Pioneer dataset and recognition of Handwritten Pashto characters using Convolution Neural Networks

Sulaiman Khan1, Abdul Hafeez2, Hazrat Ali3, Shah Nazir1 and Anwar Hussain1

Abstract
This paper presents an efficient OCR system for the recognition of offline Pashto isolated characters. The lack of an appropriate dataset makes it challenging to match against a reference and perform recognition. This research work addresses this problem by developing a medium-size database that comprises 4488 samples of handwritten Pashto character; that can be further used for experimental purposes. In the proposed OCR system the recognition task is performed using convolution neural network. The performance analysis of the proposed OCR system is validated by comparing its results with artificial neural network and support vector machine based on zoning feature extraction technique. The results of the proposed experiments shows an accuracy of 56% for the support vector machine, 78% for artificial neural network, and 80.7% for the proposed OCR system. The high recognition rate shows that the OCR system based on convolution neural network performs best among the used techniques.

Keywords
Convolution neural network, neural networks, Pashto, Handwriting recognition, support vector machine

Introduction
Handwriting refers to the writing accomplished by using pen or pencil. It differs radically from one person to another, and can may even be varies for a given individual depending on the mood and condition in which he/she is writing. Drastic differences in the handwriting of boys and girls can be observed at the stages as early as seven years. Unlike printed characters, handwritten characters are varying based on their curves or edges which make it vague by nature. Furthermore, different people have different styles of writing and a different size of characters as well. In the case of Pashto handwriting, the task of recognition becomes more complex as compared to the languages; including Arabic and English. Pashto is an important language in the Asian region as it is the official way of communication in Afghanistan as well as the major language spoken in Pakistan (northern areas i-e. Khyber Pakhtunkhwa).1

Approximately, 40 to 60 million people are the native speaker of the Pashto language.1 However, compared to the English and Arabic, the ratio of contribution regarding automatic recognition of the Pashto text is much low. However, significant contributions are added and efforts are made by experts from around the world.2

For the Arabic language, researchers exploited the Optical Character Recognition (OCR) systems based on character’s features such as its statistical measures, structure and morphological attributes. In the case of the Urdu language, OCR is exploited by including the water reservoir, topological and contour features.3 Furthermore, Kopparapu et al.4 suggested the application of machine learning techniques for recognition of handwritten Devnagari letters by using a neural classifier that uses chain line histogram, intersection, shadow, and straight-line fitting features. The fuzzy approach is adopted for the recognition of Arabic handwritten letters by using preprocessing techniques such as extracting contour, thinning and connected components detection. Pseudo-Zernike moments are used for the extraction of the features in their case. Classification in

1Department of Computer Science, University of Swabi, Swabi Pakistan
2Department of Computer Science, UET Jalozai, Pakistan
3Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad, Pakistan

Corresponding author: Sulaiman Khan, Department of Computer Science, University of Swabi, Mohallah Chino, Village & Post office Jalsai, Tehsil Lahor, Swabi, 23430, Pakistan.
Email: engr.sulaiman88@gmail.com
their approach is achieved using the Fuzzy ARTMAP neural network which is proved to be efficient as well as have support for incremental learning. Several machine learning approaches are adopted in the domain based on deep and shallow architectures. Ahmad et al.\(^1\) and Khan et al.\(^1\) proposed K-Nearest Neighbor (KNN) for the recognition of printed cursive scripts. Scale, rotation, and location invariant method are proposed by Ahmad et al.\(^8\) for the recognition of printed Pashto script.

Experts show a greater interest in enhancing the efficiency of existing recognition algorithms to provide more intelligent solutions to perform the recognition task.

From the literature cited, it is concluded that there is no significant work reported for the automatic recognition of the Pashto language. The curvy nature of the Pashto language and the inherent variations pose significant challenges for designing and developing an automated recognition system. Furthermore, previous work reported for the Pashto language is limited to the automated recognition of the printed characters only and not for the hand-written text and automatic recognition of handwritten text effectively remains unaddressed. This paper presents a decent size dataset (first of its kind) and an efficient, cost-effective and intelligent recognition framework. The contributions of this work are summarized below:

- A database of handwritten characters of Pashto language. The dataset will be made available for research purposes, free of cost.
- A benchmark on automatic recognition of handwritten characters of Pashto language. The reported results will provide a good baseline for future research work on Pashto text processing.
- The performance of three models namely: SVM, ANN, and CNN are reported for the recognition of handwritten Pashto characters. In this way, we present an insight into the performances of traditional machine learning approaches and deep learning approaches for the targeted language.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 presents the primary information regarding the algorithms used for the classification in the proposed research work. Methodology for the proposed research work is discussed in Section 4. Section 5 shows the experimental setup made for conducting the research and presents the results obtained. Finally, the concluding remarks are given in Section 6.

**Related work**

Several diverse approaches are available for text recognition systems. For the recognition of Persian/Arabic handwritten digits, Salimi and Geviki\(^7\) used singular value decomposition for feature extraction purposes and for the classification they have used KNN classification tool. Another approach exploited the support vector machine classification algorithm for the recognition of handwritten Arabic letters in combination to Orthogonal Fourier-Millen movements, T Chebyshey and Gegenbauer methods. Multiple preprocessing techniques include; centering algorithm, edge detection, thresholding, and median filter are reported in their work.\(^9\) Applications of morphological operations and genetic algorithm for the extraction and recognition of Persian letter by Keyvanpour et al.\(^9\) while feature extraction and recognition of Persian/Arabic zip code can be seen in the work of.\(^10\) For identifying the most suitable methods for the recognition among multiple alternatives, a survey is conducted by Impedovo and Pirlo,\(^11\) which analyzed several state of the art approaches such as the fuzzy approach, neural networks. Comparatively, techniques based on neural networks are considered to be efficient and reliable in the domain of the off-line recognition system. Significant recognition results are calculated for these techniques. In that regard, Ayyaz et al.\(^12\) used hybrid features extraction method for features extraction and multi-class Support Vector Machine (SVM).

Ali et al.\(^13\) developed an OCR system and a dataset for the Urdu language. Their OCR model uses convolution neural network and deep auto-encoder for the recognition purposes. Maalej and Khairallah improved the DBLSTM model for the online recognition of Arabic text. Memon et al.\(^14\) performed a systematic study of the OCR models developed for the recognition of the handwritten characters. Carbune et al.\(^15\) developed a multi-language OCR system for the recognition of online handwritten text recognition. Habib et al.\(^16\) proposed an OCR system for the recognition of degraded Urdu and Devanagari text. There model capable of performing preprocessing steps on the degraded images of the text to extract the significant features and perform the identification process. Many researchers also exploited the Zoning features extraction techniques for acquiring significant numerical values from the character images. In this regard, recognition of Thai letters is achieved using an Antiminer algorithm that uses 4 × 3 zoning grid to extract end-point and closed loop features for a handwritten image.\(^17\) A variant of the Zoning grid of 4 × 4 is utilized by Negi et al.\(^18\) to derive pixel density in each zone. Cha et al.\(^19\) exploited the application of zoning grid for extracting the information from an image based on gradient, structure and concavity features. Similarly, adopting the same approach, Kimura and Shridhar\(^20\) extract information from the contour profile of an image. In this approach, to count the number of segments, four orientation steps are used such as; 00, +45°, −45°, 90°. Liu et al.\(^21\) used a zoning grid of 4 × 4 size for the recognition of the Chinese letter with a technique known as directional decomposition.

Sharma and Gupta proposed variant zoning grids including 4 × 4, 6 × 6 and 8 × 8\(^22\) for the recognition
of text. These variants are used over a pattern image to measure the density of pixels, while, 5 × 5 grid is utilized for feature extraction. For feature extraction, the average distance is calculated between the pixel and the letter centroid in a row or column. Adapting the same configuration of the grid, local density is measured in a letter image while for profile projection features extraction, zoning grid of 1 × 10 and 10 × 1 is used for extracting these features in the vertical and horizontal directions. Studies including based on an approach that uses circular ring partitioning criteria as well as convex hull ring. Their presented approach, is shaped based where convex hull circles are made using a letter 'convex hull shape. In their approach concentric circles refer to a set of circular rings and the center of these concentric rings is the pattern’s minimum enclosing center. According to arithmetic progression, selection for the radii of circles belonging to the set is made. For making the rotation invariant for the set of the system, features are based angular information of the letter pixels’ internal and external contour.

Based on studying the literature, many interesting approaches work for the text recognition in major languages such as Urdu, Arabic, and Persian languages. However, for the recognition of handwritten letters in the Pashto language, no significant work is reported. Keeping in view the efforts made for the recognition of handwritten Pashto letters, we presented a novel framework for the recognition of handwritten characters in the Pashto language named as PHCR system.

**Background**

The section gives a background detail of characters modeling for Pashto language and a discussion on the classification techniques namely; support vector machine, artificial neural network and convolution neural network.

**Pashto script**

Several pronunciations are suggested for the Pashto such as Pukhto, Pusho, or Pakhto. It is also known as Afghani in the context of Indo-European language as well as treated as a sister language of Indo-Iranian language family too. Dialect of Pashto language are mainly categorized as “soft” and “hard” which are known as southern and northern respectively. Major differences between these dialects are on the phonological level, such as northern pronunciation references “Pashto” word as “Pakhto” or “Pukhto” and the southern refer the same word as “Pashto”. In the context of this paper “Pashto” word is used in both of the dialects that will reference the same word “Pashto”. There is one other dialect that is considered as standard and used as a spelling system or manuscript for Pashto words known as Pata Khazana.

Pashto is derived from Perso-Arabic characters as a modified pattern of it. Pashto writing can be done in a version of Persian script that is a derivation from the Arabic script as well. There are a total of 44 characters in the Pashto language as shown in Figure 1. For written Pashto script, several characters can be merged to make a word as well as a combined character. Furthermore, according to the position of a word, the shape of the character change as well. Due to some variations in character’s shape during the formation of the word make it more challenging as the shape of the character varies to its appearance in the word. For example shape of a character appearing at the start of a word is different if the same character appears in the middle or at the end of the word as shown in Table 1.

| Isolated letter | Letter’s shape at the start | Letter’s shape at the middle | Letter’s shape at the end |
|-----------------|-----------------------------|-----------------------------|--------------------------|
| ت               | ت                           | ت                           | ت                         |
| پ               | پ                           | پ                           | پ                         |

**Figure 1. Standard Pashto character set.**
process more complex. However, in this paper, we have confined our study to the recognition of isolated characters. As discussed earlier that alphabets of the Pashto language are derived from the Persian alphabets. For defining phenomena that should be more specific to Pashto, several characters are summed and modified. One of the most widely used script for writing Urdu and Persian is the Nasta’liq script. In case of Pashto language, both kind of script are used for its writing such as Nasakh and Nasta’liq. However, among these two scripts, Nasakh has used a standard to write Pashto script. The Pashto language includes the character belong to the Persian language in its characters dataset. The Pashto language includes 32 characters from the Persian language, among which it covers the 28 Arabic characters with addition to four character that is specific to the Persian language. On the other hand, the Urdu language that is a modified version of the Persian language consisting of six specific characters to the Urdu language while 32 Persian script characters. As a result, the Urdu language has a total of 38 characters. Interestingly Pashto language does include five Urdu specific characters with minor changes in their shape as shown in Table 2.

Seven characters that are specific to the Pashto language are shown in Table 3.

After accumulating all these characters, the Pashto language has a total of 44 characters in its dataset that is shown in Figure 1.

**Support vector machine**

Support Vector Machine (SVM) is followed as a successful machine learning techniques for the recognition and classification tasks. SVM is proposed by Cortes and Vapnik in the ends of 1990’s. Generally, SVM is used for the classification of linear and non-linear problems. Optimum results for the classification process can be obtained by resolving the quadratic task that depends on the regularization parameter. For their transformation, kernel functions are used such as radial basis, polynomial kernel, sigmoid, and linear.

- Linear Kernel – \( M(a, b) = a \times b \)
- Sigmoid Kernel – \( M(a, b) = \tanh(\beta_0 a b + \beta_1) \)
- RBF (Radial Basis Function) Kernel – \( M(a, b) = e^{-\gamma||a-b||^2} \)
- Polynomial Kernel – \( M(a, b) = \frac{(a \times b) + 1}{2} \)

With \( d, \beta_0, \beta_1 \) and \( \gamma \) are parameters that can be calculated empirically using equation (1).

\[
f(x) = W^T \theta x + b
\]  
(1)

Where \( w \in R \) and \( b \in R \) and \( \theta(x) \) is a feature map.

SVM is binary classifier while in our case there are 44 classes i-e. 44 Pashto characters (multi-class problem). In order to address this issue two different possibilities can be made to the SVM that are:

1. One versus all approach (1: M)
2. One versus each (iterative approach)

This research work uses one versus all approach (1: M) as shown in Figure 2.

One of the reasons for using this approach is that it is simpler in implementation as well as computationally efficient. Transformation is made by using a non-linear operator such as \( \theta(x) \) to map inputs \( a_i, b_j \) into high dimensional space. Equation (2) shows the optimal hyper-place in the described scenario.

\[
f(x) = \text{sgn}\left(\sum b_i x_i M(a_i, a) + y\right)
\]  
(2)

Where \( M(a_i, a) = e^{-\gamma||a_i-a||^2} \) the kernel function is calculated using radial basis function (RBF), while \( \text{sgn}(.) \) is sign function.

Overall SVM is a better classifier but in our case, it has poor accuracy. Furthermore, results of SVM are not good comparatively Artificial Neural network and the convolution neural network. If the results of these

---

**Table 2. Urdu character’s Pashto equivalent.**

| Urdu Letter representation | Pashto letter representation |
|---------------------------|-----------------------------|
| ت | ت |
| ڈ | ڦ |
| ڑ | ړ |
| گ | ګ |
| ے | ۍ |

**Table 3. Pashto specific characters.**

| | ي | ي | ي |
|---|---|---|---|
| ځ | څ | ږ | ښ |
| ي | ې | ﺉ |

**Figure 2. One versus all (1: M) approach for SVM.**

With \( d, \beta_0, \beta_1 \) and \( \gamma \) are parameters that can be calculated empirically using equation (1).

f\((x) = W^T \theta x + b\)

Where \( w \in R \) and \( b \in R \) and \( \theta(x) \) is a feature map.

SVM is binary classifier while in our case there are 44 classes i-e. 44 Pashto characters (multi-class problem). In order to address this issue two different possibilities can be made to the SVM that are:

1. One versus all approach (1: M)
2. One versus each (iterative approach)

This research work uses one versus all approach (1: M) as shown in Figure 2.

One of the reasons for using this approach is that it is simpler in implementation as well as computationally efficient. Transformation is made by using a non-linear operator such as \( \theta(x) \) to map inputs \( a_i, b_j \) into high dimensional space. Equation (2) shows the optimal hyper-place in the described scenario.

f\((x) = \text{sgn}(\sum b_i x_i M(a_i, a) + y)\)

Where \( M(a_i, a) = e^{-\gamma||a_i-a||^2} \) the kernel function is calculated using radial basis function (RBF), while \( \text{sgn}(.) \) is sign function.

Overall SVM is a better classifier but in our case, it has poor accuracy. Furthermore, results of SVM are not good comparatively Artificial Neural network and the convolution neural network. If the results of these
algorithms are compared, then we find the SVM accuracy much lesser against the 78% and 80.7% for ANN and CNN respectively. After much training and testing, it becomes evident that SVM performs for binary classification. However, in multi-class class problems (Pashto has 44 classes in our case), the SVM is not prominent as this is good in binary classification. The classification results of SVM are shown in Figure 14.

Artificial neural network

Artificial Neural Network works by predicting the results from the set of input variables. Inspired from the human brain where many neurons pass information to each other to perform certain pattern matching and recognition task efficiently and accurately the idea of ANN is developed. Due to the high classification and recognition abilities, ANN is exploited in the field of handwritten characters recognition.\(^{29}\) In their study, the features based on diagonals are extracted which are then used by ANN for recognition of handwritten alphabets. ANN is also used for solving quantum many-body problems.\(^{30}\) Feed forward Neural Network is presented by Zhao et al.\(^{31}\) for the recognition of characters via surface mount technology (SMT) products. The efficiency of SMT is further enhanced by introducing the additional back propagation algorithm in the approach. The approach exploits momentum item for that results in avoiding oscillations, divergences and convergence is also improved. In the proposed research work a back propagation neural network is used consists of an input layer, two hidden layers, and an output layer as shown in Figure 3.

Convolution neural network

As a result of having multi-layered and hierarchical neural network training via back-propagation procedure using architecture of deep supervised learning, Convolution Neural Network (CNN) has an automatic feature extraction capability as well as work as classifier. Main aim of the CNN is to work in the situation where we have more complex and data with many dimensional. CNN differs from other learning algorithms due to investigation of convolution layers and sub-sampling is performed. Due to the variations that occur in shapes, CNN are used for classification and recognition of objects, handwritten digits,\(^{32}\) as well as to recognize Urdu Nastaliq.\(^{33,34}\) CNN is usually used for the problems associated to pattern recognition to get the efficient performance. CNN is based on three

![Figure 3. Generalize two hidden layer neural network.](image)

![Figure 4. Conventional CNN model.](image)
main steps which is performed in hierarchy such as
weight sharing, spatial, and subsampling to insure the
invariance to distortion, shift, and scale respectively.
Model of the CNN for recognition of Pashto handwrit-
ten recognition is shown in Figure 4. In the first step,
filters (F hidden layers) are applied on the input data
which convolved the data for the object to get feature
map values. For the reduction of dimensionality such
as S hidden layer of features map’s spatial resolution,
convolution layers that are obtained from the previous
step and then engaged with the help of sub-sampling
layer. Furthermore, convolution layer trigger the
feature-extractor to get or retrieve features by altering
the sub-sample layers. Further these layers get engaged
by layers such as output and two fully connected layers
(FCL). Subsequently, the output from the previous
layer is accepted as an input by each subsequent layer.

The number of parameters used in the proposed
research work for the CNN model are:

| Input Layer – The input layer only reads the image and
| there is no learning process in this layer, so no para-
| meters learn in the input layer. |

**Convolution layer** – A filter size of $3 \times 3$ is selected with
| stride = 1. The number of parameter in the convolu-
| tion layer are $=(3 \times 3 + 1) \times 8 = 80$. |

| Pooling layer – The third POOL1 layer has no para-
| meters. It only reduce the dimension size of the image. |

| Fully connected layer – The parameters for the FCL is
| $= 256 \times 32 + 1 = 8,193$. |

| Softmax layer – Softmax layer has a total number of
| parameters $= 10 \times 256 + 1 = 2,561$. |

### Proposed methodology

Framework designed for the recognition of Pashto
character is achieved through implementing in using
three major steps or phases that is shown in Figure 5.

- Characters database development
- Features extraction
- Model training and recognition (SVM, ANN, CNN based classification)

This is discussed in the following sections.

### Character database development

To the best of our knowledge, there is no standard
data set available for handwritten Pashto letters. In
order to address this, we present a medium sized data-
base for the handwritten Pashto letters. The main steps
involved are discussed below:

**Repository of handwritten letters.** For the proposed hand-
written Pashto characters database development phase
several universities around Khyber Pakhtunkhwa,
Pakistan are visited for handwritten samples collection.
Repository or database of handwritten letter is made by
collecting samples of handwritten letter wrote by people
on a page of having A4 size. Each given page was sec-
tioned in six number of column in order to get number
of possible variations of Pashto letters or alphabets that
occur among the writing style of people. The handwrit-
ing samples were collecting from volunteers (mainly
students) who willingly participated in the samples col-
lection for use in this research study. The distribution
of volunteers is summarized in Table 4. During the
database development phase (characters accumulation
process) 98% of the writers are right handed and the

---

**Table 4.** Gender-wise collection of text samples of Pashto.

| Age (years) | Samples | Males | Females | % |
|------------|---------|-------|---------|---|
| 17–30      | 51      | 23    | 28      | 50|
| 35–40      | 31      | 13    | 18      | 30.4|
| 50–60      | 20      | 11    | 9       | 19.6|

---

**Figure 5.** Proposed PHCR system.
rest of the 2% are left handed, while the minimum age level of the volunteers at the university level was 17 years.

**Binarization.** The accumulated handwritten Pashto letter from different people are scanned in order to store it in the computer for further processing and applying machine learning algorithms. Figure 6 depicts the scanned samples for Pashto handwritten alphabets that were stored in the computer.

**Letters extraction.** After the accumulation of handwritten characters the next step is to extract each letter from the scanned samples to develop a handwritten Pashto characters database for the simulation and research purposes. The upcoming section gives information about the handwritten characters extraction and database development process. Figure 7 shows some examples of the extracted letter from scanned samples.

During the accumulation of the handwritten Pashto characters, a fixed size window is selected for each character, however, using the same algorithm to extract characters from the non-titled and title document causes many problems to perform extraction from the scanned samples. To avoid such a problem in character extraction, all the images are brought to have the same fixed size 64×64 in our case. Dark spots in the image have a high influence on the semantics of extracted words, even a single dot in the image can change the whole alphabet. To make the extracted words to be unaffected by these form of noises, we have performed noise removal for the extracted letters to train the algorithm on a more reliable set of images.

**Thresholding.** These spots are easily visible after zooming the extracted letter image as shown in Figure 8.

Removing black spots (noise) from the characters images, several methods like thresholding and average filtering are handy in such cases. In our case thresholding gives optimum results. On the other hand, averaging filter has the remove black spots along with the meaningful dots that are actually a part of the word. So removing these dots incur more problems instead of giving solutions. Therefore, in our approach, we have used thresholding for noise removal by defining the threshold to an optimum level so that the noise has to be removed but also keeping the potential dots in the images. Using different threshold values and experiments we have to find out 20 to be an optimum value in the context of the current setup. Equation (3) shows the mathematical depiction of the thresholding procedure we have applied:

\[
T(x) = \begin{cases} 
0 & \text{if } |x| \leq 20 \\ 
x & \text{otherwise} 
\end{cases}
\]  

(3)

Resultant images produced after thresholding is represented in Figure 9. The images shown in Figures 8 and 9 show that, the thresholded image is the better representation of the handwritten character.
Morphological operations and centralizing Letters

Noise can be produced due to the imperfect segmentation. So in this step, the noise in the shape of cutting edges and holes is removed from the images. Furthermore, each character or alphabet is positioned centralized to make each image equally a fair candidate for classifier as well as to eliminate image classification affected by the variations of the position in the images which is shown in Figure 10.

For the position of the letter to be centered, we have applied centralization techniques that calculate or measure the center of the images and subtract pixels such that the letter is centered now. The result of centralization is shown in Figure 12. With these steps, we finally get a dataset for the 44 letters of Pashto, where each letter appears with 102 variations. In the end, each character is stored in the dataset as an 64 × 64 image file. The dataset thus comprises of 4488 items.

With this dataset developed, we use this to train the model for automatic recognition of Pashto handwritten letters. To the best of our knowledge, this is the first available dataset for Pashto handwritten characters.

Normalization – In the normalization phase, slants are removed first from the handwritten characters and resized to a fix size (64 × 64 in our case). In the proposed PHCR system, the all the characters in the images are thinned to remove slants and resized to a fixed shape.

Feature extraction

For extracting features from handwritten characters, a zoning technique is selected. In this technique, a zoning grid is overlapped on the considered images and the significant feature values are accumulated based on the density of pixels, color or values of pixels in corresponding grids. This technique works on \( u \times v \) grid that divides an image into uniform sub-samples and the resultant sub-samples having identical shapes. These identical shapes know called zones. In Figure 11(a) we have represented an example of zoning methods of having uniform shaped on a grid with 2 × 2 dimensions. Furthermore, (Figure 11(b)) shows the variations of the technique on 3 × 2 dimension grid, effect for 3 × 3 grid is shown in (Figure 11(c)), (Figure 11(d)) represents 4 × 4, and 5 × 4 is depicted in (Figure 11(e)).

Numbers of approaches are proposed with variant size of grids in association to the evolving of distinctive part for handwritten characters. Such example includes grid of 2 × 2, 1 × 2, 2 × 1, and 3 × 2.\(^5\) For recognition of digits based on calculation counter is presented by
Morita et al.\textsuperscript{36} and Blumenstain et al.\textsuperscript{37} where they have used grid of $3 \times 2$ size. Variant of approaches such as $3 \times 2$ grid is used to examine handwritten characters.\textsuperscript{38} Zoning grids with $3 \times 2$ dimension is exploited for extracting counter-based features \textsuperscript{50} which is then further been adopted by Verma et al.\textsuperscript{39} for the recognition of handwritten character along with back-propagation neural network algorithm. Another approach uses zoning grid has $3 \times 3$ dimensions to extract distribution of geometrical features for each individual zone.\textsuperscript{40} Using the same zoning grid, features are extracted for handwritten characters and digits by Hewitt.\textsuperscript{41} Recently, advance filtering techniques are used by number of preprocessing approaches which can be seen from the studies of Imani.\textsuperscript{42}

In the proposed PHCR system, a zoning grid of $8 \times 8$ dimension is selected for feature extraction purposes as shown in Figure 12. Computationally, the grid size such as $16 \times 16$ is more expensive to be used due to the high consumption of resources. Furthermore, from experiments, it is found that for extracting a limited number of features, a grid with a size $8 \times 8$ is much faster in results. In the next step, we have also counted the density in each zone for a white pixel. Furthermore, the results of the zoning are normalized to make the values in the range $[1, -1]$. The skeleton of the image is worthy to make the efficient use algorithm.

**Classification setup and results**

Classification in the heart of any OCR system. Based on the zoning feature map, the results of the artificial neural network and support vector machine are generated and compared with the results calculated from the convolution neural network. The overall summary regarding results for each classifier is discussed below.

**Support vector machine**

The recognition results of the SVM technique are generated based on the feature map calculated in the previous step. The feature map is divided into training set and testing set based on 2:1. In 102 handwritten samples for each individual character in Pashto language, training set includes 68 samples while test set is made of 34 samples. The results of overall accuracy for SVM classifier using the selected training and test set is 56%. The accuracy results of the SVM classifier is depicted in Figure 13.

In Figure 14 the end-nodes represents the Pashto characters and the nodes at the middle represents the accuracies for the corresponding letters. It is concluded from the Figure 14 that slight change in character’s shape confuses the classifier in recognizing the characters accurately. While multi-class problem is another
big barrier to the SVM model, as the Pashto language consists of 44 characters (44 different classes).

**Artificial neural network**

In the proposed research work, the back propagation neural network is trained based on multi-layer feed forward network. To assign the input provided to the ANN model in a batch file format as the proposed problem consists of a huge featured matrix (64 × 4488). Every individual handwritten Pashto character sample consists of 64 features that are composed in a feature matrix. The results obtained for Pashto handwritten character by Artificial Neural Network (ANN) is shown in Figure 14.

Figure 14 includes the problem of over-fitting which is necessary to be minimized as possible for increasing the accuracy of the proposed PHCR system. The PHCR system is validated for varying training and test sets to identify the time consumed and the recognition rate achieved for each variation. From the experiments, it is concluded that training size has positively influenced the classifier accuracy and derive better generalization as shown in Figure 15.

The ANN classifier efficiency in terms of recognition abilities is calculated based on varying training and test set. Also, the proposed PHCR system is tested for varying training and test sizes based on epoch size. The time consumed and accuracy generated based on epoch size and training and test sets are depicted in Figure 16. From the results, it can be seen that size of the training set has a direct relationship with time consumption and classifier’s accuracy. The same relationship is followed for the variation in the number of epochs.

**Convolution neural network**

In the proposed research work a five layered CNN model is selected consist of three hidden layers and an input and output layer. Handwritten characters images are provided as an input to the first layer. While the second and fourth layers of the model act as a convolution layer with subsampling layer as an alternator, that accepts the pooled feature map as an input. CNN is an automatic feature extractor these layers automatically extract features from the raw/input images that are more and more invariant to local transformation. A fully connected layer (FCL) is the last layer of the CNN model, in our case, this is the fifth layer. This layer decides the output based on neurons selected. The overall result of the convolution neural network is shown in Figure 17 during its performance checking.

The graphs represented in Figure 17 also have over-fitting however, this over fitting is smaller as compare to the ANN results shown in Figure 15.

CNN is also tested for varying training and test sets sizes and the results are generated accordingly that is shown in Figure 18. It is evident from the Figure 18 as the training set size increases, the recognition accuracy of the CNN model also increases.

The recognition ability and time consumption of the CNN classifier is tested based on epochs size for varying training and test sets. The results are shown in Figure 19. It is evident from Figure 19 that as the training set size increases, the time (taken during a simulation process) and accuracy of the classifier increase. Same scenario is repeated for the epoch size (when epoch size increase accuracy and time increases).

After examining the results of each model (SVM, ANN, and CNN), the convolution neural network achieves prominent recognition results. The proposed CNN model is also tested for varying training sets and test sets and generated different time and accuracy charts based on these training and test sets. The results are depicted in Figure 20.

The model loss for the proposed PHCR system based on the CNN model is depicted in Figure 21.

After testing each classifier for the proposed problem it is concluded that CNN provides better results than SVM and ANN classifiers as shown in Figure 22. CNN gives 80.7% accuracy for the proposed PHCR system while SVM and ANN gives 56% and 78% recognition accuracies, respectively.

**Conclusion**

This paper presents an optimal OCR system for the recognition of handwritten Pashto characters. For the proposed research work a medium-sized dataset for handwritten Pashto characters is developed for simulation purposes. SVM, ANN and CNN models are applied to perform the automatic recognition of the Pashto characters. Experimental results have shown the accuracy of 56% for SVM, 78% for ANN and 80.7% for CNN based framework. Mainly the contributions of the proposed research work are the following: Firstly, the dataset will act as a resource for future research work on Pashto handwritten letters. This can further help in preserving the old literature of Pashto
Figure 16. Time consumption and accuracy results of ANN for varying training and test sets based on number of epochs.

Figure 17. CNN performance results for PHCR system.

Figure 18. CNN results for varying training and test sets.
Figure 19. The time consumed and accuracy measured graph for varying training and test sets based on epoch size.

Figure 20. CNN model accuracy and time with training set size.

Figure 21. CNN model loss for PHCR system.
history and culture. Secondly, the results obtained will provide a baseline accuracy for future research work on Pashto text processing. Third, since Pashto has encapsulated characters from Persian, Arabic, and Urdu languages so the OCR system developed for the Pashto language will ultimately work in recognizing the characters in these languages.

In the future, we want to test our algorithm for a larger database of handwritten Pashto letters. In addition to that, online recognition of connected words in Pashto will be our main focus. Also, we will extend the proposed research work to handwritten Pashto words recognition. The use of advanced machine learning techniques, such as deep neural networks with a higher number of hidden layers can result in robust recognition. The use of advanced machine learning techniques, such as deep neural networks with a higher number of hidden layers can result in robust recognition.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research work is supported by Department of Computer Science, University of Swabi, Swabi Pakistan.

**ORCID iDs**

Sulaiman Khan https://orcid.org/0000-0003-2562-3477

Shah Nazir https://orcid.org/0000-0003-0126-9944

**References**

1. S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, “KNN and ANN-based recognition of handwritten pashto letters using zoning features,” International Journal of Advanced Computer Science and Applications, vol. 9, no. 10, pp. 570–577, 2018.
2. Sahloul A and Suen C. OFF-line system for the recognition of handwritten arabic character. In: Fourth international conference on computer science & information technology. Dubai, UAE, 2014, pp. 227–244.
3. Pal U and Sarkar A. Recognition of printed Urdu script. In: Seventh international conference on document analysis and recognition, 2003. Proceedings. Edinburgh, UK, 2003, pp. 1183–1187.
4. Kopparapu SK and VL L. On-line handwritten devanagari character recognition using fuzzy directional features. arXiv preprint arXiv:1504.01488, 2015.
5. Ahmad N, Khan AA, Abid SAR, et al. Pashto isolated character recognition using K-NN classifier. SURJ (Science Series) 2013; 45: 679–682.
6. Ahmad R, Naz S, Afzal MZ, et al. Robust optical recognition of cursive Pashto script using scale, rotation and location invariant approach. PloS one 2015; 10: e0133648.
7. Salimi H and Giveki D. Farsi/Arabic handwritten digit recognition based on ensemble of SVD classifiers and reliable multi-phase PSO combination rule. Int J Doc Anal Recognit (IJDAR), 2013; 16: 371–386.
8. Salouan R, Safi S and Boukhaleme B. Handwritten Arabic characters recognition using methods based on Racah, Gegenbauer, Hahn, Tchebychev and Orthogonal Fourier-Mellin moments. Int J Adv Sci Technol 2015; 78, 13–28.
9. Keyvanpour MR, Azmi R, Tabatabai ZSM, et al. Handwriting Persian character recognition using optimized structural elements. Global J Information Technol 2015; 4(2): 107–113.
10. Kamranian Z, Monadjemi SA and Nematbakhsh N. A novel free format Persian/Arabic handwritten zip code recognition system. Comput Electr Eng 2013; 39: 1970–1979.
11. Impedovo D and Pirlo G. Zoning methods for handwritten character recognition: a survey. Pattern Recognit 2014; 47: 969–981.
12. Ayyaz MN, Javed I and Mahmood W. Handwritten character recognition using multiclass svm classification with hybrid feature extraction. Pak J Eng Appl Sci 2016; 10: 57–67.
13. Ali H, Ullah A, Iqbal T, et al. Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network. SN Appl Sci 2020; 2: 152.
14. Memon J, Sami M and Khan RA. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). arXiv preprint arXiv:2001.00139, 2020.
15. Carbune V, Gonnet P, Deselaers T, et al. Fast multilanguage LSTM-based online handwriting recognition. Int J Doc Anal Recognit 2020; 23: 1–14.
16. Habib S, Shukla MK and Kapoor R. Preprocessing for identification of degraded Urdu and Devanagari printed script. In: Computational intelligence in pattern recognition. Springer, 2020, pp. 519–528. Singapore: Springer.
17. Phokharatkul P, Sankhuangaw K, Somkuarnpanit S, et al. Off-line hand written Thai character recognition using ant-miner algorithm. Trans ENFORMATIKA Syst Sci Eng 2005; 8: 276–281.
18. Negi A, Shanker KN and Chereddi CK. Localization, extraction and recognition of text in Telugu document images. In: Proceedings of the seventh international conference on document analysis and recognition-volume 2. Edinburgh, UK, 2003, p. 1193.
19. Cha S-H, Tappert CC and Srihari SN. Optimizing binary feature vector similarity measure using genetic algorithm and handwritten character recognition. In: *ICDAR*, Edinburgh, UK, 2003, pp. 662–665.

20. Kimura F and Shridhar M. Handwritten numerical recognition based on multiple algorithms. *Pattern Recognit* 1991; 24: 969–983.

21. Liu C-L, Eim I-J and Kim JH. High accuracy handwritten Chinese character recognition by improved feature matching method. In: *Proceedings of the Fourth International Conference on Document Analysis and Recognition*. Ulm, Germany, 1997, pp. 1033–1037.

22. Sharma D and Gupta D. Isolated handwritten digit recognition using adaptive unsupervised incremental learning technique. *Int J Comput Appl* 2010; 7: 27–33.

23. Rajashekararadhya S, Ranjan PV and Aradhya VM. Isolated handwritten Kannada and Tamil numeral recognition: A novel approach. In: *2008 First international conference on emerging trends in engineering and technology*, Nagpur, Maharashtra, 2008.

24. Vamvakas G, Gatos B, Petridis S, et al. An efficient feature extraction and dimensionality reduction scheme for isolated greek handwritten character recognition. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Parana, Brazil, 2007, pp. 1073–1077.

25. Roy PP, Pal U, Lladós J, et al. Multi-oriented and multi-sized touching character segmentation using dynamic programming. In: *2009 10th international conference on document analysis and recognition*. Barcelona, Spain, 2009, pp. 11–15.

26. Roy PP, Pal U, Lladós J, et al. Convex hull based approach for multi-oriented character recognition from graphical documents. In: *2008 19th international conference on pattern recognition*. Tampa, FL, USA, 2008, pp. 1–4.

27. W. Contributors. Pashto. In *Wikipedia, The Free Encyclopedia*. Retrieved 17:05, February 14, 2020, (2018, May 17).

28. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.

29. Pradeep J, Srinivasan E and Himavathi S. Diagonal based feature extraction for handwritten character recognition system using neural network. In: *2011 3rd international conference on electronics computer technology*, Kanyakumari, India, 2011, pp. 364–368.

30. Carleo G and Troyer M. Solving the quantum many-body problem with artificial neural networks. *Sci* 2017; 355: 602–606.

31. Zhao H, Zhou D and Wu Z. Snt product character recognition based on bp neural network. In: *2010 sixth international conference on natural computation*. Yantai, China, 2010, pp. 589–593.

32. Ranzato MA, Huang FJ, Boureau Y-L, et al. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In: *2007 IEEE conference on computer vision and pattern recognition*. Minneapolis, USA, 2007, pp. 1–8.

33. Naz S, Umar AI, Ahmad R, et al. Urdu Nastaliq recognition using convolutional–recursive deep learning. *Neurocomputing* 2017; 243: 80–87.

34. Naz S, Umar AI, Ahmad R, et al. Urdu Nasta’liq text recognition system based on multi-dimensional recurrent neural network and statistical features. *Neural Comput Appl* 2017; 28: 219–231.

35. Suen CY, Guo J and Li Z-C. Analysis and recognition of alphanumeric handprints by parts. *IEEE Trans Syst, Man, and Cybern* 1994; 24: 614–631.

36. Morita M, Sabourin R, Bortolozzi F, et al. Segmentation and recognition of handwritten dates: an HMM-MLP hybrid approach. *Doc Anal Recognit* 2003; 6: 248–262.

37. Blumenstein M, Verma B and Basli H. A novel feature extraction technique for the recognition of segmented handwritten characters. In: *Seventh international conference on document analysis and recognition*, Edinburgh, UK, 2003, pp. 137–141.

38. Koechich AL and Kalva PR. Unconstrained handwritten character recognition using metaclasses of characters. In: *IEEE international conference on image processing 2005*, Genova, Italy, 2005, pp. II–542.

39. Verma B, Lu J, Ghosh M, et al. A feature extraction technique for online handwriting recognition. In: *2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541)*. Budapest, Hungary, 2004, pp. 1337–1341.

40. Baptista G and Kulkarni K. A high accuracy algorithm for recognition of handwritten numerals. *Pattern Recognit* 1988; 21: 287–291.

41. Singh S and Hewitt M. Cursive digit and character recognition in CEDAR database. In: *Proceedings 15th international conference on pattern recognition*. *ICPR-2000*, Barcelona, Spain, 2000, pp. 569–572.

42. Imani M and Braga-Neto UM. Maximum-likelihood adaptive filter for partially observed boolean dynamical systems. *IEEE Trans Signal Processing* 2016; 65: 359–371.