

A NOVEL METHOD FOR HANDWRITTEN MATHEMATICAL DOCUMENT BASED ON EQUATION SYMBOLS RECOGNITION USING K-NN AND ANN CLASSIFIERS

Ratnamala S.Patil¹, Shilpa²

PG Student, Dept of Digital Communication and Networking, GECW kalaburgi, Karnataka, India.
Asst .Professor, Dept of computer science and Engineering, AIET, Kalaburgi, Karnataka, India

Abstract—: Document image classification has been one of the most widely used areas of research in image processing. Various document classification techniques are being proposed in the past. Handwritten document analysis and classification is another challenging areas in document image analysis. Handwriting recognition, handwritten text OCR, signature based document classification area some of wide research areas in this paper. Mathematical document identification is a unique challenge in document image analysis that deals with first identifying mathematical symbols in a document and then classifying the document as maths and non-maths document based on density of the mathematical symbols. In this work we have proposed a novel maths document classification based on statistical features and K-NN and A-NN classifiers.

Index Terms—Document image processing, Neural Network, Statistical feature, LBP feature.

I. INTRODUCTION

The work on handwritten math symbol recognition has begun many years ago but came into light few years ago. Much of research is going in this field but most of the work is related to online. Here in this work we concentrate more on offline features. Handwritten math symbols differs with handwritten character recognition in many aspects. The responsibility of our work is to identify the documents as maths document or non maths document. In the recognition of character, signature, maths symbol etc the basic steps are preprocessing, segmentation, feature extraction, training and finally the classification. The same previously said steps are followed in our work. In our work we extracted Statistical features and LBP features. ANN and K-NN classifiers are used for classification. This is the step towards automated image document indexing where by documents are classified or searched based on things like math symbol, handwriting etc. It is a process towards by means of which automated document classification, document indexing has been carried out.

This paper [1] presents a recognition system based on HMM. Preprocessing is done to remove the noise and unwanted

things which will be helpful for further process. Totally 25 features were extracted, out of which only 5 good features were used for classification. An HMM classifier is used. Model selection, initialization is done using K-Means and Baum-Welch algorithm is used for training, therefore the recognition rate achieved is very good. This paper [2] presents the handwritten English alphabet recognition system which uses neural network for recognition. The documents are scanned and then the scanned images are preprocessed to remove all the noise and used for further processing. For recognition process the neural network is used, which contains two hidden layers. System here achieves the accuracy of 82.5% and it shows poor results for similar patterns. This paper [3] presents recognition system for English alphanumeric character recognition. Basically preprocessing is done which comprises of digitization, detection of bounding boxes etc. this system can detect 0-9 and A-Z with accuracy of 99.99% and 98% respectively and alphanumeric more than 94%. This system is independent of size and color of the character to recognize. This work [4] represents the character recognition using neural networks which uses the back propagation algorithm. Preprocessing of the images is done which includes size normalization, binarization, smoothing, edge detection, segmentation. Features are extracted which gives the perfect information for recognition. Using neural network as a classifier accuracy achieved is 85%. The one main drawback of this system is that it doesn't work well for cursive writing. This paper [5] presents the recognition system of handwritten digits using Artificial Neural Network (ANN). Many of the previous work shows that, using the large data sets results in more training time. So this work solves the problem by reducing training period by using GPU for implementation. The back propagation algorithm is used with parallelization technique which efficiently reduces the training time by using MNIST dataset. Thus they achieved the recognition rate of 98% on the test data.

In this paper we present a novel method for recognition of mathematical document by using K-NN and ANN. Statistical and LBP features are extracted.

II. OBJECTIVE

Objective of this work is to extract probable equation area from a mathematical document image, classify the block as equation or non equation and then further classify the document image as mathematical document or not based on the density of the maths symbols present.

III. METHODOLOGY

The overall technique is presented as below:

- 1) Read document image
- 2) Preprocess and remove noise
- 3) Segment characters in the document.
- 4) Extract mean and standard deviation features from the characters.
- 5) Manually store the characters into two different sets: Maths and Non-maths
- 6) Train KNN and ANN Classifiers using these features with Maths and Non-Maths
- 7) Give a test document as input.
- 8) Follow steps 2 to 4.
- 9) Classify the characters using both KNN and ANN.
- 10) Aggregate the result and calculate Maths density as:
 $Maths\ Density = \frac{No\ of\ Maths\ Symbols}{Total\ Symbols}$
- 11) Threshold Maths Density.

if Maths density > .3

Classify Document as Maths

else

Classify Document as Non Maths

The work on recognition of math symbol recognition has following steps : preprocessing, feature extraction and classification. They are elaborated in detail in following subsections.

A. Preprocessing

Operation by which the image data is improved. Filtering removes the noise and improves through filtering and ROI selection. Filtering removes the noise and improves the features of image. The resulting image after the preprocessing step, varies from the source image in terms of appearance, size. The output image will not be same as before but it has some characteristics of the earlier image. The output image is better suited for feature extraction, image feature detection and classification.

B. Need Of Preprocessing

Processing is very important because, to get the efficient characteristics from the image for the further process like feature extraction, image feature detection and classification. Without preprocessing this is not possible. Following examples shows why we go for preprocessing

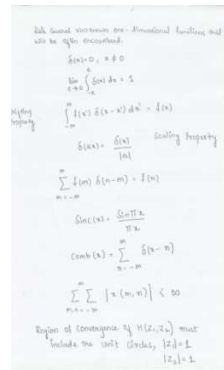


Figure 1(a): Source image

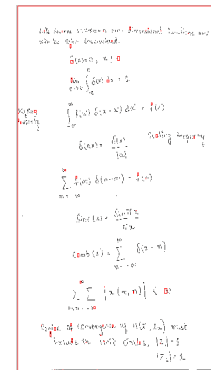


Figure 1(b): Image without preprocessing

As we can in the above image that the source image is converted to binary but the detected terms in resulting image are not so clear and not well suited for further process. So we go for preprocessing.

c. After preprocessing

After preprocessing the resulting image we get is binary image which is free from noise and is perfect image. We can locate the each and every character perfectly in the image and is very useful for the further process of our work.



Figure 2(a): Source image

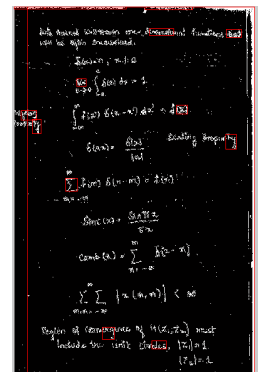


Figure 2(b): After preprocessing

D. Block Diagram Of Preprocessing

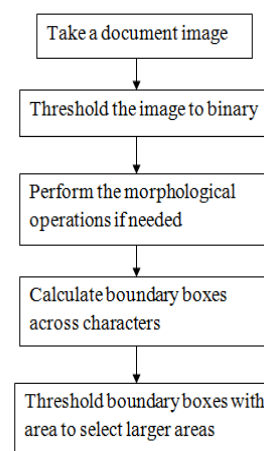


Figure 3 : Block Diagram Of Preprocessing Method

For the preprocessing of a document image, block diagram is as shown. The very first step of preprocessing is that, the document image is taken. There is no need of resizing the document image because the resizing changes the font. Here the preliminary requirement is to extract the text and remove other things. Most of document images will be of white with fonts as black.

Thresholding is done using point operation, wherein point operation is done on each and every pixel of the image. If the threshold value satisfies the condition than that pixel will be '1' otherwise It will be 0 and the image inversing is done i.e where ever the font is there the color will be white and other will be black.

Segmentation is multiplying the original image with threshold image. This threshold image is also called mask image. Problem with the segmentation is that there are more anomalies in the text. In document processing, the binary image is much more prominent than the non binary image. so we eliminate the segmented image. The resulting image will be the image where we have to find out the equations. So, to find equation morphological operation is done.

The bounding boxes are calculated using region of properties [ROI]. Region of properties identifies the block of image and returns a bounding box corresponding to every bounding boxes present, thresholding is done to select the larger area for identification of equations. The text converted within this larger area are probable equation parts.

IV. FEATURE EXTRACTION AND CLASSIFICATION

In this work the structural features and LBP features are extracted by taking region of properties. If we divide the image into 256 blocks the binary pattern of those 256 blocks represents pattern. Local binary pattern first divides the image into 256 grids and then finds out the binary pattern inside the grid

For classification purpose K-NN and ANN is used.

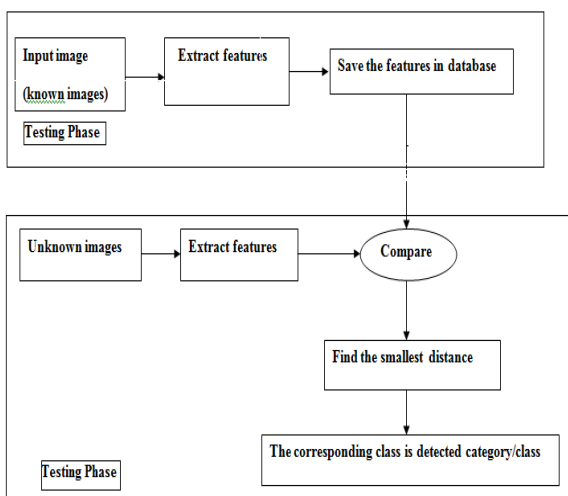


Figure4 : Block Diagram Recognition System

The basic block diagram of our system mainly has two blocks i.e Training and Testing. The input image is basically a scanned image and is preprocessed to remove noise, to convert RGB to gray image. Features are extracted. In training we take known images and extract each and every characters present in the document and store them in a database. All the extracted characters are trained and the classes are built. An unknown image is taken for testing, features are extracted and then the features of training data as well as test data are compared. After extracting the each and every characters and symbols from the document image, the maths and non maths symbols are separated manually by the user. For this, the user should have a prior knowledge about which are maths and non maths symbols.

In this work we used Artificial Neural Network (ANN) and K-Nearest Neighbour (K-NN) classifiers for recognition purpose. While training Neural Network, as there are two types of features we need to tell a Neural Network with what kind of features we are going to train a Neural Network. We first randomly construct a feed forward Neural Network of 120 hidden layers, then we train Neural Network with all our feature vector. Neural Network accepts feature vectors to be in transpose order. K-Nearest Neighbor is a distance based classifier which checks the input vector distance from maths symbols as well as non maths symbol and based on distance it will be able to tell whether a document is maths document or non maths document.

V. RESULTS AND DISSCUSION

Table 1:Document Detection Analysis

Input image	Output image	Document type	Detected document type			
			K-NN Classifier		ANN classifier	
			Statistical feature	LBP feature	Statistical feature	LBP feature
		Maths	Maths	Maths	Maths	Maths
		Non Maths	Non Maths	Non Maths	Non Maths	Non Maths
		Maths	Maths	Maths	Maths	Maths
		Maths	Maths	Maths	Maths	Maths

Table 2: Recognition Rates Obtained by K-NN Classifier





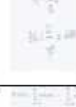


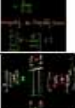
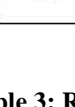











Input document image	Output image	Total number of symbols	Correctly detected		Recognition rate%	
			K-NN classifier		K-NN classifier	
			Statistical feature	LBP feature	Statistical feature	LBP feature
		17	15	15	88.23	88.23
		26	26	25	100	96.15
		8	7	7	87.5	87.5
		34	27	27	79.41	79.41
		9	5	6	55.55	66.66

Table 3: Recognition Rates Obtained by ANN Classifier

Input document image	Output image	Total number of symbols	Correctly detected		Recognition rate%	
			ANN classifier		ANN classifier	
			Statistical feature	LBP feature	Statistical feature	LBP feature
		17	15	16	88.23	94.11
		26	26	26	100	100
		8	7	8	87.5	100
		34	32	33	94.11	97.05
		9	6	7	66.60	77.78

It is observed from the above result analysis. K-NN and ANN works perfectly for the recognition of mathematical document with statistical features and LBP features. We have achieved an approximate recognition rate of 86% and 89% with K-NN and ANN respectively.

For further better results we go with Support Vector Machine (SVM). SVM with MLP and incorporating complex shape descriptors like zernike moments, Curvelet transform .

VI. CONCLUSION

Automatic document classification into Maths and Non-maths document has wide range of applications starting from document searching, archival and so on. Though there have been some past works towards maths symbol recognition, mathematical document classification hasn't attracted too many works. A document contains both mathematical as well as non maths characters and symbols. Therefore classifying such a document is extremely challenging. In this work we have proposed a novel maths document identification using KNN and ANN classifiers and using Statistical features and local binary pattern features. Our results shows that the method gives much better accuracy for ANN and KNN technique. This work can be further improved by improving the SVM with MLP and incorporating complex shape descriptors like zernike moments, Curvelet transform .

REFERENCES

- [1] Lei Hu, Richard Zanibbi, 2011. *HMM-Based Recognition of Online Handwritten Mathematical Symbols Using Segmental K-means Initialization and A Modified Pen-up/down Feature.*
- [2] Yusuf Perwej, Ashish Chaturvedi, 2011. *Neural Networks For Handwritten English Alphabet Recognition.* International Journal Of Computer Applications. Vol 20(7)
- [3] Md Fazlul Kader and Kaushik Deb, 2012. *Neural Network-Based English Alphanumeric Character Recognition.* International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.2, No.4
- [4] Ankit Sharma, Dipti R Chaudhary, 2013. *Character Recognition Using Neural Network.* International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue4- April 2013.
- [5] Suthasinee Iamsa-at, Punyaphol Horata. 2013 *Handwritten Character Recognition Using Histograms of Oriented Gradient Features in deep learning of artificial neural network.* IEEE.
- [6] K.Venkata Reddy, D.Rajeswara Rao, U.Ankaiah, K.Rajesh, 2013. *Handwritten Character And Digit Recognition Using Artificial Neural Networks.* International Journal Of Advanced Research In Computer Science And Software Engineering. Volume 3, Issue 4, April 2013.

[7] Mamatha H.R Karthik S Srikanta Murthy K, 2013.*Classifier Fusion Method to Recognize Handwritten Kannada Numerals*.

[8] Viragkumar N. Jagtap , Shailendra K. Mishra. 2014 . *Fast Efficient Artificial NeuralNetwork For Handwritten Digit Recognition*. International Journal Of Computer Science And Information Technologies, Vol. 5 (2).

[9] Fotini Simistira, Vassilis Papavassiliou, Vassili Katsouros, George Carayannis. 2014 *Recognition Of Spatial Relations In Mathematical Formulas*.IEEE, 14th International Conference On Frontiers In Handwriting Recognition.

[10] Nicolas D. Jimenez,Lan Nguyen. Recognition of Handwritten Mathematical Symbols with PHOG Features.

[11] Francisco Lvaro, Joan-Andreu SaNchez, JoseMiguel Bened. 2014, *Offline Features For Classifying Handwritten Math Symbol Recognition with Recurrent Neural Networks*. IEEE 22nd international conference on pattern recognition.