Machine Learning & Image Processing for handwritten digits and alphabets recognition from document image through MATLAB simulation

Satrughan Kumar Singh¹ and Jainath Yadav ²
¹CSIR - Central Institute of Mining & Fuel Research, Dhanbad – 826015, Jharkhand, India
²Central University of South Bihar, Gaya-824236, Bihar, India
satrughanksingh@yahoo.com, jainath@cub.ac.in

Abstract. Machine learning is currently playing a vital role in the next generation of the computer world. Automatic pattern recognition has become an important issue of image processing and machine learning. Handwritten digits and alphabets are not arranged in the same size, thickness, position and right direction. Therefore, to determine the issue of handwritten numerals and alphabet recognition, different classifications and complexity should be analyzed. The composition styles of different individuals affect mainly the patterns of alphabets and digits. An effective strategy is to understand the numbers and alphabets transferred from a document image and to make an orderly pattern. In this research work, a soft computing system has been developed using the MATLAB programming. This system uses machine learning algorithms that identify patterns by computerized estimation by identifying handwritten digits and alphabets from a document image. From the experimental results, we observed 96.24% average recognition accuracy of our proposed system.

Keywords- Image processing; machine learning; MATLAB; document image; SVM; digit & alphabet recognition; image segmentation; soft computing;

1. INTRODUCTION

Image processing and computer vision have become a key component of an emerging technology of machine learning. It has become a unique and innovative integration of aspects of machine learning with image feature extraction in image processing. Object detection and image segmentation are used in image processing and computer vision to identify well-defined patterns. The Support Vector Machine (SVM) algorithm facilitates image classification with segmentation, and it removes noise from the image and constructs an optimal hyperplane for separating different classes that are processed by the multi-dimensional phase.

The pattern recognition of handwritten numerals and alphabets is a complex task due to uniqueness and variation in the writing styles of handwriting. When we scan the handwritten digits and alphabets in the document image, a complex and large amount of noise is created which complicates the pattern recognition of the handwritten digits and alphabets. Pattern recognition has been a frontline research field in the field of human-machine interface for the last few decades. In present times, people are constantly trying to make computers intelligent so that they can do almost all the work easily like humans. This intelligent computer can not only reduce human effort but also save time. Classifying patterns from a documented image is one of the major implementations of image processing and machine learning. Offline handwritten alphabet or digit recognition is a unique process of characterization and extraction of alphabets and numbers through segmentation by automated systems.
2. LITERATURE SURVEY

The recognition accuracy of the digits and alphabets depends on the sensitivity of the selected features and the classifiers. Hence, several feature extraction and classification methods can be found in the literature. The recognition accuracy depends on the sensitivity of the extracted features using the SVM classifier. There are several numbers of feature selection and classification methods that were already performed on numeral & alphabet analysis and pattern recognition. The following research works performing the evolution of digit & character recognition for pattern recognition are shown in Table 1.

| Ref. No. | Author(s)       | Model / Classifier | Accuracy (%) | Digit | Capital Alphabet | Small Alphabet |
|---------|----------------|--------------------|--------------|-------|-----------------|----------------|
| [1]     | Gupta, A. et al.| NN & SVM           | 62.93        | -     | -               | -              |
| [2]     | Bellili, A. et al.| MLP-SVM         | 98.01        | -     | -               | -              |
| [3]     | Nasien, D. et al.| SVM               | -            | 88.46 | 86.00           | -              |
| [4]     | Priya et al. | NN & SVM           | 98.4         | -     | -               | -              |
| [5]     | Vamvakas, G. et al.| SVM         | -            | 80.19 | -               | -              |
| [6]     | Gattal, A. et al.| SVM               | 95.21        | -     | -               | -              |
| [7]     | Neves, R.F.P. et al.| SVM         | 97.94        | -     | -               | -              |
| [8]     | Kadam, D. et al.| LBP-SVM           | -            | 96.5  | 98.00           | -              |
| [9]     | Khedidja, D. et al.| SVM           | 99.89        | -     | -               | -              |
| [10]    | Mishra, A. et al.| SVM               | 96.29        | -     | -               | -              |

3. METHODOLOGY

The handwritten digits and alphabets recognition systems incorporate digitization, pre-processing, segmentation, attribute selection and tracing, training datasets, validation of datasets, testing datasets, and attribute validation steps for pattern recognition. The block diagram of the proposed model for the pattern recognition of the handwritten digit or alphabet is shown in Figure 1.
In this work, samples of different handwritten numerals and alphabets of 8 authors have been collected. Figure 2 shows a handwritten dataset of 8 authors (S1, S2… S8). The proposed method has been discussed in the following sub-sections.

3.1. Document digitalization

Document digitization is the normal process of converting handwritten and other documents into electronic and digital forms. Electronic and digital conversion is used to scan a document that is an electronic representation of the original document in the form of an image of a document or image file. Digitization is an initial step that executes before the pre-processing step. The digitization phase produces a digital image of the original document which is processed in the pre-processing phase.

3.2. Pre-processing

In the pre-processing phase, the unimportant data information is removed from the digital image of the original document which can adversely affect the recognition accuracy. This step can include steps such as binarization, noise removal, skew detection, and skeletalization that can help in completing correctly the appropriate algorithms for the decimation system and devising precise strategies. The major role of the pre-processing is to filter out the impurities from the image and also to perform smoothing and normalization[11]. Pre-processing can produce a noise-free feature as well as an appropriate image set of alphabets and numbers for effective selection. The following are the various steps of pre-processing.

3.2.1. Binarization: The binarization is the stepwise process of converting a gray-scale image into a binary image which is shown as 0 and 1. In the binary image, the visible part of the image can be represented by digit 1 and the invisible part by digit 0 as shown in Figure 3. Generally, the scanned
image of the document does not align correctly in the horizontal direction, so we need to adjust it by skew angle correction. This can reduce its ability to convert gray-scale images into binary images that are necessary for reducing dimensions to increase processing speed.

![Figure 3. The binarization process for the handwritten symbol recognition](image)

3.2.2. Noise elimination: Noise elimination is used to remove any undesirable or meaningless bit patterns and noise from binary images. Noise is an undesired disturbance that creates unwanted errors. Such interference would have to be eliminated for maintaining sufficient accuracy.

3.2.3. Skew detection: Generally, the digital image of the document does not fit directly, so the skew detection process is required for the recognition of the document's digital image.

3.2.4. Skeleton: Typically there is opposition to the width of the line from several pixels to one pixel wide. In this case, the skeletalization process can refer to reducing line width by removing many discrepancies, to simplify the classification algorithm and at the same time reduce processing time.

3.3. Segmentation
Segmentation is an impartment process of pattern recognition of any digit and alphabet symbol that determines the components of the document image. Image segmentation is a very important problem for estimating and acquiring information from an image[12]. It is essential to locate the areas of a document image where the data is printed and their figures and graphics are different. In image segmentation, the image is divided into several square pixels[12]. The segmentation technique identifies the area of interest of a picture by using criteria such as color, composition, and excess[12]. It identifies the end and appearance of the image as it appears on the computer screen[12].

3.4. Feature selection or extraction
The feature selection method can be used to choose important features and to remove inconsistencies and redundancies from these pre-processed image datasets. Diagonal features play a vital role to achieve higher accuracy of the pattern recognition system. The attribute linking shows the applied information of the shape enclosed in the pattern to classify the images to simplify the validation technique. Handwritten characters include many specific elements such as diagonal component extraction, chain code, scale-invariant, etc.

3.5. Training dataset
In this phase, the training algorithm is used to train the classifier using the input dataset to fit the model parameters such as model weights. The model is trained on training supervisors, who use a supervised learning method. The training dataset usually consists of a pair of input and related output vectors, where
the answer key is shown as the target. The current model has to extract a result by running it with the training dataset, and the model's parameters can be adjusted based on the result.

3.6. Validation of dataset
In dataset validation, the model that fits on the training dataset can be objectively evaluated by tuning the model. Validation of datasets can be used to routinely stop training when errors escalate due to overfitting of the training dataset. This process becomes complicated by the error fluctuations of the validation of the dataset during training that generates multiple local minima.

3.7. Testing dataset
It evaluates the model developed from training dataset. If the data in the test dataset has never been used in training, it can be holdout under cross-validation to be re-trained. If a model fits into the training dataset, it can be fully fitted to the test dataset with minimal overfitting. Generally, better fitting of training datasets indicates overfitting as opposed to testing datasets. The test set is therefore an example of making a proper assessment based on the performance of a fully specified classifier.

4. SUPPORT VECTOR MACHINE
Support vector machine (SVM) constructs an optimal hyperplane (multiplane) in multi-dimensional space separating different classes, which can be used to minimize an error. The main task of SVM during the model learning phase is to find the optimal hyperplane with maximum margin that can divide the dataset into classes and also estimate the width between classes. The data points which are closest to the hyperplane is called support vectors. These support vectors can better define a separate line by calculating the point margin that may be most relevant for constructing digit and alphabet classifiers. Figure 4 shows the maximum margin and optimal hyperplane between classes.

![Figure 4. Maximum margin and optimal hyperplane between classes](image)

5. RESULT AND DISCUSSION
In this work, we have used the confusion matrix terminology to evaluate the performance of the proposed method. The confusion matrix is sometimes called the error matrix, which can be used as a matrix representation on a set of test data in a problem of machine learning and statistical classification. Usually, one class is confused with another which can allow easy identification of confusion between classes. In collimation, the confusion matrix shows how the classification model is confounded when it makes predictions and then provides insight into the errors made by the classifier based on these predictions.

In our study, we have collected the scanned symbols written by eight persons. For each person, we have collected 10 instances of each symbol. In our database, the total numbers of symbols are 4960 (62 symbols × 8 persons × 10 instances). The total numbers of symbol instances used for training
and testing are 3472 and 1488 respectively. The recognition accuracy of the proposed method is shown in Table 2. From the table, we observe that the average recognition accuracy of capital alphabets, small alphabets and digits are 95.37%, 96.27% and 97.08% respectively. Among the capital alphabets, ‘U’ symbol has highest recognition accuracy and ‘B’ symbol has lowest recognition accuracy. Among small alphabet, ‘k’ and ‘p’ have highest and lowest recognition accuracy respectively. Similarly, we observed highest and lowest recognition accuracy for digits ‘7’ and ‘4’ among the digits respectively. It can be observed that for the specific pattern, the proposed method recognized all the alphabets & numerals. We have developed a technique for recognition of scanned or handwritten alphabets and digits by density method and potential application. This experimental work on handwritten pattern recognition uses sequential rules to segment alphabets and digits from document images for classification using the SVM classifier model. The classifier also describes the pattern recognition results to improve classification performance and error evaluation after the pre-processing stage. The graphical interface of the handwritten pattern recognition has been illustrated through Figure 5.

![Figure 5. GUI Screenshot of handwritten pattern recognition](image-url)

### Table 2. Accuracy of handwritten digits & alphabets of authors (S1-S8) using proposed method

| Symbol | Accuracy (in %) | Average Accuracy (in %) |
|--------|-----------------|-------------------------|
|        | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 |        |
| A      | 99.50 | 98.57 | 98.57 | 98.07 | 96.82 | 96.57 | 96.82 | 99.07 | 98.00 |
| B      | 87.07 | 89.70 | 89.70 | 89.20 | 87.95 | 87.70 | 87.95 | 90.20 | 88.68 |
| C      | 92.26 | 92.79 | 92.79 | 92.29 | 91.04 | 90.79 | 91.04 | 93.29 | 92.04 |
| D      | 97.12 | 97.78 | 97.78 | 97.28 | 96.03 | 95.78 | 96.03 | 98.28 | 97.01 |
| E      | 98.84 | 93.17 | 93.17 | 92.67 | 91.42 | 91.17 | 91.17 | 93.67 | 93.22 |
| F      | 98.46 | 97.79 | 97.79 | 97.29 | 96.04 | 95.79 | 96.29 | 98.29 | 97.22 |
| G      | 98.86 | 95.51 | 95.51 | 95.01 | 93.76 | 93.51 | 93.51 | 94.01 | 95.27 |
| H      | 87.43 | 93.95 | 93.95 | 93.45 | 92.20 | 91.95 | 92.70 | 94.45 | 92.51 |
| I      | 99.29 | 98.98 | 98.98 | 98.48 | 97.23 | 96.98 | 97.23 | 99.48 | 98.33 |
| J      | 98.66 | 97.36 | 97.36 | 96.86 | 95.61 | 95.36 | 95.86 | 97.86 | 96.87 |
| K      | 94.65 | 96.21 | 96.21 | 95.71 | 94.46 | 94.21 | 94.71 | 96.71 | 95.36 |
| L      | 96.90 | 96.40 | 96.40 | 95.90 | 94.65 | 94.40 | 94.90 | 96.90 | 95.81 |
| M      | 91.20 | 90.70 | 90.70 | 90.20 | 88.95 | 88.70 | 89.45 | 91.20 | 90.14 |
| N      | 97.76 | 97.26 | 97.26 | 96.76 | 95.51 | 95.26 | 95.76 | 97.76 | 96.67 |
| O      | 98.93 | 98.43 | 98.43 | 97.93 | 96.68 | 96.43 | 96.93 | 98.93 | 97.84 |
| P      | 93.20 | 92.70 | 92.70 | 92.20 | 90.95 | 90.70 | 90.95 | 93.20 | 92.08 |
| Q      | 94.26 | 93.76 | 93.76 | 93.26 | 92.01 | 91.76 | 92.51 | 94.26 | 93.20 |
| R      | 98.10 | 97.60 | 97.60 | 97.10 | 95.85 | 95.60 | 95.85 | 98.10 | 96.98 |
| S      | 97.73 | 97.23 | 97.23 | 96.73 | 95.48 | 95.23 | 95.73 | 97.73 | 96.64 |
This research work deals with the recognition of handwritten numerals and alphabets by applying the support vector machine technique. However, many variations of the same alphabet or number, with different text styles and sizes for recognition, become complex and difficult. Therefore, the SVM algorithm can be used for optimal pattern recognition from several attributes. In this task, offline handwritten numeral and alphabet recognition is used for training and managing the dataset, which is obtained manually or from a scanned document of handwritten numerals and alphabets. The main work focuses on classifying digits and alphabets in a digitally desirable format so that they can be easily modified and processed by the machine intelligence system. The proposed research work provides more efficient and accurate results that obtained an overall 97-99% recognition rate. Therefore, SVM works well with a clear & best margin of separation and high dimensional space. The recognition rate needs to be tested by increasing size of datasets for future work.

6. CONCLUSION

This research work deals with the recognition of handwritten numerals and alphabets by applying the support vector machine technique. However, many variations of the same alphabet or number, with different text styles and sizes for recognition, become complex and difficult. Therefore, the SVM algorithm can be used for optimal pattern recognition from several attributes. In this task, offline handwritten numeral and alphabet recognition is used for training and managing the dataset, which is obtained manually or from a scanned document of handwritten numerals and alphabets. The main work focuses on classifying digits and alphabets in a digitally desirable format so that they can be easily modified and processed by the machine intelligence system. The proposed research work provides more efficient and accurate results that obtained an overall 97-99% recognition rate. Therefore, SVM works well with a clear & best margin of separation and high dimensional space. The recognition rate needs to be tested by increasing size of datasets for future work.
References

[1] Gupta, A., Srivastava, M., Mahanta, C.: Offline handwritten character recognition using neural network. In: International Conference on Computer Applications and Industrial Electronics (ICCAIE), pp. 102-107. 2011 IEEE International Conference on IEEE (2011)

[2] Bellili, A. et al.: An hybrid MLP-SVM handwritten digit recognizer. In: Conference Proceedings of Sixth International Conference on Document Analysis and Recognition, vol. 1, pp. 0028. (2001)

[3] Nasien, D., Haron, H., Yuhaniz., S.S.: Support Vector Machine (SVM) for English handwritten character recognition. In: 2010 Second International Conference on Computer Engineering and Applications, (2010)

[4] Priya et al.: Review on handwritten digit recognition. In: International Journal of Novel Research and Development, vol. 2, Issue 4, pp. 91-94. April (2017)

[5] Vamvakas, G., Gatos, B., Perantonis, S.J.: A Novel Feature Extraction and Classification Methodology for the Recognition of Historical Documents. In: proceeding of IEEE 10th International Conference on Document Analysis and Recognition, pp. 491-495. (2009)

[6] Gattal, A. et al.: Isolated Handwritten Digit Recognition Using oBIFs and Background Features. In: Conference Proceedings 12th IAPR Workshop on Document Analysis Systems (DAS), vol. 1, pp. 305-310. (2016)

[7] Neves, R.F.P. et al.: A SVM Based Off-Line Handwritten Digit Recognizer. In: International conference on Systems, Man and Cybernetics, pp. 510-515. IEEE Xplore, (2011)

[8] Kadam, D., Chavan, P., Pandhara, P.: Literature Survey on Recognition and Evaluation of Optical Character Recognition (OCR). In: International Journal of Scientific & Engineering Research, vol. 9, Issue 2, pp. 72-75. February (2018)

[9] Khedidia, D., Hayet, M.: Multiple Classifiers and Invariant Features Extraction for Digit Recognition. In: International Journal of Computer Electrical Engineering, vol. 11, Number 1, pp. 41-52. March (2019)

[10] Mishra, A., Singh, D.: Handwritten digit recognition using combined feature extraction technique and neural network. In: Computer Modelling & New Technologies, vol. 21, issue 2, pp. 80-88.

[11] Sivaraman, P “A New Method of Maximum Power Point Tracking for Maximizing the Power Generation from a SPV Plant” Journal of scientific and Industrial Research Vol.74, No.3 pp.411 - 415 AUG 2015.

[12] Senthil kumar J Charles Raja S Dipti S Venkatesh P 2018 Hybrid Renewable Energy based Distribution System for Seasonal Load Variations International Journal of Energy Research 42 3 1066 - 1087 2018

[13] Kadam, D. et al.: Literature Survey on Recognition and Evaluation of Optical Character Recognition (OCR). In: International Journal of Scientific & Engineering Research, vol. 9, Issue 2, pp. 72-75. February (2018)

[14] Veerakumar Nnirmalkumar Sathishkumar Sathishkumar Rajesh Novel harmonic elimination technique for cascaded h-bridge inverter using sampled reference frame, Journal of Theoretical and Applied Information Technology Vol. 58 No.2 2013