraction

The moment invariant method provides moments that are invariant to translation, rotation and also for scale change [33]. For a 2D continuous function, $f(x, y)$ the moment order $(p + q)$ is defined by

$$m_{pq} = \iint x^p y^q f(x, y) \, dx \, dy, \quad \text{for } p, q = 0, 1, 2 \quad (2.1)$$

The uniqueness theorem states that if $f(x, y)$ is piecewise continuous and has a nonzero values only in a finite part of the x-y plane, moments of all the orders exist and the moment sequence $(m_{pq})$ is uniquely determined by $f(x, y)$. Conversely, $(m_{pq})$ uniquely determines $f(x, y)$. For a digital image, $f(x, y)$ is the grey value of the pixel $(x, y)$.

Moments can be given a physical interpretation if $f(x, y)$ is considered as the mass at $(x, y)$ rather than grey value. Then $m_{00}$ is the total mass of $f$, and $m_{02}$ and $m_{20}$ are moments of inertia of $f$ around x and y axes respectively. The moments can be converted into features invariant under translation, rotation and scaling. To take care of translation consider the centroid or 'centre of gravity' $(x, y)$ given by

$$\bar{x} = m_{10} / m_{00} \quad \text{and} \quad \bar{y} = m_{01} / m_{00}$$

It can be seen that if the image is translated by some amount, then its centroid is also translated by the same amount. Hence, the moments originating at the centroid are translation invariant. Such moments are called the central moments and are defined as

$$\mu_{pq} = \iint (x - \bar{x})^p (y - \bar{y})^q f(x, y) \, dx \, dy \quad (2.2)$$

Where $\bar{x} = m_{10} / m_{00}$ and $\bar{y} = m_{01} / m_{00}$

For a digital image, equation (2.2) becomes

$$\mu_{pq} = \sum \sum (x - \bar{x})^p (y - \bar{y})^q f(x, y)$$

The central moments of order up to three are

$$\mu_{10} = \sum \sum (x - \bar{x})^1 (y - \bar{y})^0 f(x, y)$$

$$= m_{10} - (m_{10} / m_{00}) (m_{00})$$
The normalised central moments, denoted $\eta_{pq}$ are defined as
$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}}$$
where
$$\gamma = \frac{p + q}{2} + 1$$
for $p + q = 2, 3, \ldots$

A set of seven invariant moments can be derived from the second and third moments
\begin{align*}
\Phi_1 &= \eta_{20} + \eta_{02} \\
\Phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\Phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\Phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\Phi_5 &= (\eta_{30} - 3\eta_{12}) (\eta_{30} + \eta_{12}) [ (\eta_{30} + \eta_{12})^2 \\
&\quad - 3(\eta_{21} + \eta_{03})^2 ] + (3\eta_{21} - \eta_{03}) (\eta_{21} + \eta_{03}) \\
&\quad [ 3 (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 ]
\end{align*}
\[ \Phi_6 = (\eta_{20} - \eta_{02}) \left[ (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \right] \\
+ 4 \eta_{11} (\eta_{30} + \eta_{12}) (\eta_{21} + \eta_{03}) \\
\Phi_7 = (3\eta_{21} - \eta_{03}) (\eta_{30} + \eta_{12}) \left[ (\eta_{30} + \eta_{12})^2 \\
- 3 (\eta_{21} + \eta_{03})^2 \right] + (3\eta_{12} - \eta_{30}) (\eta_{21} + \eta_{03}) \\
[3 (\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \]

This set of moments is invariant to translation, rotation and scale change. Moments only up to third order are used.

**Algorithm 2.2: Moment invariant method of feature extraction**

1. Read the scanned image as a matrix.
2. Threshold the input image into a two-tone image. Now only '1's or '0's are present in the image.
3. Calculate the 1st, 2nd and 3rd order moments.
   \((m_{00}, m_{10}, m_{11}, m_{20}, m_{02}, m_{03}, m_{30}, m_{12}, m_{21})\) are calculated using the equations.
   \[ m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) \, dx \, dy, \]
   where as \(f(x, y)\) is selected part \((x, y)\)
   \[ m_{00} = \Sigma \Sigma (x)^0 (y)^0 f(x, y) \]
   \[ m_{10} = \Sigma \Sigma (x)^1 (y)^0 f(x, y) \]
   \[ m_{11} = \Sigma \Sigma (x)^1 (y)^1 f(x, y) \]
   \[ m_{20} = \Sigma \Sigma (x)^2 (y)^0 f(x, y) \]
   \[ m_{02} = \Sigma \Sigma (x)^0 (y)^2 f(x, y) \]
   \[ m_{03} = \Sigma \Sigma (x)^0 (y)^3 f(x, y) \]
   \[ m_{30} = \Sigma \Sigma (x)^3 (y)^0 f(x, y) \]
   \[ m_{21} = \Sigma \Sigma (x)^2 (y)^1 f(x, y) \]
   \[ m_{12} = \Sigma \Sigma (x)^1 (y)^2 f(x, y) \]
4. Central moments are calculated by substituting the above 1st, 2nd and 3rd order moments into the following equations.
   \[ \mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q f(x, y) \]
   \(\mu_{00} = m_{00}\)
   \(\mu_{10} = 0\)
   \(\mu_{01} = 0\)
\[\mu_{10} = m_{20} - x m_{10}\]
\[\mu_{02} = m_{02} - y m_{01}\]
\[\mu_{11} = m_{11} - y m_{10}\]
\[\mu_{30} = m_{30} - 3 x m_{02} + (2 x^2 m_{10})\]
\[\mu_{12} = m_{12} - 2 y m_{11} - x m_{02} + 2 y^2 m_{10}\]
\[\mu_{21} = m_{21} - 2 x m_{11} - y m_{20} + 2 x^2 m_{01}\]
\[\mu_{03} = m_{03} - 3 y m_{02} + 2 y^3 m_{01}\]

5. Calculate normalized central moments

\[\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^p}\]
where \(\gamma = (p + q)/2 + 1\), for \(p + q = 2, 3, \ldots\)

\[\eta_{20} = \frac{\mu_{20}}{\mu_{00}^2}, \quad \eta_{02} = \frac{\mu_{02}}{\mu_{00}^2}, \quad \eta_{11} = \frac{\mu_{11}}{\mu_{00}^2},\]
\[\eta_{21} = \frac{\mu_{21}}{\mu_{00}^{2.5}}, \quad \eta_{12} = \frac{\mu_{12}}{\mu_{00}^{2.5}}, \quad \eta_{03} = \frac{\mu_{03}}{\mu_{00}^{2.5}}.\]
\[\eta_{30} = \frac{\mu_{30}}{\mu_{00}^{2.5}}.\]

6. Calculate the set of seven invariant moments using the following equations:

\[\Phi_1 = \eta_{20} + \eta_{02}\]
\[\Phi_2 = (\eta_{20} - \eta_{02})^2 + 4 \eta_{11}\]
\[\Phi_3 = (\eta_{30} - 3 \eta_{12})^2 + (3 \eta_{21} - \eta_{03})^2\]
\[\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2\]
\[\Phi_5 = (\eta_{30} - 3 \eta_{12}) + (\eta_{30} + \eta_{12}) [((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03}) (\eta_{21} + \eta_{03})]
\[+ 3(\eta_{30} + \eta_{12})^2 - (\eta_{21} - \eta_{03})]\]
\[\Phi_6 = (\eta_{20} - \eta_{02}) [((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\[+ 4 \eta_{11} (\eta_{30} + \eta_{12}) (\eta_{21} + \eta_{03})\]
\[\Phi_7 = (3 \eta_{21} - \eta_{03}) (\eta_{30} + \eta_{12}) [((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]
\[+ (3 \eta_{12} - \eta_{03}) (\eta_{21} + \eta_{03})
\[+ 3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\]

2.4 Developed method

A feature extraction method is said to be effective, if it produces features which are discriminative among classes. That is, with a single feature it should be possible to classify the different classes present in a group of patterns. The above discussed moment
invariant feature extraction method produced the features (shown in table 2.1), which are not discriminative. A simple bar mask encoder method produced slightly better features but still may not be sufficient for recognition. In order to produce more discriminative features the following new method of feature extraction is developed.

In the simple bar mask method, the features obtained present the black pixel distribution in the selected regions. The information about the directions is completely lost. If the information about the directions of the majority of pixels can be captured, that may serve as being more discriminative among the classes. Mathematical morphology [34], [35], [2] is useful to determine these types of directional features.

Among the morphological operators, the dilation operator can be used effectively. The dilation can be explained as follows. Let A and B be two subsets (in the present case, one can be a document image, the other can be a structuring element) in two-dimensional space $E^2$ representing two object regions containing elements $\{a = (a_1, a_2)\}$ and $\{b = (b_1, b_2)\}$, respectively. Here, $a_1$ and $a_2$ denote $x$ and $y$ co-ordinates of the element $a$. Then, the dilation of $A$ by $B$, denoted by $A \ominus B$, is defined as

$$A \ominus B = \{c \in E^2 | c = a + b; a \in A, b \in B\}$$

where $a + b = \{(a_1 + b_1), (a_2 + b_2)\}$ etc.

As an example consider

$A = \{(0, 1), (1, 1), (2, 1), (2, 2), (3, 0)\}$

$B = \{(0, 0), (0, 1)\}$

Then

$A \ominus B = \{(0, 1), (1, 1), (2, 1), (2, 2), (3, 0), (0, 2), (1, 2), (2, 2), (2, 3), (3, 1)\}$

This is graphically shown in figure 2.3. The method to describe this process is given as algorithm 2.3.
Algorithm 2.3: Process of dilation

1. Read the input image Im1 and the size of input image matrix $m$ rows and $n$ columns.

2. Read the structuring element matrix to decide the pixels to be modified. e.g., if
   \[
   SE1 = \begin{bmatrix}
   0 & 0 & 0 \\
   1 & 1 & 1 \\
   0 & 0 & 0 
   \end{bmatrix}
   \]
   then modify, $(i, j - 1)^{th}$ pixel and $(i, j + 1)^{th}$ pixel to 1 if $(i, j)^{th}$ pixel is 1.

3. Create the matrix Im2 of the same size with zero elements.

4. Make $i = 1$.

5. Make $j = 1$.

6. Assign element $\leftarrow Im1 (i, j)$.

7. If element $\leftarrow 1$ /* for the horizontal structuring element shown above */
   
   \[
   \begin{align*}
   x1 &= j - 1 \\
   x2 &= j + 1 
   \end{align*}
   \]
if \( x_i \geq 1 \)

\[ I_{m2}(i, j - 1) = 1 \]

eendif

if \( x_2 \leq n \)

\[ I_{m2}(i, j) = 1 \]

eendif

\[ I_{m2}(i, j) = 1 \]

}\}

8. Make \( j = j + 1 \), if \( j \leq n \) repeat steps 6 and 7.

9. Make \( i = i + 1 \), if \( i \leq m \) repeat steps 5, 6 and 7.

By combining the algorithms 2.1 and 2.3, a new method of feature extraction is developed. In this method, each sample document image is dilated in the desired number of directions (e.g. in the experimented case, each sample document image is dilated in horizontal and vertical directions) using structuring elements such as

\[
SE_1 = \begin{bmatrix}
0 & 0 & 0 \\
1 & 1 & 1 \\
0 & 0 & 0
\end{bmatrix}
\]

\[
SE_2 = \begin{bmatrix}
0 & 0 & 1 \\
1 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix}
\]

thus capturing the information about the direction of pixels in an image. \( SE_1 \) captures horizontal information and \( SE_2 \) captures vertical information from the input sample. This process of dilation yields new modified images of the same size as that of original corresponding to the number of structuring elements. For each of such a modified image, simple bar mask encoding method can be applied to get the feature values. The summary of this method of feature extraction can be presented as follows.

Algorithm 2.4: New hybrid method of feature extraction (combination of process of dilation and simple bar mask encoder)

1. Receive the input document image.

2. Threshold the input image into two tone image. Now only '1's or '0's are present in the image.
3. Dilate the original image into \( x \) number of directions with proper structuring elements to get \( x \) modified images (in the experiments it is 2).

4. Make \( \text{counter} = 1 \).

5. Divide the \( \text{counter}^{th} \) image into \( h \) horizontal regions (in the experiments \( h = 5 \)).

6. Divide the \( \text{counter}^{th} \) image into \( v \) vertical regions (in the experiments \( v = 5 \)). These regions may overlap with the above horizontal regions.

7. For first region out of the above \( h + v \) regions, count the number of 'ON' ('1's) pixels. Count the total number of pixels in the same region. Find feature 1 = number of '1's ('ON' pixels) / number of total pixels in that region.

8. Repeat step 7 for all the regions and determine feature 2, feature 3 ... feature \((h + v)\).

9. Make \( \text{counter} = \text{counter} + 1 \).

10. If \( \text{counter} \) is less than or equal to \( x \) go to step 5 else end.

### 2.5 Experiments and results

Three experiments are conducted to extract the features of 15, 64 \( \times \) 64 pixels document images by the above three methods. (i) Simple bar mask encoder method, (ii) moment invariant feature extraction method and (iii) a new hybrid method (To store these document images cell array data structure is utilized. Each location in this array is capable of storing a complete document image (matrix) of any size. The more details about this data structure can be found in [36]). Of these 15 images, 5 are English document images, 5 are Hindi document images and 5 are Kannada document images.

The results obtained by simple bar mask encoder method are shown in table 2.1. Here each image is divided into five horizontal bars and five vertical bars creating ten capture regions. Hence for each image simple bar mask encoder method (algorithm 2.1) produces 10 feature values. For each of 15 images such 10 features values \((f1 \text{ to } f10)\) are shown in table 2.1.
Table 2.1: Results obtained by a simple bar mask encoder method of feature extraction on 15, 64 x 64 pixels document images.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>f7</th>
<th>f8</th>
<th>f9</th>
<th>f10</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0.2083</td>
<td>0.0794</td>
<td>0.1419</td>
<td>0.2031</td>
<td>0.1592</td>
<td>0.1966</td>
<td>0.1497</td>
<td>0.1484</td>
<td>0.1849</td>
<td>0.1240</td>
</tr>
<tr>
<td>02</td>
<td>0.1901</td>
<td>0.0755</td>
<td>0.1367</td>
<td>0.1979</td>
<td>0.1504</td>
<td>0.1211</td>
<td>0.1667</td>
<td>0.1771</td>
<td>0.1549</td>
<td>0.1357</td>
</tr>
<tr>
<td>03</td>
<td>0.1732</td>
<td>0.0885</td>
<td>0.1250</td>
<td>0.2266</td>
<td>0.1445</td>
<td>0.1576</td>
<td>0.1289</td>
<td>0.1628</td>
<td>0.1432</td>
<td>0.1602</td>
</tr>
<tr>
<td>04</td>
<td>0.1927</td>
<td>0.1003</td>
<td>0.1445</td>
<td>0.2031</td>
<td>0.1523</td>
<td>0.1497</td>
<td>0.1758</td>
<td>0.1615</td>
<td>0.1393</td>
<td>0.1631</td>
</tr>
<tr>
<td>05</td>
<td>0.2122</td>
<td>0.0781</td>
<td>0.1406</td>
<td>0.1719</td>
<td>0.1309</td>
<td>0.1315</td>
<td>0.1419</td>
<td>0.1523</td>
<td>0.1523</td>
<td>0.1494</td>
</tr>
<tr>
<td>06</td>
<td>0.2318</td>
<td>0.1510</td>
<td>0.1146</td>
<td>0.2747</td>
<td>0.1943</td>
<td>0.2161</td>
<td>0.1810</td>
<td>0.2018</td>
<td>0.1784</td>
<td>0.1904</td>
</tr>
<tr>
<td>07</td>
<td>0.1836</td>
<td>0.2266</td>
<td>0.2044</td>
<td>0.1589</td>
<td>0.1924</td>
<td>0.2448</td>
<td>0.1250</td>
<td>0.1927</td>
<td>0.2005</td>
<td>0.2002</td>
</tr>
<tr>
<td>08</td>
<td>0.2122</td>
<td>0.1549</td>
<td>0.1875</td>
<td>0.2487</td>
<td>0.1787</td>
<td>0.1680</td>
<td>0.1680</td>
<td>0.1953</td>
<td>0.2161</td>
<td>0.2207</td>
</tr>
<tr>
<td>09</td>
<td>0.2344</td>
<td>0.2044</td>
<td>0.1758</td>
<td>0.1003</td>
<td>0.1533</td>
<td>0.1016</td>
<td>0.2214</td>
<td>0.1849</td>
<td>0.1354</td>
<td>0.2070</td>
</tr>
<tr>
<td>10</td>
<td>0.2409</td>
<td>0.1367</td>
<td>0.1406</td>
<td>0.2865</td>
<td>0.1934</td>
<td>0.2109</td>
<td>0.1927</td>
<td>0.1940</td>
<td>0.1836</td>
<td>0.2109</td>
</tr>
<tr>
<td>11</td>
<td>0.1680</td>
<td>0.1576</td>
<td>0.1172</td>
<td>0.1497</td>
<td>0.1211</td>
<td>0.0052</td>
<td>0.1758</td>
<td>0.2135</td>
<td>0.1380</td>
<td>0.1660</td>
</tr>
<tr>
<td>12</td>
<td>0.2070</td>
<td>0.1784</td>
<td>0.1549</td>
<td>0.1628</td>
<td>0.1523</td>
<td>0.1810</td>
<td>0.2031</td>
<td>0.1536</td>
<td>0.1680</td>
<td>0.1504</td>
</tr>
<tr>
<td>13</td>
<td>0.1862</td>
<td>0.1732</td>
<td>0.1576</td>
<td>0.1563</td>
<td>0.1963</td>
<td>0.1576</td>
<td>0.2227</td>
<td>0.1901</td>
<td>0.1419</td>
<td>0.1670</td>
</tr>
<tr>
<td>14</td>
<td>0.2057</td>
<td>0.1536</td>
<td>0.1458</td>
<td>0.1602</td>
<td>0.1572</td>
<td>0.1940</td>
<td>0.1602</td>
<td>0.1667</td>
<td>0.1250</td>
<td>0.1719</td>
</tr>
<tr>
<td>15</td>
<td>0.2057</td>
<td>0.1771</td>
<td>0.1693</td>
<td>0.1432</td>
<td>0.1484</td>
<td>0.1471</td>
<td>0.1641</td>
<td>0.1576</td>
<td>0.1797</td>
<td>0.1836</td>
</tr>
</tbody>
</table>

The results obtained from the same document images by the moment invariant method of feature extraction (algorithm 2.2) are shown in table 2.2. For each sample image seven features (f1 to f7) corresponding to seven invariant moments for each sample are shown.

In the experimentation of new hybrid method (algorithm 2.3) the following two structuring elements SE1 and SE2 (hence x = 2) are used. SE1 captures the horizontal information from the original image and SE2 captures the vertical information from the original image.

\[ SE1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \]
\[ SE2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \]

The two new modified images are of the same size as that of the original image (64 x 64 pixels). For each such modified image bar mask which extracts 10 features (horizontal bars, h = 5, vertical bars, v = 5) is applied. Thus there are totally 20 features for each sample image. These 20 features obtained for the 15 sample document images are shown in tables 2.3a, 2.3b and 2.3c.
Table 2.2: Results obtained by moment invariant feature extraction method on 15, 64 x 64 pixels document images.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>f7</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1.1690</td>
<td>0.0015</td>
<td>0.0031</td>
<td>0.0023</td>
<td>0.0000</td>
<td>-0.0001</td>
<td>-0.0000</td>
</tr>
<tr>
<td>02</td>
<td>1.1306</td>
<td>0.0095</td>
<td>0.0025</td>
<td>0.0018</td>
<td>-0.0000</td>
<td>-0.0002</td>
<td>-0.0000</td>
</tr>
<tr>
<td>03</td>
<td>1.2094</td>
<td>0.0091</td>
<td>0.0114</td>
<td>0.0016</td>
<td>0.0000</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>04</td>
<td>1.1773</td>
<td>0.0004</td>
<td>0.0048</td>
<td>0.0011</td>
<td>0.0000</td>
<td>-0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>05</td>
<td>1.2135</td>
<td>0.0056</td>
<td>0.0315</td>
<td>0.0075</td>
<td>-0.0001</td>
<td>-0.0002</td>
<td>-0.0001</td>
</tr>
<tr>
<td>06</td>
<td>1.1354</td>
<td>0.0279</td>
<td>0.0070</td>
<td>0.0035</td>
<td>-0.0000</td>
<td>0.0004</td>
<td>-0.0000</td>
</tr>
<tr>
<td>07</td>
<td>1.1042</td>
<td>0.0042</td>
<td>0.0098</td>
<td>0.0013</td>
<td>0.0000</td>
<td>0.0001</td>
<td>-0.0000</td>
</tr>
<tr>
<td>08</td>
<td>1.1044</td>
<td>0.0053</td>
<td>0.0068</td>
<td>0.0037</td>
<td>-0.0000</td>
<td>-0.0002</td>
<td>-0.0000</td>
</tr>
<tr>
<td>09</td>
<td>1.2264</td>
<td>0.0093</td>
<td>0.0074</td>
<td>0.0008</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>1.1313</td>
<td>0.0050</td>
<td>0.0008</td>
<td>0.0080</td>
<td>-0.0000</td>
<td>0.0005</td>
<td>0.0000</td>
</tr>
<tr>
<td>11</td>
<td>0.9801</td>
<td>0.0034</td>
<td>0.0015</td>
<td>0.0022</td>
<td>-0.0000</td>
<td>-0.0001</td>
<td>-0.0000</td>
</tr>
<tr>
<td>12</td>
<td>1.0305</td>
<td>0.0012</td>
<td>0.0012</td>
<td>0.0004</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>13</td>
<td>0.8727</td>
<td>0.0095</td>
<td>0.0082</td>
<td>0.0005</td>
<td>-0.0000</td>
<td>-0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>14</td>
<td>1.0916</td>
<td>0.0007</td>
<td>0.0086</td>
<td>0.0034</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.0000</td>
</tr>
<tr>
<td>15</td>
<td>0.9070</td>
<td>0.0020</td>
<td>0.0025</td>
<td>0.0003</td>
<td>0.0000</td>
<td>-0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 2.3a: Results obtained by the developed method, features 1 to 7.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>f7</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0.0378</td>
<td>0.0091</td>
<td>0.0195</td>
<td>0.0313</td>
<td>0.0264</td>
<td>0.0326</td>
<td>0.0234</td>
</tr>
<tr>
<td>02</td>
<td>0.0391</td>
<td>0.0143</td>
<td>0.0286</td>
<td>0.0339</td>
<td>0.0205</td>
<td>0.0195</td>
<td>0.0286</td>
</tr>
<tr>
<td>03</td>
<td>0.0208</td>
<td>0.0143</td>
<td>0.0169</td>
<td>0.0339</td>
<td>0.0234</td>
<td>0.0143</td>
<td>0.0208</td>
</tr>
<tr>
<td>04</td>
<td>0.0365</td>
<td>0.0156</td>
<td>0.0247</td>
<td>0.0391</td>
<td>0.0293</td>
<td>0.0247</td>
<td>0.0182</td>
</tr>
<tr>
<td>05</td>
<td>0.0417</td>
<td>0.0156</td>
<td>0.0260</td>
<td>0.0286</td>
<td>0.0254</td>
<td>0.0221</td>
<td>0.0273</td>
</tr>
<tr>
<td>06</td>
<td>0.1016</td>
<td>0.0716</td>
<td>0.0313</td>
<td>0.1211</td>
<td>0.0840</td>
<td>0.0951</td>
<td>0.0716</td>
</tr>
<tr>
<td>07</td>
<td>0.0430</td>
<td>0.0820</td>
<td>0.0820</td>
<td>0.0768</td>
<td>0.0742</td>
<td>0.0951</td>
<td>0.0482</td>
</tr>
<tr>
<td>08</td>
<td>0.0951</td>
<td>0.0885</td>
<td>0.0417</td>
<td>0.1029</td>
<td>0.0693</td>
<td>0.0560</td>
<td>0.0703</td>
</tr>
<tr>
<td>09</td>
<td>0.1003</td>
<td>0.0703</td>
<td>0.0703</td>
<td>0.0195</td>
<td>0.0576</td>
<td>0.0443</td>
<td>0.0768</td>
</tr>
<tr>
<td>10</td>
<td>0.1094</td>
<td>0.0951</td>
<td>0.0221</td>
<td>0.1133</td>
<td>0.0850</td>
<td>0.0990</td>
<td>0.0820</td>
</tr>
<tr>
<td>11</td>
<td>0.0313</td>
<td>0.0443</td>
<td>0.0378</td>
<td>0.0234</td>
<td>0.0186</td>
<td>0.0</td>
<td>0.0391</td>
</tr>
<tr>
<td>12</td>
<td>0.0378</td>
<td>0.0443</td>
<td>0.0547</td>
<td>0.0221</td>
<td>0.0303</td>
<td>0.0469</td>
<td>0.0391</td>
</tr>
<tr>
<td>13</td>
<td>0.0313</td>
<td>0.0339</td>
<td>0.0547</td>
<td>0.0221</td>
<td>0.0498</td>
<td>0.0260</td>
<td>0.0586</td>
</tr>
<tr>
<td>14</td>
<td>0.0365</td>
<td>0.0313</td>
<td>0.0443</td>
<td>0.0182</td>
<td>0.0322</td>
<td>0.0456</td>
<td>0.0365</td>
</tr>
<tr>
<td>15</td>
<td>0.0469</td>
<td>0.0443</td>
<td>0.0521</td>
<td>0.0117</td>
<td>0.0332</td>
<td>0.0313</td>
<td>0.0313</td>
</tr>
</tbody>
</table>
Table 2.3b (cont. of table 2.3a): Results obtained by the developed method, features 8 to 14.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>$f_8$</th>
<th>$f_9$</th>
<th>$f_{10}$</th>
<th>$f_{11}$</th>
<th>$f_{12}$</th>
<th>$f_{13}$</th>
<th>$f_{14}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0.0195</td>
<td>0.0299</td>
<td>0.0205</td>
<td>0.0781</td>
<td>0.0169</td>
<td>0.0703</td>
<td>0.0859</td>
</tr>
<tr>
<td>02</td>
<td>0.0313</td>
<td>0.0313</td>
<td>0.0244</td>
<td>0.0586</td>
<td>0.0065</td>
<td>0.0456</td>
<td>0.0742</td>
</tr>
<tr>
<td>03</td>
<td>0.0273</td>
<td>0.0247</td>
<td>0.0225</td>
<td>0.0586</td>
<td>0.0208</td>
<td>0.0651</td>
<td>0.0833</td>
</tr>
<tr>
<td>04</td>
<td>0.0299</td>
<td>0.0365</td>
<td>0.0342</td>
<td>0.0716</td>
<td>0.0195</td>
<td>0.0664</td>
<td>0.0820</td>
</tr>
<tr>
<td>05</td>
<td>0.0182</td>
<td>0.0365</td>
<td>0.0313</td>
<td>0.0729</td>
<td>0.0065</td>
<td>0.0443</td>
<td>0.0547</td>
</tr>
<tr>
<td>06</td>
<td>0.0833</td>
<td>0.0755</td>
<td>0.0840</td>
<td>0.0690</td>
<td>0.0221</td>
<td>0.0234</td>
<td>0.0404</td>
</tr>
<tr>
<td>07</td>
<td>0.0638</td>
<td>0.0794</td>
<td>0.0723</td>
<td>0.0547</td>
<td>0.0534</td>
<td>0.0521</td>
<td>0.0260</td>
</tr>
<tr>
<td>08</td>
<td>0.0846</td>
<td>0.0833</td>
<td>0.0947</td>
<td>0.0560</td>
<td>0.0208</td>
<td>0.0729</td>
<td>0.0586</td>
</tr>
<tr>
<td>09</td>
<td>0.0703</td>
<td>0.0521</td>
<td>0.0703</td>
<td>0.0742</td>
<td>0.0638</td>
<td>0.0378</td>
<td>0.0365</td>
</tr>
<tr>
<td>10</td>
<td>0.0924</td>
<td>0.0651</td>
<td>0.0859</td>
<td>0.0573</td>
<td>0.0208</td>
<td>0.0391</td>
<td>0.0794</td>
</tr>
<tr>
<td>11</td>
<td>0.0417</td>
<td>0.0352</td>
<td>0.0342</td>
<td>0.0247</td>
<td>0.0443</td>
<td>0.0026</td>
<td>0.0430</td>
</tr>
<tr>
<td>12</td>
<td>0.0221</td>
<td>0.0521</td>
<td>0.0293</td>
<td>0.0326</td>
<td>0.0586</td>
<td>0.0065</td>
<td>0.0417</td>
</tr>
<tr>
<td>13</td>
<td>0.0326</td>
<td>0.0352</td>
<td>0.0420</td>
<td>0.0339</td>
<td>0.0534</td>
<td>0.0117</td>
<td>0.0378</td>
</tr>
<tr>
<td>14</td>
<td>0.0247</td>
<td>0.0221</td>
<td>0.0332</td>
<td>0.0417</td>
<td>0.0313</td>
<td>0.0065</td>
<td>0.0365</td>
</tr>
<tr>
<td>15</td>
<td>0.0326</td>
<td>0.0456</td>
<td>0.0439</td>
<td>0.0378</td>
<td>0.0352</td>
<td>0.0169</td>
<td>0.0260</td>
</tr>
</tbody>
</table>

Table 2.3c (cont. of table 2.3b): Results obtained by the developed method, features 15 to features 20.

<table>
<thead>
<tr>
<th>Serial No.</th>
<th>$f_{15}$</th>
<th>$f_{16}$</th>
<th>$f_{17}$</th>
<th>$f_{18}$</th>
<th>$f_{19}$</th>
<th>$f_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>0.0371</td>
<td>0.0742</td>
<td>0.0534</td>
<td>0.0521</td>
<td>0.0599</td>
<td>0.0459</td>
</tr>
<tr>
<td>02</td>
<td>0.0557</td>
<td>0.0482</td>
<td>0.0599</td>
<td>0.0508</td>
<td>0.0378</td>
<td>0.0469</td>
</tr>
<tr>
<td>03</td>
<td>0.0498</td>
<td>0.0768</td>
<td>0.0443</td>
<td>0.0482</td>
<td>0.0443</td>
<td>0.0605</td>
</tr>
<tr>
<td>04</td>
<td>0.0439</td>
<td>0.0508</td>
<td>0.0781</td>
<td>0.0677</td>
<td>0.0430</td>
<td>0.0439</td>
</tr>
<tr>
<td>05</td>
<td>0.0400</td>
<td>0.0456</td>
<td>0.0508</td>
<td>0.0664</td>
<td>0.0313</td>
<td>0.0283</td>
</tr>
<tr>
<td>06</td>
<td>0.0371</td>
<td>0.0560</td>
<td>0.0313</td>
<td>0.0378</td>
<td>0.0404</td>
<td>0.0293</td>
</tr>
<tr>
<td>07</td>
<td>0.0469</td>
<td>0.0508</td>
<td>0.0391</td>
<td>0.0443</td>
<td>0.0391</td>
<td>0.0566</td>
</tr>
<tr>
<td>08</td>
<td>0.0332</td>
<td>0.0365</td>
<td>0.0391</td>
<td>0.0664</td>
<td>0.0495</td>
<td>0.0459</td>
</tr>
<tr>
<td>09</td>
<td>0.0293</td>
<td>0.0195</td>
<td>0.0794</td>
<td>0.0482</td>
<td>0.0482</td>
<td>0.0420</td>
</tr>
<tr>
<td>10</td>
<td>0.0586</td>
<td>0.0443</td>
<td>0.0599</td>
<td>0.0482</td>
<td>0.0378</td>
<td>0.0635</td>
</tr>
<tr>
<td>11</td>
<td>0.0254</td>
<td>0.0013</td>
<td>0.0313</td>
<td>0.0417</td>
<td>0.0326</td>
<td>0.0313</td>
</tr>
<tr>
<td>12</td>
<td>0.0244</td>
<td>0.0352</td>
<td>0.0391</td>
<td>0.0286</td>
<td>0.0378</td>
<td>0.0234</td>
</tr>
<tr>
<td>13</td>
<td>0.0381</td>
<td>0.0365</td>
<td>0.0534</td>
<td>0.0404</td>
<td>0.0221</td>
<td>0.0264</td>
</tr>
<tr>
<td>14</td>
<td>0.0273</td>
<td>0.0299</td>
<td>0.0286</td>
<td>0.0313</td>
<td>0.0221</td>
<td>0.0303</td>
</tr>
<tr>
<td>15</td>
<td>0.0371</td>
<td>0.0339</td>
<td>0.0286</td>
<td>0.0326</td>
<td>0.0299</td>
<td>0.0303</td>
</tr>
</tbody>
</table>
2.6 Comparisons and discussions

The three methods of feature extraction discussed above can be compared with the help of the results shown in tables 2.1, 2.2 and 2.3. Any feature is said to be good if it exhibits the changes for different classes and remains consistent in the same group or same class. An ideal feature should have discriminative ability among the classes.

To analyze the capability of the extracted features, the simplest way is to plot the values obtained. Thus the obtained real values of the features are plotted on the $y$ axis and image serial numbers are indicated on the $x$ axis. Figure 2.4 shows the plot of features extracted by the simple bar mask encoder method. In this graph, even though there is a distinction between the first 5 images (class 1 images) and the second 5 images (class 2 images), there exists only slight distinction between the second 5 and the third 5 images (class 3). Thus there is a chance of confusion between class 2 and class 3 images.

The above fact can be visually seen in figure 2.4.

![Figure 2.4: Plot of features extracted by a simple bar mask encoder method.](image)
Similarly figure 2.5 shows the plot of features extracted by the moment invariant feature extraction method. In this graph there is no distinction between the classes and the method has not produced any interesting features. This can be seen in figure 2.5.

![Figure 2.5: Plot of features extracted by moment invariant method.](image)

The features extracted from the developed hybrid method are plotted in figure 2.6. It can be observed in the graph that compared to above two methods (figure 2.4 and figure 2.5) this method yields more discriminative features. The values of features obtained for images 1 to 5 (class 1), 6 to 10 (class 2), and 11 to 15 (class 3) are visually distinct and are most useful in classification.
2.7 Conclusions

In this chapter document image pre-processor, that is feature extraction methodologies are discussed. Existing methods which are more often used with neural networks are explained and implemented. The newly developed, hybrid method of simple bar mask encoder and morphological operators considered in various directions, yields very effective features. The results obtained by this method are compared with the results obtained by the simple bar mask encoder method and the moment invariant method and are found to have more discriminative ability.