IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.
PP 148-152
www.iosrjournals.org

Handwritten Sentence Recognition Using Independent Component Analysis (Ica)

Beena Kamel[1], Shriram S[2]
ME II Year Student (ECE), MAM College of Engineering, Tiruchirapalli
Assistant Professor (ECE, MAM College of Engineering, Tiruchirapalli

***Abstract**: Handwriting is one of the most important means of daily communication. Although the problem of handwriting recognition has been considered for many years there are still many open issues, especially in the task of unconstrained handwritten sentence recognition. This article focuses on the automatic system that recognizes continuous English sentence based on ICA. The input provided is an image of some text, and the system produces, as output, an ASCII transcription of the input. This task involves a number of processing steps, some of which are quite difficult. Typically, preprocessing, segmentation, feature extraction, classification, and post processing operations are required.*

***Keywords:** Handwritten sentence recognition, ICA, preprocessing, segmentation, text identification*

I. Introduction

The Handwriting recognition refers to the identification of written characters. Handwriting recognition has become an acute research area in recent years for the ease of access of computer applications. Numerous approaches have been proposed for character recognition and considerable successes have been reported [1]. Traditional handwritten character recognition techniques enable a computer to receive and interpret intelligible handwritten input from sources such as papers, documents, touch-screens or pictures. During the last years, many popular studies and applications merged for bank check processing, mailed envelops reading, and handwritten text recognition in documents and videos [2]. Until now, it is still a difficult task for a machine to recognize human handwritings with significant accuracy, especially under variable circumstances such as variations in writings, variable sizes, and different patterns for different people etc.

English comprises of 26 basic alphabets which are simple to write and recognize but becomes very complex when they are handwritten. The work presented in this article is an effort towards the recognition of offline handwritten English sentence recognition by using an Independent component analysis (ICA) based feature extraction.

Independent component analysis (ICA) is a method for finding underlying factors or components from multivariate (multi-dimensional) statistical data. What distinguishes ICA from other methods is that it looks for components that are both statistically independant, and non Gaussian. ICA is important because of its potential applications in signal and image processing. The goal of ICA is to separate independent source signals from the observed signals, which is assumed to be the linear mixtures of independent source components. The mathematical model of ICA is formulated by mixture processing and an explicit decomposition processing. ICA decomposition enables to separate reflections, shadows and specularities from natural scene texts [3].
This work will help in better interface between human beings and computers and also facilitate the progress of the expansion of such systems for recognition of handwritten texts of other languages.

II. English Written Characteristics

Modern English (1500 A D to the present): Modern English developed after William Caxton established his printing press at Westminster Abbey in 1476. Johann Gutenberg invented the printing press in Germany around 1450, but Caxton set up England's first press. The Bible and some valuable manuscripts were printed. The invention of the printing press made books available to more people. The books became cheaper and more people learned to read. Printing also brought standardization to English. Since around the 9th century, English has been written in the Latin script, which replaced Anglo-Saxon runes. The modern English alphabet contains 26 letters of the Latin script. Early Modern English and Late Modern English vary essentially in vocabulary. Late Modern English has many more words, arising from the Industrial Revolution and the technology that created a need for new words as well as international development of the language [4].

III. Methodology

Handwritten sentence recognition involves 5 major steps
1) Preprocessing; 2) Segmentation; 3) Feature Extraction; 4) Classification; 5) Post processing.

3.1 Preprocessing: The preprocessing stage is a collection of operations that apply successive transformations on an image. It takes in a raw image and enhances it by reducing noise and distortion, and hence simplifies segmentation, feature extraction, and consequently recognition. Preprocessing operations are usually specialized image processing operations that transform the image into another with reduced noise and variation. Those operations include binarization, filtering and smoothing, thinning, alignment, normalization, and baseline detection. Ideally, preprocessing should remove all variations and detail from a text image that are meaningless to the recognition method.

3.2 Segmentation: The segmentation stage takes in a page image and separates the different logical parts, like text from graphics, lines of a paragraph, and characters (or parts thereof) of a word. After the preprocessing stage, most Optical Character Recognition (OCR) systems isolate the individual characters before recognizing them. Segmenting a page of text can be broken down into two levels: page decomposition and word segmentation, When working with pages that contain different object types like graphics, headings, mathematical formulas, and text blocks, page decomposition separates the different page elements, producing text blocks, lines, and sub-words. While page decomposition might identify sets of logical components of a page, word segmentation separates the characters of a sub-word.

3.3 Feature Extraction: The feature extraction stage analyzes a text segment and selects a set of features that can be used to uniquely identify the text segment. These features are extracted and passed in a form suitable for the recognition phase. Feature extraction is used to extract relevant features for recognition of characters based on these features. First features are computed and extracted and then most relevant features are selected to construct feature vector which is used eventually for recognition. The computation of features is based on structural, statistical, directional, moment, transformation like approaches. Feature extraction is extracting information from raw data which is most relevant for classification purpose and that minimizes the variations within a class and maximizes the variations between classes [5]. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, thinned (skeletons sentence) or gray-level sub images of each individual character. On feature extraction stage each character is represented by a feature vector, which becomes its identity [6]. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of confusion. The captured picture is normalized and thinned, it's distinct features are extracted using Independent Component Analysis (ICA).

3.3.1 Independent Component Analysis (ICA): ICA is the unsupervised computational and statistical method for discovering intrinsic hidden factors in the data. ICA exploits higher-order statistical dependencies among data and discovers a generative model for the observed multi dimesional data. In the ICA model, observed data variables are assumed to be linear mixtures of some unknown independent sources (independent components). A mixing system is also assumed to be unknown. Independent components are assumed to be non gaussian and mutually statistically independent. The purpose of ICA is to estimate both the mixing matrix H and the sources (independant components) s using sets of observed vectors x. The ICA model for the set of N patterns x, represented as columns in matrix X, can be given as X = HS,where S = [s1, s2, · · · , sN] is the $m \times N$ matrix which columns correspond to independent component vectors si = [$s1$, $s2$, · · · , sm]^T discovered from the observation vector xi. Once the mixing matrix H has been estimated, we can compute its inverse B = H^{-1}, and then the independent component for the observation vector x can be computed by s = Bx. The extracted independent components si are as independent as possible.

3.3.1.1 ICA Preprocessing: ICA is preceded by preprocessing, including centering and whitening.Centering of x is the process of subtracting its mean vector $\mu = E\{x\}$ from x: x = x − $E\{x\}$.The second preprocessing step in ICA is decorrelating (and possibly dimensionality reducing), called whitening. In whitening the sensor signal vector x is transformed using formula y = Wx.The purpose of whitening is to transform the observed vector x linearly so that we obtain a new vector y (which is white) which elements are uncorrelated and their variances are equal to unity. Whitening allows also dimensionality reduction, by projecting of x onto first l eigenvectors of the covariance matrix of x. Whitening is usually realized using the eigen-value decomposition (EVD) of the covariance matrix $E\{xx^T\} \in R^{n \times n}$ of observed vector x . Rxx = $E\{xx^T\}$ = Ex $\Lambda_x^{1/2}$ $\Lambda_x^{1/2}$E$_x^T$ Here, Ex ∈ R$n \times n$ is

the orthogonal matrix of eigenvectors of Rxx = $E\{xx^T\}$ and Λ is the diagonal matrix of its eigenvalues x = diag($\lambda_1, \lambda_2, \ldots \lambda_n$) with positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$. The whitening matrix can be computed as $W = \Lambda_x^{-1/2} E_x^T$ and consequently the whitening operation can be realized using formula $y = \Lambda_x^{-1/2} E_x^T x = Wx$. Recalling that x = Hs, we can find from the above equation that $y = \Lambda_x^{-1/2} E_x^T Hs = H_w s$. We can see that whitening transforms the original mixing matrix H into a new one, $H_w = \Lambda_x^{-1/2} E_x H$. Whitening makes it possible to reduce the dimensionality of the whitened vector, by projecting observed vector into first l ($l \cdot n$) eigenvectors corresponding to first l eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_l$ of the covariance matrix Ex. Then, the resulting dimension of the matrix W is $l \times n$, and there is reduction of the size of observed transformed vector y from n to l. Output vector of whitening process can be considered as an input to ICA algorithm. The whitened observation vector y is an input to unmixing (separation) operation s = By, where B is an original unmixing matrix. An approximation (reconstruction) of the original observed vector x can be computed as $\tilde{x} = Bs$, where $B = W_w^{-1}$

3.3.1.2 ICA Algorithm: ICA rotates the whitened matrix back to the original space. It performs the rotation by minimizing the Gaussianity of the data projected on both axis [7].

3.3.2 Image Representation: Training set of M images of size NxN are represented by vector of size N^2. Feature vector of a image is stored in a NxN matrix. Those two dimensional vector is changed to one dimensional vector. Each image is represented by the vector Γ_i.

Mean Image: It is calculated by $\Psi = (1/M) \sum_{i=1}^{M} \Gamma i$, each image differs from the average by $\phi_i = \Gamma_i - \Psi$ called mean centered image.

3.3.3 Covariance Matrix: A covariance matrix is constructed as $C = AA^T$ where $A = [\phi_1, \phi_2 \ldots \ldots \ldots \phi_M]$ of size $N^2 x N^2$. In statistics C captures the correlation between all possible pairs of measurements for which the diagonal and off diagonal terms are called the variance and covariance respectively. The correlation value reflects the noise and redundancy in the measured data. Eigen vectors corresponding to AA^T can easily be calculated with reduced dimensionality where AX_i is the Eigen vector and λ_i is the Eigen value.

3.3.4 Eigen Space: The Eigen vectors of the covariance matrix resemble the input image but look ghostly called Eigen face. Eigen vectors correspond to each Eigen face and discard the faces for which Eigen values are zero thus reducing the Eigen space to an extent. A face image can be projected into the face space by $\omega_i = U_i (\Gamma i - \Psi)$; i=1, 2, ...M, where U_i is the i^{th} Eigen face. The weight is obtained as above form a vector, $\Omega_i = [\omega_1, \omega_2, \ldots \ldots \ldots \omega_M]$.

Testing Sample Classification: Read the test image and separate the character from it. Calculate the feature vector of the test face. The test image is transformed into its Eigen vector components. First we compare the line of our image with mean image and multiply their difference with each Eigen vectors. Each value would represent a weight and would save on a vector Ω_{test}. $\Omega_{test} = U_{test}(\Gamma_{test} - \Psi)$ $\Omega_{test} = [\omega_1, \omega_2, \ldots \ldots \ldots \omega_M]$. Compute the average distance (Euclidean distance) between test feature vector and all the training feature vectors. Mathematically recognition is finding the minimum Euclidean distance $\varepsilon_M = \sqrt{(\Omega_{test} - \Omega_i)^2}$ where i=1,2, ...M. The Euclidean distance between two weight vectors thus provides a measurement of similarity between the corresponding images. The class with minimum Euclidean distance shows similarity to test image.

All methods for handwritten character, word or sentence recognition need to be trained. As a rule of thumb, the larger the training set, the better is the recognition performance of the system. This empirical finding has been confirmed in a number of experiments [8, 9, 10].

3.4 Classification: The Classification in an OCR system is the main decision making stage in which the features extracted from a pattern are compared to those of the model set. Based on the features, classification attempts to identify the pattern as a member of a certain class. When classifying a pattern, classification often produces a set of hypothesized solutions instead of generating a unique solution. The (subsequent) post-processing stage uses higher level information to select the correct solution. When input image is presented to HCR system, its features are extracted and given as an input to the trained classifier. Classifiers compare the input feature with stored pattern and find out the best matching class for input. Euclidean distance between two vectors of size N is given by: ED = $\sqrt{\sum_{i=1}^{N} (Ai - Bi)2}$, Euclidean distance between two dimensional images can be given by: ED = $\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (A(i,j) - B(i,j))2}$

3.5 Post Processing: The post-processing stage, which is the final stage, improves recognition by refining the decisions taken by the previous stage and recognizes words by using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies,

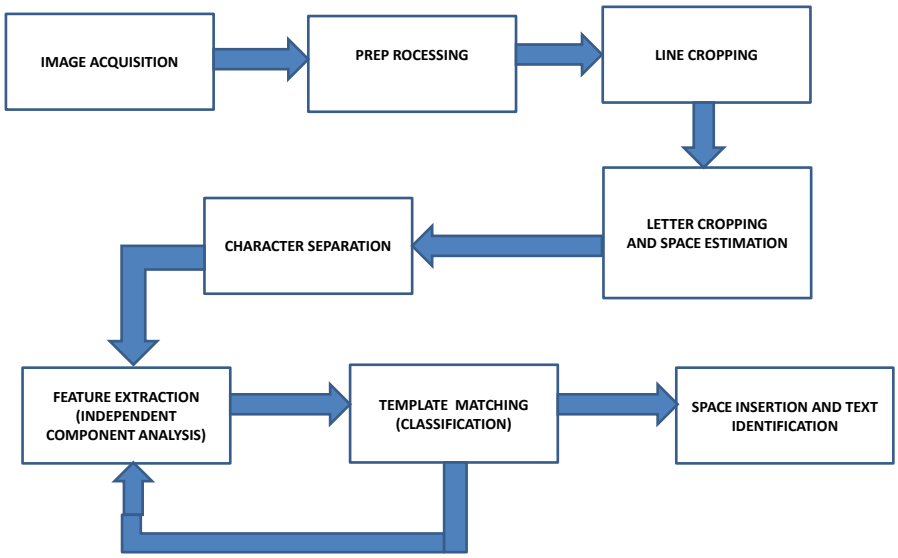
lexicons, and other context information. The most common post-processing operations are spell checking and correction. Spell checking can be as simple as looking up words in a lexicon.

IV. Software Tool

4.1 MATLAB: MATLAB is the high-level language and interactive environment used by millions of engineers and scientists worldwide. It lets the user to explore and visualize ideas and collaborate across disciplines including signal and image processing, communications, control systems, and computational finance. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, Fortran and Python.

4.2 Graphical User Interface Development Environment (Guide): MATLAB's Graphical User Interface Development Environment provides a rich set of tools for incorporating GUIs in M functions. The resulting graphical M function is composed of two identically named files, A file with extension .fig called FIG file contains a complete graphical description of all the functions of GUI objects and their spatial arrangements. A file with extension .m called a GUI M file contains the code that controls the GUI operation. This file include functions that are called when the GUI is launched and excited and call back functions that are executed when a user interacts with GUI objects eg. pushing a button.

V. Block Diagram Of Proposed Methodology



VI. Conclusion

The focus of attention in handwriting recognition has been shifting from isolated character recognition to more complex tasks, such as recognition of words and unconstrained text. It can be expected that a significant percentage of future systems for cursive handwriting recognition will be personal systems serving a single user. Such systems will perform best when they are trained with data provided by the future user. However it can be quite cumbersome for an individual to provide a sufficiently large body of handwritten samples to train a system [11]. Natural language processing techniques, and machine translation [12] are very promising to improve the recognition performance of today's handwriting recognition procedures. In addition they would naturally lead to tools for content based search and retrieval in the context of archives of handwritten texts. The ultimate goal of handwriting recognition is to have machines which can read any text with the same recognition accuracy as humans but at a faster rate [13]. Sentence recognition is important in a number of future applications, for example, the transcription of personal notes, faxes, and letters, or the electronic conversion of historical handwritten archives in the context of the creation of digital libraries [14].

Acknowledgments

The author wants to thank her Head of Department Dr. S. Shanmughapriya for valuable contributions to this article.

References

[1]. R. Plamondon and S.N. Srihari., 2000, "Online and off-line handwriting recognition: a comprehensive survey.", Pattern Analysis and Machine Intelligence, IEEE Transac-tions on, vol. 22(1), pp 63–84.

[2]. A. L. Bianne-Bernard, F. Menasri, R. A. H. Mohamad, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem., 2011 , " Dynamic and contextual information in hmm modeling for handwritten word recognition", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33(10), pp 2066–2080.

[3]. Independent Component Analysis Final version of 7 March 2001 Aapo Hyvarinen, Juha Karhunen, and Erkki Oja ¨ A Wiley-Interscience Publication JOHN WILEY & SONS, INC. https://www.cs.helsinki.fi/u/ahyvarin/papers/bookfinal_**ICA**.pdf

[4]. Algeo, and John, 2010, "The Origins and Development of the English Language", Boston, MA: Wadsworth. pp. 182-187

[5]. O. D. Trier, A. K. Jain and T. Taxt, 1996, "Feature Extraction Methods for Character Recognition A Survey", Pattern Recognition, vol. 29, pp. 641-662.

[6]. S. V. Rajashekararadhya, and P. VanajaRanjan , 2005, "Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts", Journal of Theoretical and Applied Information Technology, pp 1171- 1181

[7]. Aapo Hyvärinen, Erkki Oja (2000), 'Independent Component Analysis: Algorithms and Applications', Neural Networks, Vol. 13(4-5) P. No. 411-430

[8]. J. Cano, J.-C. Perez-Cortes, J. Arlandis, and R. Llobet. Training set expansion in handwritten character recognition. In T. Caelli, A. Amin, R. Duin, M. Kamel, and D. de Ridder, editors, Structural, Syntactic and Statistical Pattern Recognition, pages 548– 556. LNCS 2396, Springer, 2002.

[9]. H. Rowly, M. Goyal, and J. Bennet. The effect of large training set sizes on online Japanese Kanji and English cursive recognizers. In 8th Int. Workshop on Frontiers in Handwriting Recognition, pages 36–40, 2002.

[10]. S. Smith. Handwritten character classification using nearest neighbour in large databases. IEEE Trans. on Pattern Analysis and Machine Intelligence, 16:915– 919, 1994

[11]. G. Lorette. Handwriting recognition or reading? what is the situation at the dawn of the 3rd millenium? Int. Journal on Document Analysis and Recognition, 2:2– 12, 1999

[12]. R. Rosenfeld. Two decades of statistical language modeling: Where dowe gofrom here? Proc. of the IEEE, 88:1270–1278, 2000

[13]. A. Vinciarelli and S. Bengio. Writer adaptation techniques in HMM based off-line cursive script recognition. Pattern Recognition Letters, 23:905–916, 2002.

[14]. C. Tomai, B. Zhang, and V. Govindaraju. Transcript mapping for historic handwritten document images. In 8th Int. Workshop on Frontiers in Handwriting Recognition, pages 413–418, 2002.