Research Article

A Novel Two-Stage Spectrum-Based Approach for Dimensionality Reduction: A Case Study on the Recognition of Handwritten Numerals

Mohammad Amin Shayegan,1 Saeed Aghabozorgi,2 and Ram Gopal Raj3

1 Image Processing and Pattern Recognition Research Lab, R&D Center, Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
2 Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
3 Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Correspondence should be addressed to Mohammad Amin Shayegan; mashaygan@siswa.um.edu.my

Received 29 January 2014; Accepted 11 April 2014; Published 12 May 2014

Academic Editor: Feng Gao

Copyright © 2014 Mohammad Amin Shayegan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Dimensionality reduction (feature selection) is an important step in pattern recognition systems. Although there are different conventional approaches for feature selection, such as Principal Component Analysis, Random Projection, and Linear Discriminant Analysis, selecting optimal, effective, and robust features is usually a difficult task. In this paper, a new two-stage approach for dimensionality reduction is proposed. This method is based on one-dimensional and two-dimensional spectrum diagrams of standard deviation and minimum to maximum distributions for initial feature vector elements. The proposed algorithm is validated in an OCR application, by using two big standard benchmark handwritten OCR datasets, MNIST and Hoda. In the beginning, a 133-element feature vector was selected from the most used features, proposed in the literature. Finally, the size of initial feature vector was reduced from 100% to 59.40% (79 elements) for the MNIST dataset, and to 43.61% (58 elements) for the Hoda dataset, in order. Meanwhile, the accuracies of OCR systems are enhanced 2.95% for the MNIST dataset, and 4.71% for the Hoda dataset. The achieved results show an improvement in the precision of the system in comparison to the rival approaches, Principal Component Analysis and Random Projection. The proposed technique can also be useful for generating decision rules in a pattern recognition system using rule-based classifiers.

1. Introduction

Pattern recognition (PR) is one of the most attractive branches in the artificial intelligence field. In all PR systems, the quantity, quality, and diversity of training data in the learning process directly affect the final result. The training dataset size has double importance in this regard, because the training phase—related to large datasets—is often a time-consuming process.

Nowadays, the emergence of the big-data issue has resulted in specific attention on size reduction and also dimensionality reduction to save time and memory usage. In addition, the demand for employing various applications running on limited-speed and limited-memory devices, such as mobile phones and mobile scanners, is growing dramatically [1]. Hence, the necessity to find efficient techniques for reducing the volume of data in order to decrease the overall processing time and also the memory requirements is considered more important than in the past.

Two general approaches are utilized for the dataset volume reduction in the literature: size reduction and dimensionality reduction.
The available size reduction techniques try to reduce the number of objects or observations in a dataset. They find and remove two groups of samples from a dataset: samples far from a class centroid (outlier samples or support vector samples) [2–5] and samples close to each class centroid (e.g., using k-means clustering technique) [6]. However, the outlier and support vector samples are necessary to evaluate system efficiency and system functionality. In addition, the samples close to a class centroid include important information about the various characteristics of a class and are necessary to make a system model.

The dimensionality reduction techniques have been used to identify and remove the less important features, extracted from the dataset samples. They are widely employed in different areas, such as biological data clustering [7], EMG signal feature reduction [8], face recognition [9], blog visualization reduction [10], and gene expression databases reduction [11]. Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Random Projection (RP) are some of the well-known methods in this part [12]. Concerning the dimensionality reduction in the Optical Character Recognition (OCR) applications, PCA has been used to compress the features space in the numeral part of the CEDAR database [13], MNIST database [14], and also the Tamil handwritten character classes [15]. Nevertheless, finding an optimal, effective, and robust feature set out of a large initial extracted feature is usually a heuristic and hard task [16].

This paper contributes to the corpus of knowledge in dimensionality reduction as follows: (1) introducing a new preprocessing method to stick disconnected parts of an image and (2) proposing a novel two-stage spectrum-based method to reduce the number of features in the feature space. The mentioned method is based on analysing one-dimensional and two-dimensional spectrum diagrams for standard deviation and minimum to maximum distribution corresponding to the existing features in the initial feature vector. Unlike the other available techniques for dimensionality reduction, such as PCA, the proposed method can keep every feature in the final feature vector, based on some characteristics of a specific feature or even based on user opinion. To investigate the efficiency of the proposed dimensionality reduction technique, two large standard benchmark handwritten OCR datasets—MNIST and Hoda—are employed. The empirical results have shown the effectiveness of the proposed method. Although the results have been reported for OCR application databases, the salient point of the proposed approach is that it can also be used for other datasets with numerical feature vectors.

The rest of the paper is organized as follows. Section 2 discusses the background of the research topic and introduces the related works for dimensionality reduction operation. Thereafter, Section 3 presents the proposed dimensionality reduction technique, including the definition of one- and two-dimensional standard deviations and also minimum to maximum spectrum diagrams. Section 4 presents the research method and the experimental results and comparisons, and, finally, Section 5 presents the conclusion of the paper.

2. Background and Related Works
2.1. Feature Extraction in OCR Application. Among the various stages in PR systems, feature extraction plays a vital role in building system models, the recognition process, and system accuracy [17]. Feature extraction is a task to detect/extract the maximum amount of various desired attributes and characteristics from the input data. The features are the information that is fed to the recognizer to build a system model [18]. They should be insensitive to irrelevant variability in the input as much as possible, limited in number to permit for effective computation of discriminant functions and should not be similar, redundant, or repetitive. Various kinds of features can be found and/or calculated for an object in a PR system. Usually, features are categorized into global transformations [19], structural [20], statistical [21], and template-based matching [22].

The structural features describe the geometrical and topological characteristics of patterns using their global and local properties [23]. They are the most popular features investigated by researchers in OCR systems, because they are intuitive aspects of writing [24]. Structural features are less influenced by sources of distortion, but they are highly dependent on the style of writing. They may be extracted from each row, column, skeleton, or contour of an image. However, extracting this kind of features from the pattern images is not usually an easy task.

The statistical features are derived from the statistical distribution of the image's pixels and describe the characteristic measurements of a pattern. They include numerical values computed from a part or a whole of an image. Although these features are easy to extract, they can lead the system to the wrong way, because most of them are very sensitive to noise, scale, rotation, and other changes in the patterns.

Transformation features are derived from the transformed representation of an image in a new space by transformation operators. The transformation process maps an image from one space to another space, and it usually reduces the dimensionality and order of computing in the new space. They provide features that are invariant to global deformation like translation, dilation, and rotation [24].

Template-based features are usually created by matching predefined templates on graphical input data. However, they are completely data dependent, and they cannot transfer from a PR system to another.

Some of the most used structural features [14, 17, 20, 25, 26], statistical features [18, 21, 27, 28], and transformation features in OCR applications are shown in Table 1.

2.2. Dimensionality Reduction in OCR Application. Various features are computed and/or extracted in the feature extraction block of an OCR system. However, it is possible that some of the extracted features correspond to very small details of the patterns or that some of them are a combination of other features (nonorthogonal features), while others might not have any efficacy in the recognition stage. Irrelevant or redundant features may degrade the recognition results, reduce the speed of learning algorithms, and significantly increase the time complexity of the recognition.
Table 1: Some of the most used features in OCR applications.

<table>
<thead>
<tr>
<th>Type of features</th>
<th>Some of the most used features</th>
</tr>
</thead>
</table>
| Structural       | (i) Simple, double, and complex loops  
(ii) Loops positions, their types, and their relative locations 
(iii) Hills and valleys 
(iv) Open curves in different directions 
(v) Ascenders and descenders 
(vi) Numbers and locations of dots in each character 
(vii) Location of dots relevant to baselines 
(viii) Starting, ending, branching, crossing, turning, and corner points in image skeleton 
(ix) Curvature and length of the image’s segments 
(x) Length of a character segment relative to other segments 
(xi) Location of a character segment relative to center of mass image’s skeleton |
| Statistical      | (i) Normalized central, Zernike, pseudo Zernike, fast Zernike, Legendre, orthogonal Fourier-Mellin, rotational, and complex moments extracted from the whole body, only from the main body, or only from the secondary parts of an image 
(ii) Fourier and gradient descriptors 
(iii) Image area and image perimeter 
(iv) Pixels distribution in left, right, up, and down halves of the image 
(v) Image density 
(vi) Aspect Ratio 
(vii) Mean, mode, variance, and 2D standard deviation 
(viii) Average and variance of X and Y changes in portions of the image skeleton 
(ix) Local maxima points in horizontal and vertical projection histograms 
(x) Ratio of horizontal variance histogram to vertical variance histogram 
(xi) Ratio of upper half variance to lower half variance of an image 
(xii) Thinness ratio 
(xiii) The ratio of pixel distribution between two or more parts of the image 
(xiv) Center of mass (COM) (center of gravity) |
| Transformation   | (i) M-band packet wavelet coefficients 
(ii) Fourier, DCT, and radon coefficients |

Feature selection is typically a search problem to find an optimal subset with \( m \) features out of the original \( M \) features, to exclude irrelevant and redundant features from the initial feature vector. It reduces problem dimensionality, reduces system complexity and processing time, and increases the system accuracy [30]. In this respect, some feature subset selection algorithms have been proposed. According to the criterion function used for finding one \( m \) member subset out of \( 2^M \) possible subsets (\( M \) is the number of initial features), two general categories are introduced for this important task: Wrapper algorithm and Filtering algorithm. In the Wrapper algorithm, the classifier performance is used to evaluate the performance of a feature subset. In the Filtering algorithm, the features evaluation function is used rather than optimizing the classifier performance. In this category, by
using a special feature evaluation function, the best individual features are found one by one. However, the \( m \) best features are not the best \( m \) features [31]. Usually, the Wrapper methods are slower but perform better than the Filtering methods [32].

Based on the removing strategies, the feature selection methods are categorized into three groups. The first category is the Sequential Backward Selection (SBS) technique [33]. In this approach, the features are deleted one by one and the systems’ performance is measured to determine feature performance. However, finding the correct sequence to delete the features one by one is also very important. This means that a system’s derived efficiency after deleting features \( A, B, \) and \( C \) is not equal to the same system’s derived efficiency after deleting the features in order \( A, C, B \), or \( B, C, A \), and so on. Due to their nature, some features are relevant to others from different points of view. In this case, the SBS technique does not help to find the best subset of features. El-Glaly and Quek [33] extracted 4 set features \( S_1, S_2, S_3, S_4 \) to use them in an OCR application. They trained an OCR system with those 4 feature sets separately. After that, sets \( S_1 \) to \( S_4 \) were delivered to a Principal Component Analysis (PCA) algorithm, which rearranged the features in each set, based on their importance in the recognition system. The results showed that a feature (e.g., \( X \)) in rank 23 in set \( S_3 \) (23rd feature in feature vector \( S_3 \) after applying PCA) took rank 7 in set \( S_1 \) (7th feature in feature vector \( S_1 \) after applying PCA), and so on. This experiment indicates that if feature \( X \) is deleted for the sake of feature reduction, it may cause a large error in the final results.

The second group of feature selection methods comprises the random search methods, such as Genetic Algorithms (GA), which keep a set of the best answers in a population. Bahmani et al. [34] used GA in a handwritten OCR system. The initial number of features in their proposed system was 81 and the accuracy was 77%. After applying GA, the number of features was reduced to 55 and the accuracy increased from 77% to 80%. Soryani and Rafat [35] employed GA to carry out feature subset selection in a typical printed OCR system. They tested the proposed method to identify 5 fonts and 5 sizes of printed alphabet characters, achieved a reduction in the number of features from the initial 256 loci features to 146, and enhanced the system accuracy by 4.07% as well. The GA methods always select chromosomes one by one with the best recognition percentage and move those chromosomes (features) to the next stage. However, it is possible that when a good characteristic feature gets combined with another feature, the overall performance is not as good as the individual performances.

The third method for feature selection operation is a group of methods that have been applied to find important patterns in high-dimensional input data such as Principal Component Analysis (PCA) [12] and Random Projection (RP) [36]. PCA is a statistical method that tries to convert a correlated features space into a noncorrelated features space. In the new space, features are reordered in decreasing variance value such that the first transformed feature accounts for the most variability in the data. Hence, PCA overcomes the problem of high dimensionality and collinearity [37]. RA technique is a simple powerful dimension reduction technique that uses Random Projection matrices to map the data from a huge-dimensional space to a lower space one [36]. To achieve this aim, a mapping matrix \( R \) is used, where the columns of \( R \) are realizations of independent zero-mean normal variables, scaled to have unit length.

Some researchers employed PCA for various applications in PR systems. Gesualdi and Seixas [38] used PCA for data compression in the feature extraction part of a licence plate recognition system. They reduced the number of features from 30 to 4. The reported results showed that the achieved accuracy for digits recognition was acceptable but that the accuracy for characters recognition was degraded significantly. Ziaratban et al. [39] extracted a set of feature points including terminal, two-way, and three-way branch points from the skeletons of the characters in an OCR application. Finally, each skeleton was decomposed into some primitives, which are curved lines between any two successive feature points. Since the number of primitives varies from character to character, they used PCA algorithm to reduce and equalize the lengths of the feature vector. Using a postprocessing stage, they achieved 93.15% accuracy on a dataset with 7647 test samples. To recognize isolated handwritten Arabic letters, Abandah et al. [16] extracted 95 features from all feature categories. After that, only the first 40 features were selected from the PCA process result. Finally, five different classifiers were employed and, on average, 87% accuracy was achieved in the best case. Although PCA has been widely utilized in different PR applications, it suffers from high computational cost. In addition, after applying PCA, finding the order of the most effective to least effective features in the initial feature vector is not possible.

### 3. The Proposed Model

Figure 1 depicts the overall structure of the proposed dimensionality reduction model. First of all, some common pre-processing operations were carried out on the input images to enhance the quality of the input patterns. Then, the most used features—in the literature—were extracted as the initial feature vector denoted by \( S_{\text{Initial}} \). Dimension reduction was taken in two stages. In the first stage, using the proposed tools, One-Dimensional Standard Deviation (1D_SD) and One-Dimensional Minimum to Maximum (1D_MM) spectrums, the number of features was decreased from \( n \) in the \( S_{\text{Initial}} \) to \( k \) in the first reduced version of feature vector \( S_1 \). Then, in the second stage, and by employing the proposed tools Two-Dimensional Standard Deviation (2D_SD) and Two-Dimensional Minimum to Maximum (2D_MM) spectrums, the number of features was decreased again from \( k \) in feature vector \( S_1 \) to \( p \) in the final reduced version of feature vector \( S_2 \). The operations mentioned are demonstrated in the following subsections.

#### 3.1. Preprocessing

The performance of an OCR system depends very much upon the quality of the original data. In this context, we took into consideration that the proposed algorithm should be nonsensitive with respect to the scaling, rotation, and transformation of patterns. Hence, some
important preprocessing operations, such as noise removal, dimension normalization, and slant correction using common powerful techniques, are first performed on the samples. We applied a median filter with a 3 × 3 window and also morphological opening and closing operators using dilation and erosion techniques for high-frequency noise removal [23]. The image size was normalized without making any changes to the image Aspect Ratio, and, as a result, the width or height (or both) was changed to 50 pixels and the image was located in the center of a 50 × 50 pixels bounding box.

The body of every English and Arabic/Farsi digit is constructed using only one component. Thus, after the preprocessing operations, if it is found that there is still more than one group of connected pixels in the image of the digit, the extra blocks are considered as noise or separate components of the initial image. To find and remove the rest of the noise, the pen width is estimated using three different methods, and then the average of those values is considered to be the final pen width. To achieve this, we compute

(a) the mode of image vertical projection;
(b) (the value of image density)/(the number of image skeleton pixels);
(c) [(the value of image density)/(the number of image outer profile pixels)] × 2.

The results of the experiments showed that the average of three values is a more accurate estimate of the pen width than each of the values alone. After finding the pen width, all small components with a pixel density that is less than two times of the pen width can be considered to be noise and can be deleted from the input image. The threshold of 2 was obtained experimentally. The rest of the connected components are considered as broken parts of the digit image.

In order to connect the broken image segments together, we used a new approach. By using connected component analysis, we named the biggest available part as the main part \( M \) of the image. The outer contour of the main part \( M \) was then extracted and the coordinates of its pixels were saved in array \( MAIN \). Thereafter, for all of the rest secondary components \( S_i \) (which are smaller than the main part \( M \)), we found the outer contour and saved the pixels coordinate of those outer contours in another array \( SEC \). Then, we computed the Euclidean distance between all elements of array \( MAIN \) with all elements of array \( SEC \). The smallest value of the computed distance indicates the shortest path between contour \( M \) and one of the secondary contours \( S_i \). Finally, we drew a line with thickness equal to the estimated pen width along the shortest path between \( M \) and \( S_i \). As a result, the main part \( M \) is connected to a secondary part \( S_i \). This process was repeated until there is not another secondary component. A new version of main part \( M \) is used in each iteration of the algorithm, because, in each iteration, one secondary part is connected to the old version of main part. Algorithm 1 demonstrates the pseudocode for this process.

The images in the Hoda dataset are in binary format, while the images in MNIST are in grey level format. Hence, by analysing the grey level histogram for the image and using the standard global Otsu’s method [23], the MNIST samples were also changed to bilevel images. The method proposed by Hanmandlu et al. [40] was used to correct the slant angle of each image. First, an image is divided into upper and lower halves. Afterwards, the centres of mass points for these two parts are calculated. The slope of a line which connects these two mass point centres is considered to be the slant angle and the image is rotated in the reverse direction of this value.

We applied the proposed method on the digits part of the MNIST dataset [41] and the Hoda dataset [42] to connect the broken parts of the digit images. The results were encouraging as we were able to achieve 95.11% and 97.16% successful connections for the MNIST and the Hoda datasets, respectively. Figures 2(a) and 2(b) show examples of the above-mentioned preprocessing operations on two sets of training digits from the Hoda and MNIST datasets, in order.

3.2. Feature Extraction. Due to the vast diversity in writing style, handwritten characters are placed in the high-dimensional dataset category. Hence, finding an optimal, effective, and robust feature set to utilize in the recognition phase of an OCR system is usually a complex task. In this research, based on the literature, an initial feature vector, \( Initial_S \), including 133 of the most used features for both English and Arabic/Farsi digits, was extracted from the input images. Some of the extracted features are

(i) Aspect Ratio;
(ii) image area, perimeter, diameter, extent, eccentricity, and solidity;

Figure 1: The 2-stage proposed model for dimensionality reduction (feature selection).
while (there is another secondary component in input image) do
{
    find outer contour of the main part $M$;
    save the pixels coordinate of $M$ in array $MAIN$;
    repeat
    {
        find outer contour of an image secondary part $S$;
        save the pixels coordinate of $S$ in array $SEC$;
    } until (there is not another secondary parts in image);
    for (each pixel $A$ in array $MAIN$)
    {
        for (each pixel $B$ in array $SEC$)
        {
            compute the distance $d$ between pixels $A$ and $B$;
            save ($d$, coordinate of pixel $A$, coordinate of pixel $B$) in array $D$;
        }
    }
    $d_{min}$ = smallest value $d$ in array $D$;
    $A_{min}$ = coordinate of pixel $A$, corresponding to $d_{min}$;
    $B_{min}$ = coordinate of pixel $B$, corresponding to $d_{min}$;
    draw (a straight line with pen_width thickness from $A_{min}$ to $B_{min}$);
}

Algorithm 1: The proposed procedure to connect the broken image segments.

<table>
<thead>
<tr>
<th>Initial shape</th>
<th>Enhanced shape</th>
<th>Initial shape</th>
<th>Enhanced shape</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
<td><img src="a" alt="Image" /></td>
<td><img src="b" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 2: Applying proposed connecting procedure on input images.

(iii) Euler’s number;
(iv) centre of mass (COM) and centroid distance;
(v) pixel distribution density in up, down, left, and right halves of the normalized image;
(vi) pixel distribution density in upper and lower main diagonals of the normalized image;
(vii) ratio of pixel distribution in different quarters of the normalized image to each other;
(viii) ratio of horizontal variance histogram to vertical variance histogram;
(ix) ratio of upper half variance to lower half variance of an image;
(x) normalized horizontal and vertical transitions;
(xi) maximum horizontal and vertical crossing counts;
(xii) average of multiplication distances $X$ and $Y$ from COM;
(xiii) average of distances $X$ and $Y$ from boundary;
(xiv) ratio of major to minor axes lengths;
(xv) convex area;
(xvi) number and location of start, end, branch, corner, and crossing points in image skeleton;
(xvii) normalized invariant moments to order 7;
(xviii) discrete cosine transform coefficients related to the main image to order 9;
(xix) top, down, left, and right concavities in the image skeleton;
(xx) number of modified horizontal and vertical transitions;
(xx) average distance and average angular distance of each foreground pixel in a subimage from a virtual origin.
3.3. Dimensionality Reduction (Feature Selection). In our approach, we proposed one-dimensional and two-dimensional spectrum diagrams for standard deviation and minimum to maximum distribution.

3.3.1. Stage 1: Reduction Using 1D_SD and 1D_MM. In this stage, some candidate features were selected through 1D_SD and 1D_MM spectrums. In digit recognition domain, there are 10 classes corresponding to 10 digits, 0 to 9. Hence, for each feature from the initial feature set, a 1D_SD diagram is plotted with 10 spectrum lines corresponding to digits 0 to 9. In the 1D_SD plot, a spectrum line, corresponding to a specific feature, is drawn from the mean $-\text{std}$ to the mean $+\text{std}$ for each class. Figures 3(a) and 3(b) show the 1D_SD distribution diagrams corresponding to the “X Coordinate Centre of Mass” and “Normalized Vertical Transition” features for English digits, respectively. In Figure 3(a), the majority of spectrum lines are in an overlapping range $[20, 25]$, meaning that the “X Coordinate Centre of Mass” feature alone cannot discriminate existing classes from each other in the feature space. In Figure 3(b), the spectrum line corresponding to class (digit) 1 is completely separate from other spectrum lines, indicating that the “Aspect Ratio” feature can completely discriminate digit 1 (class 1) from other Arabic/Farsi digits set (other classes). Therefore, it can be considered as a candidate feature in the final features vector.

Finding a set of separated spectrum lines using only 1D_SD distribution diagrams is not enough to create an optimum feature vector, because the outlier samples in each class are not placed in the range of 1D_SD spectrum lines. Indeed they are put in One-Dimensional Minimum to Maximum (1D_MM) spectrums.

In the 1D_MM plot, a spectrum line corresponding to a specific feature is drawn from the minimum to the maximum value of that specific feature for each class. A shorter spectrum line corresponding to a specific feature indicates that the existing samples in a particular class have more similarity (less diversity) to each other in respect to that feature. Hence, a shorter spectrum line is better than a longer one. In addition, a distribution diagram with class centres (locations of the means of the classes) further apart is better than one with closer class centres. In this case, a classifier separates the existing clusters better.

Figures 5(a) and 5(b) illustrate 1D_MM spectrum lines for the same features, “Normalized Vertical Transition” (Figure 3(b)) and “Aspect Ratio” (Figure 4(b)), for the English and Arabic/Farsi digit sets, respectively. It is obvious that, in Figure 5(a), some samples of class 1 overlap with some samples in all the rest of the classes. This means that, in the recognition phase, these samples may be misclassified as belonging to other classes and vice versa, if only the “Normalized Vertical Transition” feature is employed. In addition, in Figure 5(b), some samples of class 1 overlap...
with some samples in classes 2 or 9. In other words, in the recognition phase, it is possible that some samples of class 1 are misclassified into classes 2 or 9 and vice versa, if only the "Aspect Ratio" feature is utilized.

In our proposed dimensionality reduction method, 1D-MM is utilized to find the maximum allowable overlapping threshold $T_1$ to create the first reduced feature vector $S_1$ from the initial features set, Initial S. By investigating the overlapping values of the spectrum lines in the 1D-MM diagram for each feature in Initial S, the value of threshold $T_1$ is selected. In this study, the $T_1$ threshold was set to 30%, experimentally.

3.3.2. Stage 2: Reduction Using 2D SD and 2D MM. Similar to the 1D SD and 1D MM distribution diagrams, the Two-Dimensional Standard Deviation (2D SD) distribution diagram and the Two-Dimensional Minimum to Maximum

**Figure 4: 1D_SD spectrum distribution diagrams for the Arabic/Farsi digits set.**

**Figure 5: 1D_MM spectrum distribution diagrams.**
In our proposed dimensionality reduction method, 2D_MM is utilized to find a maximum allowable overlapping threshold $T_2$ to create the final reduced feature vector $S_2$ from the first reduced feature vector $S_1$. By investigating the overlapping values of the spectrum ellipse (rectangular) in the 2D_MM diagram for the pair features in $S_1$, the value of the threshold $T_2$ is selected. In this study, the $T_2$ threshold was set to 20%, experimentally.

### 3.3.3. Creating Final Feature Vector

For the dimensionality reduction process, we defined the value of a specific feature as $f_k(S_i)$, where $f_k$ is the value of the $k$th feature from the initial feature vector. Initial $S$, and $S_{ij}$ represents the $j$th sample of class $i$. Subsequently, using all samples in the training part of each class, the values of the minimum, maximum, mean, and standard deviation for all the features in the initial feature vector were computed.

To find the first reduced subset of feature vector, the 1D_SD distribution diagram along with the 1D_MM spectrums was generated for all 133 features in initial features set, $Initial_S$. The system selected every feature for which its 1D_SD spectrum line had a maximum of 30% overlapping (threshold $T_1$) with the other 1D_SD spectrum lines of the other classes. The output of this stage was the first reduced version of the feature vector $S_1$, which satisfied the criteria necessary for membership in the final feature vector.

Finally, by using the 2D_SD distribution diagrams and also the 2D_MM spectrums on the first reduced version of the feature set $S_1$, the final reduced versions of feature vectors $S_2$ were selected. In this stage, a couple of features were selected, if the 2D_SD had a maximum of 20% overlapping (threshold $T_2$) with the other 2D_SD distribution diagrams. In stage 2, it was possible that a feature was added to $S_2$ more than once. Hence, in the final step, the repetitive features in $S_2$ were removed to create the smallest size of $S_2$.

In our proposed method, the 1D_SD and 2D_SD spectrums were utilized to decide whether or not a feature was suitable for including in the final feature vector. 1D_MM and 2D_MM were used to find the best value for thresholds $T_1$ and $T_2$. It is completely clear that these threshold values are dependent on characteristics of training dataset samples. Algorithm 2 explains the proposed method for the dimensionality reduction operation.

The mentioned operations created feature vectors $Initial_S$, $E-S_1$, and $E-S_2$ for the English dataset MNIST and feature vectors $Initial_S$, $A/F-S_1$, and $A/F-S_2$ for the Arabic/Farsi dataset Hoda. Table 2 shows the number of features in each stage for these datasets.

The following are among the selected features in the final feature vectors $E-S_2$ and $A/F-S_2$:

(i) X Coordinate Centre of Mass;
(ii) Number of Foreground Pixels in Upper Half;
Extract most-used features (in literature) from input dataset, and Create initial features set Initial $S_1$; 
$n := \text{Number of features in initial features set Initial $S_1$};$
first reduced version of feature vector $S_1 := \text{null};$

Stage 1: 
Choosing threshold $T_1$, by investigating 1D/MM spectrum diagrams.
for $(k = 1 : n)$
| Compute the coordinate of all 1D SD spectrum lines corresponding to feature $f_k$; |
| for $(c = 1 : \text{number of classes})$
  | If (overlapping of spectrum line of class $c$ with all the rest spectrum lines has the value less than threshold $T_1$) then |
  | Insert feature $f_k$ to $S_1$; |
  | goto L1; |
| } |
L1: continue; |
$\text{m} := \text{Number of features in first reduced version of features vector } S_1;$
final reduced versions of feature vectors $S_2 := \text{null};$

Stage 2: 
Choosing threshold $T_2$, by investigating 2D/MM spectrum diagrams.
for $(k = 1 : \text{m})$
  | for $(h = 1 : \text{m})$
    | Compute the coordinate of all 2D SD spectrum ellipses corresponding to features pair $(f_k, f_h)$; |
    | for $(c = 1 : \text{number of classes})$
      | if (overlapping of spectrum ellipses of class $c$ with all the rest spectrum ellipses has the value less than threshold $T_2$) then |
      | Insert feature pair $f_k$ and $f_h$ to $S_2$; |
      | goto L2; |
    | goto L2; |
| delete the repetitive features from $S_1$; |

Algorithm 2: Two-stage dimensionality reduction procedure.

Table 2: Number of features in initial feature vector, first reduced version of feature vector, and final reduced version of feature vector.

<table>
<thead>
<tr>
<th></th>
<th>Initial feature vector</th>
<th>First reduced versions of feature vectors $S_1$</th>
<th>Final reduced versions of feature vectors $S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>English dataset MNIST</td>
<td>133</td>
<td>$E-S_1$: 103</td>
<td>$E-S_2$: 79</td>
</tr>
<tr>
<td>Arabic/Farsi dataset Hoda</td>
<td>133</td>
<td>$A/F-S_1$: 94</td>
<td>$A/F-S_2$: 58</td>
</tr>
</tbody>
</table>

(iii) Number of Foreground Pixels in Lower Half;
(iv) ratio of foreground pixels to area of bounding box;
(v) ratio of number of foreground pixels upper main diagonal to number of foreground pixels under main diagonal;
(vi) Aspect Ratio;
(vii) normalized horizontal transition;
(viii) maximum horizontal crossing count;
(ix) Normalized Vertical Transition;
(x) variance for vertical histogram;
(xi) solidity;
(xii) perimeter;
(xiii) ratio of major to minor axes lengths;
(xiv) convex area;
(xv) number of end points;
(xvi) number of end points in different zones of image bounding box;
(xvii) number of branch points;
unbalanced dataset. This means that the sample frequencies
training and 10,000 test samples [41]. This dataset is an
Standards and Technology (MNIST) dataset contains 60,000
4.1.1. MNIST Dataset.
The Modified National Institute of
ations demonstrate these datasets briefly.
particulate matter, are Al-Isra, ARABASE, IFHCDB, CENPARMI, Hadaf, LMCA, and Hoda.

In order to test the effectiveness of the proposed method, the digit parts of two big handwritten standard benchmark datasets were utilized, namely, MNIST, for English numerals, and Hoda, for Arabic/Farsi numerals. The following subsections demonstrate these datasets briefly.

4.1.1. MNIST Dataset. The Modified National Institute of Standards and Technology (MNIST) dataset contains 60,000 training and 10,000 test samples [41]. This dataset is an unbalanced dataset. This means that the sample frequencies
for different classes in training—and also the testing part—are not equal (Table 3). All the digits have been stored in 28 × 28 image pixels, with intensities from 0 to 255. Figure 8 shows some sample digits from this dataset.

4.1.2. Hoda Dataset. The Hoda dataset is a very large corpus of Arabic/Farsi handwritten alphanumeric characters [42]. It has two parts: digits and characters. The digit section of the Hoda dataset was prepared in 2007 by extracting the images of the digits from 11,942 registration forms related to university entrance forms. Those forms were scanned at 200 dpi in 24-bit colour format. The digits were extracted from the postal code, national code, record number, identity certificate number, and phone number fields of each form. The digit section of the Hoda dataset has 80,000 samples and has been divided into two parts, namely, training (60,000 samples) and testing (20,000 samples). This dataset is a balanced dataset. It includes 6,000 and 2,000 samples for each digit in the training and testing parts, respectively. Figure 9 shows some sample digits from this dataset.

Table 3 includes the distribution of digits in training and testing parts of MNIST and Hoda datasets.

4.2. Proposed Method. In this research, the same operations were carried out in the preprocessing step on the training and testing samples. The outputs were noise filtered, reslanted, relocated, and dimension normalized.

Several experiments were carried out to test the effectiveness of the proposed method for dimensionality reduction.

In the first part, we applied the proposed approach for the recognition of handwritten English digits. In all experiments, a multilayer perceptron neural network with backpropagation was trained with 103 (or 79) neurons in the input layer (corresponding to the number of features in sets $E_{S1}$ and $E_{S2}$), 30 neurons (found experimentally) in the hidden layer, and 10 neurons (corresponding to 10 different classes of digits 0 to 9) in the output layer, respectively.

In the first experiment, a neural network was employed with 103 (number of features in set $E_{S1}$) neurons in the input layer. The network was trained with all 60,000 samples from the training part of the MNIST dataset and was then tested with all 10,000 samples from the testing part of the MNIST dataset. This operation was repeated 10 times, and, finally, 93.17% accuracy was achieved on average in this stage.

To compare the performance of the proposed method against other well-known feature selection techniques, a general Principal Component Analysis (PCA) technique and a Random Projection (RP) technique [36] were applied on the initial reduced feature set $E_{S1}$ with 103 features. PCA changed the order of the features in the new orthogonal feature space—based on the derived eigenvectors—and generated a new reordered feature set $E_{S2}$ with the same 103 features. The complete reordered feature set $E_{S2}$ was fed into the same MLP-NN and a final accuracy of 93.82% was achieved. This result was 0.65% higher than the 93.17% result, portraying the superiority of the PCA technique for feature selection. Similarly, we employed RP dimensionality reduction technique to create a new smaller feature space.
Table 4: Recognition rate corresponding to different feature vectors for the English dataset MNIST and Arabic/Farsi dataset Hoda.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Number of features in feature vector</th>
<th>Feature selection method</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Proposed method</td>
<td>PCA method</td>
</tr>
<tr>
<td>Initial S</td>
<td>133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;1&lt;/sub&gt;</td>
<td>103</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;2&lt;/sub&gt;</td>
<td>79</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>English dataset MNIST</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;3&lt;/sub&gt;</td>
<td>103</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;4&lt;/sub&gt;</td>
<td>79</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;5&lt;/sub&gt;</td>
<td>103</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>E-S&lt;sub&gt;6&lt;/sub&gt;</td>
<td>79</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Arabic/Farsi dataset Hoda</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial S</td>
<td>133</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;1&lt;/sub&gt;</td>
<td>94</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;2&lt;/sub&gt;</td>
<td>58</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;3&lt;/sub&gt;</td>
<td>94</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;4&lt;/sub&gt;</td>
<td>58</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;5&lt;/sub&gt;</td>
<td>94</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>A/F-S&lt;sub&gt;6&lt;/sub&gt;</td>
<td>58</td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9: Some sample digits in the Hoda dataset.

(with 103 features) from the initial feature set, Initial S. The output of this stage was the new feature set E-S<sub>2</sub>. The feature set E-S<sub>2</sub> was fed into the same MLP-NN, too. In this experiment, a final accuracy of 93.51% was achieved that is 0.31% lower than the achieved result by using PCA.

In the second experiment, the system was trained with the proposed final version of feature set E-S<sub>2</sub> with only 79 features. On average, the correct recognition rate increases from 93.17% to 94.88%, clearly indicating the superiority of the reduced feature set E-S<sub>2</sub> of the proposed method against the initial reduced feature set E-S<sub>1</sub> with 103 features. To find the superiority of the our proposed feature set, E-S<sub>2</sub>, compared to the other subsets of Initial S, which have 79 members, we made set E-S<sub>4</sub> with the first 79 members of set E-S<sub>3</sub> (generated by PCA) and set E-S<sub>6</sub> with 79 members using RP technique. The recognition rate declined dramatically from 93.82% to 90.71% by using set E-S<sub>4</sub> and from 93.51% to 88.39% by using set E-S<sub>6</sub>. These results obviously show the effectiveness and superiority of our proposed technique in comparison with PCA and RP as two of the popular techniques for feature selection operation. The obtained results also show the superiority of PCA technique compared to RP technique for dimensionality reduction purpose. However, it is worth mentioning that some researchers have shown the superiority of RP against PCA—for dimension reduction purpose—in high-dimensional feature space condition [43, 44]. The outcome results are reported in rows 2 to 8 of Table 4.

In the second part, we repeated all the experiments in the first part, for the Arabic/Farsi digits, using the Hoda dataset. A/F-S<sub>1</sub>, A/F-S<sub>2</sub>, A/F-S<sub>3</sub>, A/F-S<sub>4</sub>, A/F-S<sub>5</sub>, and A/F-S<sub>6</sub> were the first reduced feature vectors with 94 members, the final reduced feature vector created by our proposed method including 58 features, the reordered initial reduced feature vector by PCA including 94 features, the first 58 features from the A/F-S<sub>3</sub>, the smaller version of feature space with 94 features created by using RP technique, and also another smaller version of feature space with 58 features created by using RP technique, in order. It is worth mentioning that not only the trend was completely similar to previous experiments for the English digits dataset, but also the achieved results for the Arabic/Farsi dataset Hoda were better than the achieved results for the English dataset MNIST. In this case, using the proposed feature selection method, the number of features was reduced from 113 to 58; meanwhile, the final accuracy was increased from 90.41% to 95.12%. In this experiment, when we used the PCA technique for feature selection, the accuracy was decreased significantly from 94.04% to 89.00%, and when we employed the RP technique for feature selection, the accuracy was decreased from 91.07% to 83.66%. Here, we only employed 1/2 of training and testing samples from Hoda dataset. The corresponding results are shown in the last seven rows of Table 4.

5. Conclusion

In this paper, a new method for dimensionality reduction (feature selection) in pattern recognition systems was introduced. To begin with, an initial set of the most used features was extracted from the training patterns of two handwritten
digit standard datasets: MNIST for the English digits and Hoda for the Arabic/Farsi digits. Then, by using the proposed 1D_SD and 1D_MM distribution diagrams methods, the initial feature vector was reduced to a smaller version based on the maximum allowable overlap between the spectrum lines using the threshold $T_1$. Thereafter, by using the proposed 2D_SD and 2D_MM spectrums, a final reduced feature vector was selected. In this stage, another threshold, $T_2$, was used to guarantee that the overlaps between the spectrum diagrams were not more than the maximum threshold.

The mentioned algorithm was implemented in an OCR application system to reduce the dimension of the initial feature vector. For English MNIST dataset, the feature vector was decreased from 100% (133 elements) to 59.40% (79 elements); meanwhile, the accuracy was increased 2.95% (from 91.93% to 94.88%). The accuracy was 4.17% more than the outcome results in a similar experiment that used PCA (90.71%) and 6.49% more than outcome results in a similar experiment that used RP (88.39%) as two of the common techniques for the feature selection operation. For this experiment, we only employed 1/2 of the training samples (30,000 samples) and 1/2 of the testing samples (10,000 samples).

The results clearly indicate the superiority of the proposed method for dimensionality reduction (feature selection). According to the results, the proposed technique is completely effective for OCR application as a subcategory of PR systems. Nevertheless, the proposed new method can be used for other PR systems with different database types.

Along with proposing a new method for dimensionality reduction, this paper introduced a new method to connect the broken parts of an image in the preprocessing stage of an OCR system. This technique estimates the pen width in three different ways. By utilizing the connected component analysis, it traverses the outer contour of the separated blocks in an image and connects them together.

The proposed feature selection method can be considered as an approach to infer the appropriate rules for creating a decision tree classifier in a PR system. In other words, using the 1D_MM and 2D_MM spectrum diagrams, the necessary rules in a decision tree classifier are found more accurate and faster. This is a salient feature of our proposed approach.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This paper is supported by University of Malaya Research Grant BK026-2013.

References


[38] D. R. Gesualdi and J. M. Seixas, Character Recognition in Car License Plates Based on Principal Components and Neural Processing, The VII Brazilian Symposium on Neural Networks, 2002.


Submit your manuscripts at http://www.hindawi.com