Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN : 2348-4470
p-ISSN : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3,Issue 5,May -2016

A Review onText Detection and Character Recognition for Printed Document

Aayushi J Solanki,[1],HareshChande[2]

[1]*Computer Engineering Department, HJD Institute of Technical Education and Research,Kera-Kutch,*
[2] *Computer Engineering Department, HJD Institute of Technical Education and Research,Kera-Kutch,*

***Abstract*—**Textrecognitioninimagesisaresearcharea which attemptstodevelopa computersystem withtheabilityto automaticallyread thetextfrom images.Thesedaysthere isa hugedemand in storing theinformation availableinpaper documents formatin toacomputerstoragediskandthen later reusingthisinformationbysearchingprocess.Onesimple wayto store informationfrom thesepaperdocumentsin tocomputer system istofirstscanthedocumentsandthenstorethemas images.Buttoreusethisinformationitisverydifficulttoread theindividualcontentsandsearchingthecontentsformthese documentsline-by-lineand word-by-word.Thechallenges involved in thisthefontcharacteristicsofthecharactersin paper documentsand qualityofimages.Duetothesechallenges, computer isunable torecognize thecharacters while reading them.Thusthereisaneedofcharacterrecognitionmechanisms toperform DocumentImageAnalysiswhichtransforms documentsin paperformattoelectronicformat.In this paperwe havediscussmethodfortextrecognitionfrom images.

***Keywords*-** *pre-processing, segmentation,character recognition,classification*

I. INTRODUCTION

Text Recognition isthemechanicalorelectronic conversionof imagesof typed,handwrittenorprintedtextintomachine encodedtext.Itiswidelyused asaformofdataentryfromprintedpaperdatarecords,whetherpassportdocuments,invoices,bank statements,computerizedreceipts,businesscards,mail,printoutsofstatic data,oranysuitable documentation.Itisacommonmethodofdigitizingprintedtextssothatitcanbeelectronicallyedited, searched,storedmorecompactly,displayedon-line,andusedinmachineprocessessuchasmachine translation,text-to-speech,keydataandtextmining.OCRisafieldofresearchinpatternrecognition, artificialintelligenceandcomputervision.

Text recognitionhasbecome oneof themostsuccessfulapplicationsoftechnology inthefieldof patternrecognition andartificialintelligence.Many commercialsystems for performingOCRexist foravariety ofapplications, althoughthemachinesarestillnotableto compete with human readingcapabilities. Optical CharacterRecognition dealswiththeproblem of recognizingoptically processedcharacters. Now-a-days,thereis growingdemand for the software systems to recognize charactersin computer system when informationisscannedthroughpaperdocuments asweknow thatwehavenumberofnewspapers andbookswhicharein printedformatrelatedtodifferentsubjects.Thesedays thereis ahugedemandin"storingtheinformation availableinthese paperdocuments intoacomputerstoragediskandthenlater reusingthisinformation bysearchingprocess".Onesimple waytostoreinformation inthesepaperdocumentsinto computer systemistofirstscanthedocuments. Wheneverwe scanthedocuments throughthescanner,thedocuments are storedas images formatin thecomputersystem.Theseimages containing textcannotbeeditedbytheuser.

Thereasonforthis difficultyisthefontcharacteristicsofthecharactersinpaper documents aredifferenttofontofthecharactersincomputer system.Asaresult,computer isunabletorecognize the characterswhilereading them.Thisconceptofstoring the contents ofpaperdocuments incomputer storageplace and thenreading andsearching thecontentiscalleddocument processing.Sometimesinthisdocumentprocessingweneed toprocesstheinformation thatisrelatedtolanguages other than the Englishin theworld.Thisprocessis also called DocumentImageAnalysis(DIA).Thus ourneed istodevelop sometextrecognitionalgorithm toperformDocument Image Analysiswhichtransforms documents inpaperformatto electronicformat.

II. LITERATURE SURVEY

A Neural Network BasedHandwritten Meitei MayekAlphabetOpticalCharacterRecognition System [1]

Handwrittencharacter recognition is apart of optical character (OCR) system.OCR canbeappliedtoboth printed textandhandwritten documents. Inthispaperwe discussedthehandwrittencharacter recognition ofMeiteiMayek (Manipuri script). Although OCR has been studied and developedformany Indianscript veryfew works have been reportedsofarforMeitei-Mayek. Thispaperdescribesthe handwritten MeiteiMayek(Manipuriscript)alphabets recognition (HMMAR)usinganeuralnetworkapproach.The alphabetdatabaseispre-processed andtheextracted featureis senttoaneuralnetwork systemfortraining.Thetrainedneural networkisfurthertestedandperformance analysisisobserved.

Theemphasisisgivenontheprocessofcharacter segmentation fromawholedocument i.e.isolatingasinglecharacterimage fromacompletescanneddocument.

Text RecognitionfromImages [2]

Textrecognitioninimagesisaresearcharea which attemptstodevelopa computersystem withtheabilityto automaticallyread thetextfrom images.Thesedaysthere isa hugedemand in storing theinformation availableinpaper documents formatin toacomputerstoragediskandthen later reusingthisinformationbysearchingprocess.Onesimple wayto store informationfrom thesepaperdocumentsin tocomputer system istofirstscanthedocumentsandthenstorethemas images.Buttoreusethisinformationitisverydifficulttoread theindividualcontentsandsearching thecontentsformthese documentsline-by-lineand word-by-word.Thechallenges involved in thisthefontcharacteristicsofthecharactersin paper documentsand qualityofimages.Duetothesechallenges, computer isunable torecognize thecharacters while reading them.Thusthereisaneedofcharacterrecognitionmechanisms toperform DocumentImageAnalysis(DIA)whichtransforms documentsin paperformattoelectronicformat.In this paperwe havediscussmethodfortextrecognitionfrom images.

Recognition of Off-line Hand printed EnglishCharacters, Numerals and Special Symbols [3]

Optical CharacterRecognitioncan improvetheinteractionbetweenman andmachinein variousapplicationsincludingdataentry,office automation, digital library, banking applications, healthinsuranceandtaxformsetc.Muchofworkhasbeendonein the recognitionofmachineprinted charactersin variouslanguages with considerablygoodefficiencies,howevermakingrobust recognitionenginesthatcanbeputto recognizehandwrittenand hand printeddatawith commendablerecognitionratesstill remainsas an activeareaofresearchowingtothechallengeslike diversehumanhandwritingstyle,variationin shape,angleand styleof characters. Takingintoaccountthechallengesand scope forimprovementinthisdomain, theworkofoff-linecharacter recognitionofhand printeddocumentimagescontainingEnglish Characters-Uppercaseand Lowercase,Numeralsand Special Charactershasbeen presented.Statistical,Geometricand DirectionalFeature Extraction techniques have been applied over segmentedcharacterimage.Classificationwas doneusing Multilayer perception neural network (NN) with back propagationandSupportvectormachine(SVM)classifier.

Neural NetworkbasedHandwrittenCharacter Recognitionsystemwithout featureextraction [4]

Handwriting recognition has been one of the activeandchallenging researchareasinthefieldofimage processingandpatternrecognition.It hasnumerous applicationswhichinclude,readingaid forblind,bank chequesand conversionof anyhandwrittendocumentinto structuraltextform.Inthispaper anattempt ismadeto recognizehandwrittencharactersfor Englishalphabets withoutfeatureextraction usingmultilayer FeedForward neuralnetwork.Eachcharacterdataset contains26 alphabets. Fiftydifferent character datasetsareusedfor training theneuralnetwork. Thetrained network isused forclassification andrecognition.

Scanning NeuralNetwork for TextLineRecognition [5]

This paper describessegmentationfree text linerecognitionapproachusing multilayer perceptron(MLP) and hiddenmarkovmodels (HMMs).Aline canningneural networktrainedwith characterlevel contextualinformation andaspecialgarbageclass–isusedtoextractclassprobabilities at every pixel succession.The outputofthis scanningneural networkisdecoded byHMMs toprovide characterlevel recognition.

A Robust Approach for Offline English CharacterRecognition [6]

In thispaper,weproposearecognitionmodelbasedonArtificial NeuralNetwork(ANN)supportedby novelfeatureextraction technique.Hand written datahascontinued to persist asameans ofrecordinginformation in day-to-daylifewith theintroduction of latest technologies. The constant development of computer toolslead to therequirement ofeasierinterface between human and computers.Recognitionofhandwritten charactersby computersiscomplicated taskascomparedtotyped character. Theproposedsystem isbeen implementedusing MATLAB successfully. TheANNacceptstheinputasascannedimage. This input undergoes a sequence ofpre-processingsteps; binarization andnormalization.Then featuresareextractedand matchedfrom thestoreddatainthedatabase.

OpticalCharacterRecognition [7]

Characterrecognition techniquesassociatea symbolicidentitywith theimageof character. Ina typicalOCRsystemsinputcharactersare digitizedbyanopticalscanner.Each character isthen located andsegmented,and theresultingcharacterimageisfed intoa pre-processorfornoise reduction and normalization.Certain characteristicsaretheextracted from thecharacterfor classification.

III. GENERAL OVERVIEW

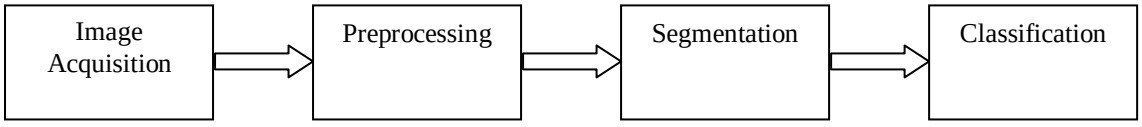

Figure.1- A general overview

CharacterRecognition proceed in mainly four steps.

(1) Image Acquisition (2) Preprocessing (3) Character Segmentation and (4) Classification.

IV. IMAGE ACQUISITION

In Imageacquisition,therecognitionsystemacquiresa scanned imageasaninput image. The imageshould have a specificformatsuchasJPEG,BMTetc.Thisimageis acquired throughascanner,digitalcameraoranyothersuitable digitalinputdevice.

A. Image Pre-processing

Theprocessofenhancing theimage,whichshould beusedforfurtherprocessing, iscalledpreprocessing. Preprocessingisthemajorstepinhandwriting recognition system. Noiseinadocument imageisduetopoorly photocopiedpages.Thescanned images notonlyhavenoises whichare inbuilt within it, but also the noise may be during the scanningof thatimage.

a)Gray scale processing

It includes converting RGB to gray for noise removal, border enhancement for brightness. Colored images are also complex in space and time, therefore it is necessary to convert them to gray scale to reduce time and space complexity. The basic idea behind gray conversion is to eliminate hue and saturation by not affecting its luminance. For this, we compute the threshold of an image by using the suitable gray scale value. This separates the object of interest from background. Thresholding is important to provide sufficient contrast of an image such that, varying level of intensity between foreground and background are considered. Gray scale conversion enhances the quality of an image for later computational processes. Gray scale images consist of different ranges of gray values; from 0 to 255.

b)Binarization

Theconversionofthegrayscaleimagetoblack and whiteiscalledbinarization.Binaryimagesarealsocalledas Bi-levelortwo-level.First,theoriginalRGBimageshould beconvertedtograyscale andthentheimageisconvertedto blackandwhiteimage.MostoftheOCRpackages work on thebinarizedimages.Theconversionispossible becauseofthethreshold valuesandthevalueswhicharehigherthanthethresholdare whiteandthevalueswhicharelowerthanthisthreshold are black.Otsu'smethodisusedtoperform thresholdbasedon clusteri.e. fromgraylevelimageto binaryimage.

c) Normalization

Theprocessofchangingtheintensity valueofthe pixeltotherangeof[0, 1]iscallednormalization inimage processing.The conversionof variousdimensionimages intofixeddimensions isalsocalledasnormalization.

d)Thinning

Thinningisapre-process whichresultsinsingle pixelwidthimagetorecognizethe character easily.Itisappliedrepeatedlyleavingonlypixel-widelinear representations oftheimagecharacters.

e) Median Filtering

It comes under the category of nonlinear filters. It changes the gray value of the pixels to the median of the gray value of surrounding pixels. We use a 3x3 mask and calculate the corresponding gray value of each pixel using the 8 neighboring pixels. This helps in noise removal. Median filtering gives advantages such as no reduction in contrast since output values are its neighborhood values, boundaries remain unchanged. Median filters are very useful in the presence of impulse noises also called salt and pepper noise because of its appearance as white and black dots superimposed on an image.

f)Morphological Operations

They are generally used to remover noise from the imperfect segmentation. Morphological operations are especially suited for binary images. So they are performed on output image of thresholding. Here Dilation and erosion are performed. Dilation and erosion are used to remove holes in the detected foreground. In the process of dilation the size and shape determination of structuring element is very important

V. SEGMENTATION

A)Line Segmentation

Thesegmentation isthemostimportantprocessintext recognition.Segmentation isdonetomaketheseparation betweentheindividualcharactersofanimage. Line segmentation process is segmenting the text region into lines, also called as line segmentation

*B***)** *Character Segmentation*
Wordsegmentationis easierthanlinesegmentationand character segmentation.Space betweentwowords isgenerallymorethanthreepixels.Wordsare segmentedby the projectionbasedmethod.

C) Skew Detection
The skew detectionused in character recognitionwhereinthewholewordisinscribed inapolygonwithatleast twodimensions. The centerofgravityoftheimageisconsideredas asinglelineextending acertainanglewith the horizontal.Theangleismeasuredwhichgives theangleby whichthe wordordocumentis rotatedandalsosignifiesthedirection andangle by whichisshouldberotatedforittobeatextin readableandnormalform.

D) Hough Transform
This is a technique for feature detection. Initially it was usedfor line detection. Now it has been extended to find positionof different shapes like circle, square, oval etc. Therefore we can use it to detect number plates from the image based on their rectangular shape.

E)Bounding Box
Segmentation is one of the most important processes in the number plate recognition, because all further steps rely on it. If the segmentation fails, a character can be improperly divided into two pieces, or two characters. The ultimate solution on this problem is to use bounding box technique. The bounding box is used to measure the properties of the image region. Once a bounding box created over each character and numbers presented on image, each character & number is separate out for recognition.

VI. CLASSIFICATION

Theclassification istheprocessofidentifyingeach characterandassigningtoitthecorrect characterclass,sothat texts inimages areconverted intocomputer understandable form.

A) Template matching
It is useful when we have characters of fixed size. Thistechnique is differentfromtheothersinthatno features are actually extracted. Instead the matrix containing theimageoftheinputcharacterisdirectly matched witha setofprototypecharactersrepresenting eachpossibleclass.Thedistancebetween thepatternand eachprototype iscomputed,andtheclass oftheprototype giving thebestmatch isassignedtothepattern.

B)Neural network
Character Recognition algorithm can also be used to recognize the vehicle number. The character can be recognized using neural network like artificial neural network(ANN), back propagation neural network, kohenen neural network etc.
IntheArtificialNeuralNetwork (ANN)for classificationneural networkcangetitselftrained automatically onthebasisofefficienttoolsforlearninglarge databasesand examples.Thisapproachisnon-algorithmicand trainable. TheKohonenneuralnetworkworksdifferently than thefeed forwardneuralnetwork.The Kohonenneuralnetwork containsonlyan inputand outputlayer of neurons.Thereis no hidden layerinaKohonenneuralnetwork.

C) Support Vector Machine (SVM)
SVMs have become more and more important in the field of pattern recognition. SVM is forcefully competing with many methods for classification. An SVM is a supervised learning technique. SVM takes Statistical Learning Theory (SLT) as its theoretical foundation, and the structural risk minimization as its optimal object to realize the best generalization. They are based on some simple ideas and provide a clear intuition of what learning from examples is all about. More importantly, they possess the feature of high performance in practical applications. The SVMs use hyper planes to separate the different classes. Many hyper planes are fitted to separate the classes, but there is only one optimal separating hyper plane. The optimal one is expected to generalize well in comparison to the others. A new data sample is classified by the SVM according to the decision boundary defined by the hyper plane. Among many classification methods, SVM has demonstrated superior performance.

VII. CONCLUSION

Character recognition has different phases and accuracy of each phase dependent on previous phase. Different kinds of page or worst document and skew lines or characters under which the systems are designed to operate are the main challenges. From this survey it can be concluded that various pre-processing methods like median filter, morphological operation that are erosion and dilation used

to improve the image quality. Segmentation can be done by various techniques like line and word segmentation. For recognition process, various classification methods like neural network, SVM can be used to improve the result.

REFERENCES

[1] RomeshLaishram,PheiroijamBebisonSingh, ThokchomSukaDebaSinghandSapamAnilkumar,AngomUmakantaSingh," A Neural Network BasedHandwritten Meitei MayekAlphabetOpticalCharacterRecognition System", 2014 IEEE International Conference on Computational Intelligence and Computing Research.

[2] Mr.PratikMadhukarManwatka,Mr. ShashankH. YadavDepartment," Text RecognitionfromImages", *IEEESponsored2nd InternationalConference onInnovations inInformation,EmbeddedandCommunicationsystems (ICIIECS)2015*

[3] NishaSharma,Bhupendra Kumar ,Vandita Singh," Recognition of Off-line Hand printed EnglishCharacters, "Numerals and Special Symbols", *20145thInternationalConference-ConfluenceTheNextGenerationInformationTechnologySummit(Confluence)*

[4]J.Pradeep,E.Srinivasan,S.Himavathi," Neural NetworkbasedHandwrittenCharacter Recognitionsystemwithout featureextraction", InternationalConferenceon Computer,CommunicationandElectricalTechnology –ICCCET2011, 18th& 19thMarch, 2011

[5] SheikhFaisalRashi,FaisalShafaitandThomasM.Breuel,"Scanning NeuralNetwork forTextLineRecognition",10th IAPR InternationalWorkshop on DocumentAnalysisSystems 2012

[6] SumanAvdheshYadav,SmitaSharma, ShipraRaviKumar,"A Robust Approach For Offline English CharacterRecognition, 20151stInternationalconferenceonfuturistictrendincomputationalanalysisandknowledgemanagement(ABLAZE2015)

[7] JagrutiChandarana[1],MayankKapadia," OpticalCharacterRecognition", International Journal ofEmerging Technology andAdvancedEngineering(ISSN2250-2459,ISO9001:2008CertifiedJournal,Volume4, Issue5, May2014)

