Barcode Recognition Using Principal Component Analysis and Support Vector Machine

Clarin Mulyaningtyas  
Department of Mathematics  
Universitas Negeri Surabaya  
Surabaya, Indonesia

Elly Matul Imah  
Department of Mathematics  
Universitas Negeri Surabaya  
Surabaya, Indonesia  
ellymatul@unesa.ac.id*

Abstract—Barcode is visual code to identify the symbols of the data in the form of one or two-dimension image contains lines and spaces based on detecting the edges. The use of barcode has significantly contributed for warehouses and retail product. Nowadays, the research about barcode is still an interesting topic especially from blurry, low contrast, low resolution, rotated barcode and fixed-focus lenses. Datasets of barcode are taken from WWU Muenster Barcode Database University of Muenster Germany as many as 142 images consisting 13 types of barcode EAN-13. This research aims to investigate the possibilities of one-dimensional barcode recognition in image region using Support Vector Machine (SVM) multiclass one-against-all with feature extraction using Principal Component Analysis (PCA) variation of principal component are 8, 12, 17, 25, 38, and 70 features. Dataset were randomly separated into data train and data test using cross validation repeated five times with ratio 2:1 of 95 images data train and 47 images data test. Based on the best performance result, SVM was capable for classifying barcode accurately with accuracy $0.92 \pm 0.02$. Based on computation time, the average of training time is about 3.21 seconds and testing time is about 0.66 seconds.

Keywords— Barcode, Principal Component Analysis (PCA), Support Vector Machine (SVM)

I. INTRODUCTION

Barcode is visual representation of information in the form of bars and spaces. Bars and spaces are designed with different thickness which representing numbers, letters, and symbols. Barcode are utilised by packaging products to store production code, identity number, and company license [1]. Nowadays, barcode recognition technology is not only for labeling an item but also developed in many application such as identification [2], tracking [3], inventory of goods [4] and others. Some commonly barcode used are Code 128, Code 39, UPC, EAN etc [1]

Basically, the method used to read barcode is scanline that works by scanning barcode lines using utilising laser light [2]. Firstly, the laser beam gets the barcode information then passed on to a computer with a simpler data format to detect the barcode symbols. The use of laser light aims to barcode with low intensity or blurry can be read properly by the process of converting image to a binary value. There are many methods and algorithms that can be used to read barcode such as research by Bathia et al that used template matching method. Template matching is a technique in digital image processing to find small parts of the image that match with the image template [4]. Tribak et al used the PCA method to recognize QR-Code. PCA developed a method that extracts a feature vector which important to recognise the code. PCA analysed the feature of each QR-code characteristics component to distinguish identical image. The method of distinguishing is measured using the entire QR-Code position and angle to identification [3]. Noce et al used mobile devices with Hough transform algorithms to detect barcode areas with average detection time of 270 seconds [10]

Based on the research that has been done by Asraf et al [7], the classification problem was a problem that often encountered in daily life, which is to determine an object whether a type of particular object or not. Based on the problem, this research aims to build a barcode classification method using Support Vector Machine (SVM) with Principal Component Analysis (PCA) feature extraction. PCA is chosen for feature extraction because it is a simple feature extraction algorithm. PCA is the method of unsupervised learning which theoretically can take data information maximally with small dimension results. PCA developed a method that extracts a feature vector which important to recognise the code. PCA is a dimensional reduction technique for simplifying data by transforming data linearly to form small dimensionless data that does not correlate each other without losing the data characteristics. PCA works by finding the principal component value by calculating eigenvalues and eigenvectors of the covariance matrix by performing Eigen Value Decomposition (EVD) [3]. The general PCA method is known as the Karhunen-Loven method, which attempts projection from the high dimensional image space to the characteristic space with lower dimensions. Dimensional reduction aims to improve efficiency in the computing process and reduce the required memory.

SVM is preferred for barcode classification because SVM is still become state of the art in machine learning. SVM is known as the best machine learning technique after previous machine learning theory called Neural Network (NN) [7]. SVM and NN training process used the pair of input data and
output as the target. Many researchers use SVM in health [5], signals [6], disease in plants [7], and others. SVM works by finding the best hyperplane by measuring the distance of maximum margin from a data point. The best hyperplane is hyperplane located in the middle of two classes. Process of finding best hyperplane is the basic process in SVM. While the testing process using model from training [9].

Barcode is chosen for experimental object because research about barcode recognition using SVM is rarely to find. This paper is organised as follows: section 1 consists of background and literature review related to SVM and PCA, section 2 is about research methods and describe some concepts of PCA and SVM, section 3 describes the flowchart of experiment, dataset used and cross validation. The experimental results are given in section 4 and section 5 presents the conclusion of the research.

II. METHOD

A. Preprocessing

Preprocessing is the process of converting image into numeric because the brightness and color of the image distinguishes image type that represented by numeric [8]. The purpose of preprocessing is to shorten the process of data. Preprocessing is shown in Fig.1 which consist of cropping, grayscale, sobel edge detection, and convert to one dimensional vector. Here is the block diagram of preprocessing.

Fig 1. Block Diagram of Preprocessing

The first step of preprocessing was cropping the barcode image background. The cropping process aims to eliminate the unnecessary parts of images such as package background, production dates, presentation methods and more. Fig 2 shows the cropping results.

Fig 2. Cropped Image

The next process was grayscale. Grayscale conversion is a technique for converting colored image having the matrix values of each R, G and B (Red, Green, Blue) into grayscale images. The conversion can be done by taking the mean of R, G and B [8].

\[(i, j) = \frac{R(i, j) + G(i, j) + B(i, j)}{3}\]  

Sobel edge detection is finding and marking the small part of the image border details and also for fixing the blurry image.

Process of sobel edge detection among the pixel \((x, y)\) is formulated as:

\[G = \begin{bmatrix} a_0 & a_1 & a_2 \\ a_3 & a_4 & a_5 \\ a_6 & a_7 & a_8 \end{bmatrix}\]

The last process was transform image into one-dimensional form. A barcode image in the form of a two-dimensional image can be seen as a one-dimensional vector if the width of the image is \(h\) and the length of the image is \(w\), then the process of converting the image into a 1-dimensional vector form is \((h \times w)\).

B. Feature Extraction

PCA is an image dimensional reduction technique that converts a large number of variables that correlate into a small number of uncorrelated variables without losing the information of the data [3]. The first step of PCA was finding the mean value. Mean is the process of finding the average row by summing the entire value of each line then divided by all of data. The mean can be found by using this equation.

\[
\bar{\mu} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}\]

\(x_i\) is the input data and \(n\) is the number of data. The next process was data normalization. Data normalization aims to minimize redundancy of data and time complexity. The next process was finding the covariance matrix. The covariance matrix equation is given by:

\[
cov(X_i, Y_i) = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \mu_i)(Y_i - \mu_i)^T\]

\(T\) is transpose mean matrix. The dimension of covariance matrix is very large hence generating the eigenvectors so to reduce the calculation, it needs to calculate the eigenvectors and eigenvalue. Total number of column in covariance matrix gives the total number of eigenvectors. Then the corresponding eigenvector \(\lambda\) can be searched as follows.

\[
(C - \lambda I) \cdot V = 0\]

The vector eigenv is sorted by the eigenvalues from the largest to the smallest which is called the characteristic matrix. So the principal component value are selected by k% of the total eigenvalues. Furthermore, the result of the matrix is multiplied by the previous normalization matrix [3].

C. Support Vector Machine (SVM)

SVM is used to separate data linearly 2 classes. Suppose \(\{x_1, \ldots, x_n\}\) is dataset with \(y_i \in \{+1, -1\}\) is class label. SVM will find the best hyperplane in input space to separate the classes [11]. Data is separated by hyperplane located in the middle of classes. The best hyperplane is a hyperplane located halfway between two sets of objects from two classes determined by measuring the maximum margin. The margin value determines the distance between classes. While Support vector is the closest point to the hyperplane in each class [12].

Data form class +1 and -1 are formulated as data that satisfy the equations:

\[(w \cdot x_i) + b \geq +1 \text{ for } y_i = +1\]

\[(w \cdot x_i) + b \geq -1 \text{ for } y_i = -1\]
w is the weight and b is bias. Since there are two inputs (x₁ and x₂), then w also has two attributes (w₁ and w₂) that is to minimize value [13].

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}(w_1^2 + w_2^2)$$  \hfill (8)

for \(y_i = (w, x_1 + b) \geq 1, i = 1, 2, 3, ..., N\)

The optimization problem was solved by using lagrange multiplier and satisfies the condition of Karush-Kuhn-Tucker (Summit) [14]. The main idea of lagrangian is to minimize an estimate of number of positive multiplier which improves bounds on the generalization error. This points are termed as Support Vector. The hyperplane is determined by separated training set. SVM can be used to summarise the information contained in dataset by the the SV [15]. After completing the optimization problem, then the class of the testing data can be determined based on the value of the decision function [16].

$$f(x_d) = \sum_{i=1}^{m} \alpha_i y_i x_i x_d + b$$  \hfill (9)

m is total number of support vector, \(x_i \) is support vector, and \(x_d\) is classified data. Hence can be inferred that support vectors are \(\alpha_i\) with positive value. SVM do not require a reduction in the number of features in order to avoid over fitting data classification. Another advantage of SVM is the low probability of generalization errors [17].

### III. EXPERIMENTAL SET UP

This research consists of some processes; preprocessing for converting images into numeric for shorten the learning time. Cross validation was used to separate data into two parts: training data and testing data. PCA feature extraction obtains the data characteristics by reduce the dimensionality [18]. The last process was classification using SVM. For the testing process, data input will be processed using model from training process. Fig.3 is the flowchart of the barcode recognise implementation.

![Flowchart of Experiment](Image)

#### TABLE I. INPUT IMAGE

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of Data</th>
<th>Sample Barcode Image</th>
<th>Resolution (pixels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>6</td>
<td>24</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>7</td>
<td>24</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
<tr>
<td>13</td>
<td>8</td>
<td><img src="Image" alt="Image" /></td>
<td>20x10</td>
</tr>
</tbody>
</table>

### A. Dataset

Dataset were taken from WWU Muenster Barcode Database University of Muenster Germany. These images are in color and converted to gray levels. The edge detection following the sobel edge detection algorithm to keep only the outer border having less than fixed value of pixels. The dataset used consists of 142 RGB images with thirteen class categories, size 648x488 and resolution 72 pixels [19].

Table 1 shows the dataset of thirteen subject of barcode which every class were taken 6-24 images total 142 images. Some of barcode subject have similar appearance but they were taken in different lighting, blurry, rotation and angle. Data input that used for this study resized into 20x10 pixels. Dataset was separated into data training and data testing using cross validation. The dimensionality reduction using PCA feature extraction, and the last process is SVM will be classify barcode into thirteen class based on thirteen different subject. Due to imbalance samples, the random selection is carried out in each barcode.

### B. Cross Validation

Cross validation type hold out is the process of dividing the dataset into two parts: training data and testing data [20]. This
study randomly select ratio 2:1 and 4:1 images to build datasets. The illustration of cross validation hold out is shown in Fig.4.

Table II shows the best average accuracy with feature extraction was from 70 features with accuracy 0.92 ± 0.02. While the best average accuracy without feature extraction or full features was 0.93 ± 0.03. The result of experiment based on computation time is presented by Fig. 5.

V. CONCLUSION

In this study, the classification was used for barcode recognition using SVM with PCA feature extraction. Experimental results show that PCA algorithm can be used for barcode feature extraction by calculating mean, covariance, eigenvalues, eigenvectors, and principal component values. The barcode recognize using SVM was done by inputting several parameters of μ (parameters for quadratic programming (QP)), and C (maximum penalty limit for the lagrange multiplier duality process). The result of SVM gives the best average accuracy of linear kernel is 0.92 ± 0.02 with an average time about 0.72 seconds. While the average accuracy without feature extraction with is 0.93 ± 0.03 with an average time about of 2.26 seconds.

REFERENCES


