Abstract

OASIS is an interactive question - answering system, which answers questions in common English. Developed as a result of on-going research at the Tata Infotech's, Cognitive Systems Research Lab, OASIS is specifically oriented towards answering questions and providing information on a given topic, and is the first product of this type to be developed in India. The engine has been set-up to answer questions about Tata Infotech.

The system adopts a user-friendly conversational style. Each answer is preceded by a conversational 'prelude' which confirms to the user that the question has been properly understood. It also suggests other topics on which the user can ask questions. In its present form, OASIS will not deal with involved questions of certain types and extremely complicated sentences.

A variant of this has been implemented which answers questions relating to textual information as well as information from standard databases. The question in natural language is in this case converted into a query in SQL and the information obtained from the database is presented to the user, embedded in an English sentence. OASIS is proposed to be made available both as a product and as a solution

b) Optical Character Recognition (OCR)

5.1.6 A Hybrid Scheme For Handprinted Numeral Recognition Based On A Self-Organizing Network And MLP-Based Classifiers

U. Bhattacharya, T. K. Das, A. Datta, S. K. Parui and B. B. Chaudhuri, *International Journal of Pattern Recognition & Artificial Intelegence, vol 16, no 7 (2002) pp 845-864.*

Abstract

This paper proposes a novel approach for automatic recognition of handprinted Bangla (an Indian script) numerals. A Topology Adaptive Self Organizing Neural Network is first used to extract vector skeleton from a binary numeral pattern. Simple heuristics are considered to prune artifacts, if appeared, in such a skeletal shape. Certain structural and topological features like loops, junctions, positions of terminal nodes etc. are used along with a hierarchical tree classifier to classify handwritten numerals into smaller subgroups. A multilayer perceptron network (MLP) is used to classify different numerals in each subgroup uniquely. The system is trained using a sample data set of 1800 numerals and we obtained 93.26% correct recognition rate and 1.71% rejection on a disjoint test set of another 7760 samples. In addition, a separate validation set consisting of 1540 samples has been used to determine the termination of the training of the associated MLP networks. The proposed scheme is sufficiently robust with respect to considerable object noise and it does not consider any size normalization.

5.1.7 On Developing High Accuracy OCR Systems For Telugu And Other Indian Scripts

Chakravarthy Bhagavati, Tanuku Ravi, S. Mahesh Kumar, Atul Negi, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

In this paper, we list a number of factors that are important in achieving high recognition accuracy in OCR systems for Telugu and other Indian scripts. While it is relatively easy to obtain 85% - 93% accuracy, it becomes increasingly difficult to improve the performance further. We discuss how the factors presented in this paper helped achieve an accuracy of nearly 97% with our OCR system for Telugu script. It is expected that these factors are specific not only to Telugu but also work for other Indian scripts in general and south Indian scripts in particular.

5.1.8 Recognition Of Handprinted Bangla Numerals Using Neural Network Models

U. Bhattacharya, T. K. Das, A. Datta, S. K. Parui, and B. B. Chaudhuri, *Advances in Soft Computing - AFSS 2002, Springer Verlag, Lecture Notes on Artificial Intelligence, Eds. N.R. Pal and M. Sugeno, LNAI 2275, 2002, pp. 228-235.*

Abstract

This paper proposes an automatic recognition scheme for handprinted Bangla (an Indian script) numerals using neural network models. A Topology Adaptive Self Organizing Neural Network is first used to extract from a numeral pattern a skeletal shape that is represented as a graph. Certain features like loops, junctions etc. present in the graph are considered to classify a numeral into a smaller group. If the group is a singleton, the recognition is done. Otherwise, multilayer perceptron networks are used to classify different numerals uniquely. The system is trained using a sample data set of 1880 numerals and we obtained 90.56% correct recognition rate on a test set of another 3440 samples. The proposed scheme is sufficiently robust with respect to considerable object noise.

5.1.9 Self-Organizing Neural Network-Based System For Recognition Of Handprinted Bangla Numerals

U. Bhattacharya, T. K. Das, A. Datta, S. K. Parui, and B. B. Chaudhuri, *Proceedings of XXXVI Annual Convention, Computer society of India, 2001, Kolkata, pp. C-92 - C-96.*

Abstract

This paper proposes an automatic recognition scheme for handprinted Bangla (an Indian script) numerals using a Topology Adaptive Self Organizing Neural Network (TASONN) model. The Neural Network model is used to extract from a numeral pattern a skeletal shape that is represented as a planar straight line graph. Certain features like loops, junctions etc. present in the graph are considered to classify a numeral into smaller groups. If the group is a singleton, the recognition is done. Otherwise the graph is subjected to various feature extraction procedure depending on the group. Recognition is done using a look-up table of specific ranges of the extracted feature values. This scheme does not require normalization with respect to size and it performs satisfactorily even in the presence of considerable noise.The system tested on a test set of 3330 samples with 89.63% correct recognition rate.

5.1.10 A Complete Printed Bangla OCR System

B. B. Chaudhuri and U. Pal, *Pattern Recognition, vol. 31, pp. 531-549, 1998.*

Abstract

A complete Optical Character Recognition (OCR) system for printed Bangla, the fourth most popular script in the world, is presented. This is the first OCR system among all script forms used in Indian sub-continent. The problem is difficult because (i) there are about 300 basic, modified and compound character shapes in the script, (ii) the characters in a word are topologically connected and (iii) Bangla is an inflectional language. In our system the document image captured by Flat-bed scanner is subject to skew correction, text graphics separation, line segmentation, zone detection, word and character segmentation using some conventional and some newly developed techniques. From zonal information and shape characteristics, the basic, modified and compound characters are separated for the convenience of classification. The basic and modified characters which are about 75 in number and which occupy about 96% of the text corpus, are recognized by a structural feature based tree classifier. The compound characters are recognized by a tree classifier followed by template matching approach. The feature detection is simple and robust where preprocessing like thinning and pruning are avoided. The character unigram statistics is used to make the tree classifier efficient. Several heuristics are also used to speed up the template matching approach. A dictionary based error correction scheme has been used where separate dictionaries are compiled for root word and suffixes that contain morpho-syntactic information's as well. For single font clear documents 95.5% word level (which is equivalent to 99.10% character level) recognition accuracy has been obtained. Extension of the work to Devnagari, the third most popular script in the world, is also discussed.

5.1.11 An OCR System To Read Two Indian Language Scripts: Bangla And Devnagari (Hindi)

B. B. Chaudhuri and U. Pal, *Proc. Fourth Int. conf. on Document Analysis and Recognition, IEEE Computer Society Press, pp. 1011-1016, 1997.*

Abstract

An OCR system is proposed that can read two Indian language scripts: Bangla and Devnagari (Hindi), the most popular ones in Indian subcontinent. These scripts, having the same origin in ancient Brahmi script, have many features in common and hence a single system can be modeled to recognize them. In the proposed model, document digitization, skew detection, text line segmentation and zone separation, word and character segmentation, character category are done for both scripts by the same set of algorithms. The feature sets and classification tree as well as knowledge base required for error correction (such as lexicon) differ for Bangla and Devnagari. The system shows a good performance for single font scripts printed on clear document.

5.1.12 Automatic Recognition Of Printed Oriya Script

B. B. Chaudhuri, U. Pal and M. Mitra, Sadhana, *(a journal of Indian Academy of Sciences) vol.27, part 1. pp.23-34, 2002.*

Abstract

This paper deals with an Optical Character Recognition (OCR) system for printed Oriya script. The development of OCR for this script is difficult because a large number of character shapes in the script have to be recognized. In the proposed system, the document image is first captured using a flat-bed scanner and then passed through different preprocessing modules like skew correction, line segmentation, zone detection, word and character segmentation, etc. These modules have been developed by combining some conventional techniques with some newly proposed ones. Next, individual characters are recognized using a combination of stroke and run-number based features, along with features obtained from the concept of water overflow from a reservoir. The feature detection methods are simple and robust, and do not require preprocessing steps like thinning and pruning. A prototype of the system has been tested on a variety of printed Oriya material, and currently achieves 96.3% character level accuracy on average.

5.1.13 Development Of A Page Layout Analyzer For Multilingual Indian Documents

Ray Chaudhuri, A.K.Mandal, B.B.Chaudhuri *(Fellow IEEE), Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

An advanced Optical Character Recognition (OCR) system is equipped the module of the page layout analyser. It separates textual zones from non-textual zones. It identifies textual blocks from multicolumn documents and groups them into homogenous regions in terms of geometric shape and spatial distribution. All existing OCR modules developed for various Indian scripts can handle text only single-column documents. In this paper, a page layout analyser that uses typical common features present in most of the Indian scripts is introduced. A simple compatibility criterion that allows various degrees of homogeneity is defined. The page-analyser is robust in the sense that it can distinguish text regions from non-textual entities such as images, rulers, and noisy signals due to smudges and poor quality of the paper. Test results are shown in two most popular Indian Scripts, Devnagari (Hindi) and Bangla.

5.1.14 Skew Angle Detection Of Digitized Indian Script Documents

B.B. Chaudhuri and U. Pal, *IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 19, pp.182-186, 1997.*

Abstract

Skew angle detection of scanned documents containing most popular Indian scripts Devnagari and Bangla is considered. Most characters in these scripts have horizontal lines at the top, called head lines. The character head lines mostly join one another in a word and the word appears as a single component. In the proposed method the components are at first labeled. The upper envelope of a component is found by column-wise scanning from an imaginary line above the component. Portions of upper envelope satisfying the properties of digital straight line are detected. They are clustered as belonging to single text lines. Estimates from individual clusters are combined to get the skew angle. Apart from accuracy and efficiency, an advantage of the method is that character segmentation and zone detection can be readily done from head line information, which is useful in Optical Character Recognition approaches of these scripts.

5.1.15 Optical Character Recognition (OCR) System For Malayalam Language

K. Jithesh, K.G. Sulochana, R Ravindra Kumar *National Workshop on Application of Language Technology in Indian Languages, to be held in Hyderabad, March 6-8, 2003*

Abstract

In this paper we present a brief description of a Malayalam Optical Character Recognition (OCR) system. The presence of two different scripts (Old Script and New Script) and a large number of characters (including conjuncts) makes the Malayalam OCR a complex system. The proposed system is based on the **Feature Extraction** method of character recognition. Feature extraction can be considered as finding a set of vectors, which effectively represent the information content of a character. In character recognition, it is desirable to extract features, which are focused on discriminating between classes. The features are identified after the careful study of Malayalam writing system. A two level segmentation scheme, feature extraction method and classification scheme, using binary decision tree, for Malayalam characters is described. Different pre-processing modules like noise removal, skew detection and correction, line, word and character segmentation are also dealt with. The presence of touching characters particularly in the case of consonant-vowel modifier combinations renders the character segmentation process difficult. The development of the post processor, a spell checker tuned for the OCR system, is in progress. The objective of the post processing is to correct errors in OCR output by using Malayalam grammar rules, lookup table (word list) and statistical information collected from the corpora. The beta version of the System gives an accuracy of 97% at character level for good quality printouts. We are now working on Document layout analyser module, the addition of which will enable the OCR to reproduce the document in its original layout.

5.1.16 A Complete OCR System For Gurmukhi Script

G S Lehal and Chandan Singh, *Proceedings SSPR2002, Windsor, Canada, Lecture Notes in Computer Science, Vol. 2248, Springer-Verlag, Germany, pp. 344-352, (2002)*

Abstract

Recognition of Indian language scripts is a challenging problem. Work for the development of complete OCR systems for Indian language scripts is still in infancy. Research in the field of recognition of Gurmukhi script faces major problems mainly related to the unique characteristics of the script like connectivity of characters on the headline, characters in a word present in both horizontal and vertical directions, two or more characters in a word having intersecting minimum bounding rectangles along horizontal direction, existence of a large set of visually similar character pairs, multi-component characters, touching characters which are present even in clean documents and horizontally overlapping text segments. This paper

addresses the problems in the various stages of the development of a complete OCR for Gurmukhi script and discusses potential solutions. A multi-font Gurmukhi OCR for printed text with an accuracy of more than 97% at character level is presented. The recognition system presented in this paper operates at connected component level. The segmentation process decomposes the text image into connected components. After feature extraction, the connected components are fed to a classifier, which recognizes the connected component. The connected components are then combined to form Gurmukhi characters. A set of very simple and easy to compute structural features is used and a hybrid classification scheme consisting of binary decision trees and nearest neighbours is employed. To further improve the results at word level, each word is fed to a post processor, which uses the statistical information of Punjabi language syllable combination, corpora look up and certain heuristics based on Punjabi grammar rules.

5.1.17 A Range Free Skew Detection Technique For Digitized Gurmukhi Script Documents

G S Lehal and Renu Dhir, *Proceedings 5th International Conference of Document Analysis and Recognition, Bangalore, pp. 147-152, (1999)*

Abstract

In this paper, a range free skew detection technique for machine printed Gurmukhi documents has been presented. Most characters in Gurmukhi script have horizontal lines at the top called headlines. The characters forming a word are joined at top by headlines, so that the word appears as one single component with headline. The ratio of pixel density above and below the headline of any word in Gurmukhi script is always less than one. These inherent characteristics of the script have been employed and a new algorithm based on projection profile method has been devised. The skew angle is determined by calculating horizontal and vertical projections at different angles at fixed interval in range [0°, 90°]. Under such projections, for an image with no skew, headlines appear as distinct peaks while gaps between successive text rows will be represented by valleys. The bitmapped image is partitioned into ten equal sized horizontal and vertical zones and the highest peaks and valleys are determined for projections in each zone. The angle at which the difference of the sum of heights of peaks and valleys is maximum is identified as the skew angle. To decrease the computational cost, first the course skew angle is calculated by taking the angle interval 3°. Once the course skew angle is found, the accurate skew angle q is determined by looking in the range [q - 3°, q + 3°] at intervals of 0.25°. The image is then rotated over -q, where q is the skew angle

Since the skew angle is checked only in the range [0°-90°] and the image can be skewed at any angle in the range [-180°, 180°], the rotated image may need another additional rotation by 90°, -90° or 180°. If the rotated image is skewed at 90° or -90°, then the highest peaks and valleys would be present in vertical projection else they will be reported in horizontal projection. To determine the skew angle of the image aligned with y-axis, if the foreground pixel density on the left side of headlines is greater than pixel density on right side for text rows then the image is skewed at -90 else it is skewed at 90°. Similarly for the image aligned with x-axis, if the foreground pixel density above the headlines is lesser than pixel density below then the image is straight else it is upside down. In the end the image is rotated by the second rotation angle to completely remove any skew present in the image.

5.1.18 A Recognition System For Devnagri And English Handwritten Numerals

G. S. Lehal and Nivedan Bhatt, *Advances in Multimodal Interfaces – ICMI 2001, T. Tan, Y. Shi and W. Gao (Editors), Lecture Notes in Computer Science, Vol. 1948, Springer-Verlag, Germany, pp. 442-449. (2000).*

Abstract

Handwritten numeral recognition has been extensively studied for many years and a number of techniques have been proposed [1-5]. However, handwritten character recognition is still a difficult task in which human beings perform much better. The problem of automatic recognition of handwritten bilingual numerals is even more tough. In this paper a bilingual OCR system for handwritten numerals of Devnagri(Hindi) and Roman scripts has been presented. It is assumed at a time the numerals will be of one of the above two scripts and there are no mixed script numerals in an input string. A set of global and local features, which are derived from the right and left projection profiles of the numeral image, are used. For identification of script, no separate routine is used. During the recognition process when the first numeral of a numeric string is recognized correctly, the context (Devnagri/ English) is set to the domain of that particular numeral's character set. Subsequent identification of the remaining numerals is carried out in that context only which drastically reduces the search space and hence increases the performance of the system. It was observed that the Devnagri numeral set had a very good recognition and rejection rate, as compared to the English set. Also the Devnagri numeral set's recognition module had good rejection rates for the numerals of the English character set. This property was exploited by adopting a *polling strategy* in which the input numeral is first tested by the Devnagri module.

If the numeral is recognized then the context is set to Devnagri, else it is tested for English set and on recognition, the context is set to English. In case the numeral is rejected by both script sets, then the next numeral is tested, and this continues till one of the numeral is recognized. Subsequent identification of the other numerals is carried out for character set of the recognized numeral. Numeral 0 is the same for both the character sets, thus in the case when the first numeral encountered is a zero, subsequent numeral is checked before deciding the context. The system was tested on 1000 samples of both the Devnagri and English character set. For the Devnagri numeral set, a recognition rate of 89% and a confusion rate of 4.5% were obtained. For the English numeral set we had a recognition rate of 78.4%, confusion rate of 18 % and rejection rate of 3.6%.

5.1.19 A Shape Based Post Processor For Gurmukhi OCR

G S Lehal, Chandan Singh and Ritu Lehal, *Proceedings of 6th International Conference on Document Analysis and Recognition, Seattle, USA, IEEE Computer Society Press, USA, pp. 1105-1109, (2001)*

Abstract

The objective of post processing is to correct errors or resolve ambiguities in OCR results by using contextual information. In this paper a shape based post processing system for an OCR of Gurmukhi script has been presented. The Punjabi corpus developed at MIT under TDIL project has been used, which serves the dual purpose of providing data for statistical analysis of Punjabi language and also checking the spelling of a word. The corpus has been partitioned at two levels. At the first level the corpus is split into seven disjoint subsets based on the word length. At second level the shape of the word is used to further segment the subset into a list of visually similar words. A set of robust, font and character size independent features are used for identification of visually similar words. These features are available more or less as a by-product of the on-going recognition process and do not necessitate any additional computation. For each set of visually similar word the percentage frequency of occurrence of character in all the positions is recorded. This list is combined with the confidence rate of recognition of the recognizer to correct the mistakes of the recognizer. Holistic recognition of most commonly occurring words derived from the corpora is also carried out. An improvement of 3% in recognition rate from 94.35% to 97.34% has been reported on machine printed images using the post processing techniques.

5.1.20 A Structural Feature Based Approach For Script Identification Of Gurmukhi And Roman Characters And Words

G S Lehal, Chandan Singh and Renu Dhir, *Proceedings SPIE, (Jan. 2003)*

Abstract

Roman script words are now commonly being used in Gurmukhi script documents. An OCR developed for the Gurmukhi script will wrongly recognize these words in Roman script. So it is necessary to filter out these Roman script words before feeding the Gurmukhi script words to the OCR. Considering the nature of many documents in the Indian context, where the script could change at the word level or even single characters of different script may be present, there is need for development of method to identify the script of a character or a word. In this paper we have proposed a method to automatically differentiate between Gurmukhi and Roman script words and characters based on a combined analysis of several discriminating features. After a careful study of shapes of Gurmukhi and Roman script characters and words we have developed nine features for automatic classification of Roman and Gurmukhi scripts. Some of these features are common for identification at both character and word level, while some features are suitable only for either word or single characters only. These features are *Headline pixel count , Inter character gap, Bottom Projection Profile, Protruding Regions Beyond Headline, Right Vertical Bar, Loop in lower half, Left Vertical bar, C shape in lower half* and *U like shape*. This method has been implemented and tested on about 100 documents, and the experimental results indicate that this method is effective and reliable. The recognition accuracy of the system is 98.29% for Gurmukhi script words and 99.02% for Roman script words. For single characters, the accuracy is 96.81% and 95.47% for Roman and Gurmukhi scripts respectively. This is the first time that such a script recognition system has been developed for Roman and an Indian language script, which works down to word and character level.

5.1.21 A Technique For Segmentation Of Gurmukhi Text

G S Lehal and Chandan Singh, *Computer Analysis of Images and Patterns,* W. Skarbek *(Ed.), Lecture Notes in Computer Science, Vol. 2124, Springer-Verlag, Germany, pp. 191-200 (2001)*

Abstract

In this paper a technique for text segmentation of machine printed Gurmukhi script documents is discussed. Research in the field of segmentation of Gurmukhi script faces major problems mainly related to the unique characteristics of the script like

connectivity of characters on the headline, two or more characters in a word having intersecting minimum bounding rectangles, multi-component characters, touching characters which are present even in clean documents and horizontally overlapping text segments. After digitization of the text, the text image is subjected to pre-processing routines such as noise removal, thinning and skew correction. The thinned and cleaned text image is then sent to the text segmenter, which segments the text image into connected components. The text image is sliced into horizontal text strips using horizontal projection in each row. The gaps on the horizontal projection profile are taken as separators between the text strips. But this step does not always results in a single text row in each horizontal strip. Usually a text line is broken up into two or more horizontal strips. We call these strips as zones. The text line is broken into a single core zone, which is made up of characters of middle and upper zone and optionally lower zones, followed by other minor zones representing the lower zones. In some rare cases the core zone is also split into two zones representing the upper and middle zones. Some other problems also occur such as the lower characters and vowels of some of the text lines intruding into the core zone of the successor text line and more than one text line being present in a horizontal strip because of overlap. A statistical analysis of strip heights and position of headline is used to identify if a strip contains more than one text line or only a lower zone or middle and upper zone or only upper zone of a text line. The text strips are next decomposed into connected components. In case of multi strips, which contain multiple text lines, horizontal cuts are made at appropriate locations and the connected components are located in each horizontal cut. Simple heuristics are also used to detect and split touching connected components in upper and middle zones.

5.1.22 Feature Extraction And Classification For OCR Of Gurmukhi Script

G S Lehal and Chandan Singh, Vivek, *Vol. 12, No. 2, pp. 2-12 (1999).*

Abstract

In recent years, there has been a renewed attempt to reformat the classification approaches to the recognition of difficult character sets. It has been found that a multiple classification character recognition scheme has the potential of outperforming individual stand-alone classifiers because of its ability to handle extreme variance in the training and testing samples. In this paper, a feature extraction and a hybrid classification scheme, using binary decision trees and nearest neighbours, for machine recognition of Gurmukhi characters is described. The classification process is carried out in three stages. In the first stage the characters are grouped into three sets depending on their zonal position (upper zone set, middle zone set and lower zone). In the second stage the characters in middle zone set are further distributed into smaller sub-sets by a binary decision tree using a set of robust and font independent features at each node of the tree. The terminal node of the binary decision tree contains a subset of Gurmukhi characters and the cardinality of the set varies from 1 to 8. The final categorization of the input sample is then easily tackled by using a suitable recognition schemes for each subset, considering the special features and peculiarities of the characters in each subset. In the third stage the nearest neighbour classifier is used and the special features distinguishing the characters in each subset are used. One significant point in this scheme is that in contrast to the conventional single-stage classifiers where each character image is tested against all prototypes, a character image is tested against only certain subsets of classes at each stage. This eliminates unnecessary computations.

5.1.23 A New Algorithm For The Restoration Of Characters In Old Noisy Document With Varying Level Of Intensities

S. Mohanty, K. Sahoo and H. K. Behera, *Indian Science Congress, 2003.*

Abstract

An image can have noise and interference from several sources like electrical sensor, photographic grain, and channel errors. Noise cleaning is the process of removing unwanted noise from an image and is necessary for better recognition of characters in a document. These noise effects must be reduced for better analysis of character documents, which will lead to efficient character extraction and recognition process. We have developed a novel algorithm, which is simple and clears noise from character document images to a very good extent. In the proposed methodology, the algorithm is based on intensities, which clears the noises and isolated points present on the document.

5.1.24 A Solution For Ligatures During Optical Recognition of Oriya Characters

S. Mohanty, K. Sahoo and H. K. Behera, *Published at: Proceedings of IEMCT 2002, CDAC, Pune.*

Abstract

Recognition of alphabetic characters is a basic need in incorporating intelligence to computers. Machine intelligence involves several aspects among which optical recognition is a tool, which can be integrated to text recognition and speech recognition. To make these aspects effective character recognition with better accuracy is important.

Very often even in printed text, adjacent characters tend to be touched or connected. Generally, through binarization of images, useful information for the segmentation of touched character are lost in many cases which leads to the accuracy being handicapped when problem like ligatures predominates during recognition. We in our study have developed an algorithm, which provides a solution to overcome the complexity of ligatures especially for connected characters during segmentation leading towards more accurate recognition.

From the Histogram of different size neighborhoods of characters the optimal character size found out and tested on further segmentation. This technique is a recognition-based segmentation for connected characters.

The algorithm has successfully been tested for Oriya characters and the output has been integrated with Oriya Text-to-Speech system.. This technique can be applied for other Indian as well as some foreign languages during Optical Character Recognition.

5.1.25 Pattern Recognition In Alphabets Of Oriya Language Using Kohonen Neural Network

S. Mohanty, *Published at International Journal of Pattern Recognition and Artificial Intelligence Vol.12 No.7 (1998).*

Abstract

Here a computerized reading of alphabets of Oriya language is attempted using the Kohonen neural network and its unsupervised competitive learning capacity as self- organizing map or the Kohonen feature map. The proposed pattern recognition does not treat a pattern as an n-dimensional feature vector or a point in n-dimensional space as is done in the traditional pattern recognition theory. We have tried with all the Oriya alphabets and have presented the study with respect to five of them in this paper along with their average distance per pattern in each cycle till we reach the permissible average distance. In the output picture the variation of the weight vector with respect to the alphabets is clearly observed.

5.1.26 Page Layout Analyzer For Multilingual Indian Documents

A.Mandal and Prof. B. B. Chaudhuri, *that will be published in the Proceedings of the Language Engineering Conference 2002 by IEEE CS Press.*

Abstract

An advanced Optical Character Recognition (OCR) system is equipped the module of the page layout analyser. It separates textual zones from non-textual zones. It identifies textual blocks from multicolumn documents and groups them into homogenous regions in terms of geometric shape and spatial distribution. All existing OCR modules developed for various Indian scripts can handle text only single-column documents. In this paper, a page layout analyser that uses typical common features present in most of the Indian scripts is introduced. A simple compatibility criterion that allows various degrees of homogeneity is defined. The page-analyser is robust in the sense that it can distinguish text regions from non-textual entities such as images, rulers, and noisy signals due to smudges and poor quality of the paper. Test results are shown in two most popular Indian Scripts, Devnagari (Hindi) and Bangla.

5.1.27 Automatic Identification of English, Chinese, Arabic, Devnagari And Bangla Script Line

U. Pal and B. B. Chaudhuri, *In Proc. Sixth Int. Conf. on Document Analysis and Recognition, IEEE Computer Society Press, pp.790-794, 2001.*

Abstract

In a general situation, a document page may contain several script forms. For Optical Character Recognition (OCR) of such a document page, it is necessary to separate the scripts before feeding them to their individual OCR systems. In this paper, an automatic technique for the identification of printed Roman, Chinese, Arabic, Devnagari and Bangla text lines from a single document is proposed. Shape based features, statistical features and some features obtained from the concept of water reservoir, have been used for script identification. The proposed scheme has an accuracy of about 97.33%.

5.1.28 Automatic Recognition Of Unconstrained Off-Line Bangla Hand-Written Numerals

U. Pal and B. B. Chaudhuri, *Advances in Multimodal Interfaces, Springer Verlag Lecture Notes on Computer Science (LNCS-1948), Eds. T. Tan, Y. Shi and W. Gao, pp. 371-378, 2000.*

Abstract

This paper deals with an automatic recognition method for unconstrained off-line Bangla handwritten numerals. To take care of variability involved in the writing style of different individuals a robust scheme is presented here. The scheme is based on new features obtained from the concept of water overflow from the reservoir as well as topological and statistical features of the numerals. If we pour water from upper part of the character, the region where water will be stored in the character is imagined as a reservoir of the character. The direction of water overflow, height of water level when water overflows from the reservoir,

position of the reservoir with respect to the character bounding box, shape of the reservoir etc. are used in the recognition scheme. The proposed scheme is tested on data collected from different individuals of various background and we obtained an overall recognition accuracy of about 91.98%.

5.1.29 Automatic Separation Of Words In Indian Multi-Lingual Multi-Script Documents

U. Pal and B. B. Chaudhuri, *Proc. Fourth Int. Conf. on Document Analysis and Recognition, IEEE Computer Society Press, pp. 576-579, 1997.*

Abstract

In a multi-lingual country like India, a document page may contain more than one script forms. For such a document it is necessary to separate different script forms before feeding them to OCRs of individual script. In this paper an automatic word segmentation approach is described which can separate Roman, Bangla and Devnagari scripts present in a single document. The approach has a tree structure where at first Roman script words are separated using the 'headline' feature. The headline is common in Bangla and Devnagari but absent in Roman. Next, Bangla and Devnagari words are separated using some finer characteristics of the character set although recognition of individual character is avoided. At present, the system has an overall accuracy of 96.09%.

5.1.30 Machine-Printed And Hand-Written Text Lines Indentification

U. Pal and B. B. Chaudhuri, *Pattern Recognition Letters, Vol.22, No. 3-4, pp. 431-441, 2001*

Abstract

There are many types of documents where machine-printed and hand-written texts intermixedly appear. Since the optical character recognition (OCR) methodologies for machine-printed and hand-written texts are different, to achieve optimal performance it is necessary to separate these two types of text before feeding them to their respective OCR systems. In this paper, we present a machine-printed and hand-written text classification scheme for Bangla and Devnagari, the two most popular Indian scripts. The scheme is based on the structural and statistical features of the machine-printed and hand-written text lines. The classification scheme has an accuracy of 98.6%.

5.1.31 Multi-Skew Detection Of Indian Script Documents

U. Pal, M. Mitra and B. B. Chaudhuri, *In Proc. Sixth Int. Conf. on Document Analysis and Recognition, IEEE Computer Society Press, pp. 292-296, 2001.*

Abstract

There are many documents where text lines are not parallel to each other i.e. these lines have different inclinations with the horizontal lines (multi-skew documents). For the OCR of such a document we have to estimate the skew angle of individual text lines because a single rotation cannot de-skew all text lines of the document. In this paper, we describe a robust technique for multi-skew angle detection from Indian documents containing the most popular Indian scripts Devnagari and Bangla. Most characters in these scripts have horizontal lines at the top, called head-lines. The character head-lines usually connect one another in a word and the word appears as a single component. In the proposed method, the connected components are at first labeled and selected. The upper envelopes of selected components are found by column-wise scanning from the top of the component. Portions of the upper envelope satisfying the properties of a digital straight line are detected. They are then clustered into groups belonging to single text lines. Estimates from these individual clusters give the skew angle of each text line. The proposed multi-skew detection technique has an accuracy about 98.3%.

5.1.32 OCR Error Correction Of An Inflectional Indian Language Using Morphological Parsing

U. Pal, P. K. Kundu and B. B. Chaudhuri, *Journal of Information Science and Engineering, Vol.. 16. No.6. pp. 903-922, 2000.*

Abstract

This paper deals with an OCR (Optical Character Recognition) error detection and correction technique for a highly inflectional Indian language, Bangla, the second-most popular language in India and fifth-most popular language in the world. The technique is based on morphological parsing where using two separate lexicons of root words and suffixes, the candidate root-suffix pairs of each input string are detected, their grammatical agreement are tested and the root/suffix part in which the error has occurred is noted. The correction is made on the corresponding error part of the input string by a fast dictionary access technique. To do so, the information about the error patterns generated by the OCR system are examined and some alternative strings are generated for an erroneous word. Among the alternative strings, those satisfying grammatical agreement in root and suffix are finally chosen as suggested words. In the list of suggested words generated by the system, the desired word is available in 84.22% cases.

5.1.33 OCR In Bangla : An Indo-Bangladeshi Language

U. Pal and B. B. Chaudhuri, *Proc. of 12th Int. Conf. on Pattern Recognition, IEEE Computer Society Press, pp. 269-274, 1994.*

Abstract

In this paper a complete OCR system is described for documents of single Bangla (Bengali) font. The character shapes are recognized by a combination of template and feature matching approach. Images digitized by flatbed scanner are subjected to skew correction, line word and character segmentation, simple and compound character separation feature extraction and finally character recognition. A Feature based tree classifier is used for simple character recognition. Preprocessing like thinning and skeletonization is not necessary in our scheme and hence the system is quite fast. At present, system has an accuracy of about 96%. Also, some character occurrence statistics have been computed to model an error detection and correction technique in near future.

5.1.34 On The Development of An Optical Character Recognition (OCR) System For Printed Bangla Script

U. Pal, *Ph.D. Thesis, Indian Statistical Institute, 1997.*

Abstract

This thesis is devoted to the development of a complete OCR system for printed *Bangla* script. The content of the thesis is divided into three major divisions. They are (a) Preprocessing Division (b) Recognition Division (c) Postprocessing Division.

In Preprocessing Division, at first, some statistical studies have been made on Bangla script characters. Also, the application potentials of these studies in OCR design have been described. Next, binarization, smoothing, skew detection and correction as well as text/graphics separation techniques have been elaborated. The skew correction method proposed here is simple, fast and robust. The method has been developed using inherent characteristics of the script form. Finally, techniques for segmentation of text into lines, zone detection, word segmentation from line and character segmentation from word have been described.

In recognition division at first feature selection and detection procedure as well as analysis on binary tree classifier have been presented. Next, automatic character recognition procedure using a hybrid method is described. Combination of a feature based tree classifier and a run length based template matching approach has been used for the purpose. For convenience of classification we have initially classified the characters into three categories namely basic, modified and compound character. The basic and modified class of characters have been recognized by a feature based tree classifier while the compound character recognition involves a template matching approach preceded by a feature based subgrouping.

In postprocessing division we proposed an OCR error detection and correction technique for a highly inflectional language like Bangla. Using two separate lexicons of root words and suffixes, we detect candidate root-suffix pair of each word and test their grammatical agreement, and note the root suffix part in which the error has occurred. The correction is done on the corresponding error part of the input string by a fast dictionary access technique.

5.1.35 Printed Devnagari Script OCR System

U. Pal and B. B. Chaudhuri, Vivek, *vol.10, pp.12-24, 1997.*

Abstract

An Optical Character Recognition (OCR) system for printed Devnagari script is presented in this paper. Development of an OCR system for Devnagari is difficult because (i) there are about 350 basic, modified (matra) and compound character shapes in the script and (ii) the characters in a word are topologically connected. In our system the document image captured by a flatbed scanner is subjected to noise cleaning, skew correction, line segmentation, zone detection, and word and character segmentation using some conventional and some newly developed techniques. From zonal information and shape characteristics, the basic, modified and compound characters are separated for the convenience of classification. Modified and basic characters are recognized by a structural feature based tree classifier while compound characters are recognized by hybrid approach. At present, the system has an accuracy of about 96% at the character level.

5.1.36 Script Line Separation From Indian Multi-Script Documents

U. Pal and B. B. Chaudhuri, *IETE Journal of Research*

Abstract

In a multi-lingual country like India, a document page may contain more than one script form. Under the three-language formula, the document may be printed in English, Devnagari and one of the other Indian official languages. For Optical Character Recognition (OCR) of such a document page, it is necessary to separate these three script forms before feeding them to the OCRs of individual scripts. In this paper an automatic technique of separating the text lines is presented for almost all triplet of script forms. To do

so, the triplets are grouped into five classes according to their characteristics and shape based features have been employed to separate them without any expensive OCR-like algorithms. The proposed approaches are tested on many documents and the experimental results are presented. At present, the system has an overall accuracy of about 98.5%.

5.1.37 Script Line Separation From Indian Multi-Script Documents

U. Pal and B. B. Chaudhuri, *Int. Conf. on Document Analysis and Recognition, IEEE Computer Society Press, pp 406-409, 1999.*

Abstract

In a multi-lingual country like India, a document page may contain more than one script form. Under the three-language formula, the document may be printed in English, Devnagari and one of the other official Indian languages. For OCR of such a document page, it is necessary to separate these three script forms before feeding them to the OCRs of individual scripts. In this paper, an automatic technique of separating the text lines using script characteristics and shape based features is presented. At present, the system has an overall accuracy of about 98.5%.

5.1.38 Connected Script Synthesis By Character Concatenation - An Overlap And Weighted Average Formulation

V. Ramasubramanian and P.V.S. Rao, *Proceedings of SEARCC-88, New Delhi, Nov. 28-Dec.1, 1988, pp. 163-176.*

Abstract

In this paper we address the problem of synthesizing connected handwritten script from individual characters written isolation. Connected writing is viewed as a natural evolution from writing the characters in isolation, characterized by the use of continuous pen-down connecting movement from one character to the next. The problem is one of concatenation of individual character shapes to generate the connected script and consists in synthesizing a so called transition stroke from one character to the next. Particular emphasis is laid on recreating the context effect underlying the pen-down transition stroke and in preserving the continuity of motion and shape in the transition. Under this framework, we propose an approach based on the temporal description of the context effect as an overlap and weighted addition of appropriate segments of the temporally adjacent characters. The weighting (or blending) functions are designed and solved analytically such that the resulting transition curve satisfies a specified order of continuity at the transition points. We show that the weighting function is equivalent to a non-parametric Bezier curve with Bernstein polynomials as the basis function and a simple control point configuration which is a function only of the order of continuity sought. This equivalence result provides additional insight to the solution and a very simple way of generating the weighting function and the transition curve for any required order of continuity. The proposed approach is shown to generate connected script having a convincingly high degree of naturalness.

5.1.39 A Knowledge-Based Approach For Script Recognition Without Training

P.V.S. Rao, *IEEE Transactions on Pattern Analysis & Machine Intelligence (PAMI), Vol. 17, No. 12, pp. 1233-1239, December, 1995.*

Abstract

The approach is based on an empirical parametric model for the writing hand system. The parameters are so chosen and quantized as to retain only broad shape information ignoring writer-dependent and other variability. Concatenation of character prototypes generates archetypal reference words for recognition; training is unnecessary. Recognition scores exceed 90%.

5.1.40 Cursive Script : Recognition And Synthesis-Recent Trends

P.V.S. Rao, *Sectional Presidential Address at the 78th Science Congress, Indore, Jan.3-9, 1991.*

Abstract

It has been the usual practice in the Science Congress for Sectional students to choose as the topic for their Presidential Address, either an area in which they have themselves made contributions or an area which relates to the main theme chosen for that year.

Since this is the first Presidential Address to be delivered for the Computer Science Section, I thought it would be appropriate to choose a subject which is very much a current area for research. I felt also that it would be appropriate if the area has practical applications. In addition, it would be desirable to choose a subject that is not esoteric and highly theoretical, but one that is of consequence even to the non-specialist. It would then lend itself to ready appreciation without substantial effort to follow the jargon or the methodology. It would be an added attraction if the area relates to applications where computers appear to display capability which, if humans had it, would be called 'intelligence': i.e. the area of artificial intelligence. Accordingly, I have chosen the area of script recognition by computer: an area which is challenging, has numerous practical advantages, falls in the field of artificial intelligence,

and does not require much specialized knowledge to appreciate. It also happens to be an area in which I have been working in the recent past and to which I have made some non-trivial contributions. I shall therefore be talking about synthesis of cursive script words out of individual characters, characters of simpler elements and about recognition of characters and words by computers.

5.1.41 Script Recognition

P.V.S. Rao, Sadhana, *vol. 19 part 2, pp. 257-270, April, 1994.*

Abstract

This paper describes an approach for word-based on-line and off-line recognition of handwritten cursive script composed of English lower-case letters. The system uses simple and easily extractable features such as the direction of movement and curvature and the relative locations of regions where these suffer discontinuities.

Our approach was evolved based on our concept of 'shape vectors' introduced earlier. We visualize script characters as having shapes which are composed of comparatively straight segments alternating with regions of relatively high curvature. We derive the shape vectors from each script character essentially by identifying regions of least curvature and approximating these by straight lines. That these shape vectors carry adequate information about the identity of the character is established by showing that the original character can be faithfully reconstructed from the shape vectors.

We thus use slopes of the shape vectors and relative locations of points of maximum curvature (both highly quantized) as parameters for recognition. The system extracts parameters for individual characters from single specimens written in isolation and uses these to construct feature matrices for words in the vocabulary. These are used for matching with the feature matrices of test words during the recognition phase.

The advantage of the system is that it does not require elaborate training. Recognition scores are in the neighborhood of 94% for vocabulary sizes of 200 words. The approach has been extended for off-line information as well and performs quite well even in this case.

5.1.42 SHAPE VECTORS : An Efficient Parametric Representation For The Synthesis And Recognition Of Hand-Script Characters

P.V.S. Rao, Sadhana, *Vol. 18, Part 1, pp. 1-15, March 1993.*

Abstract

Earlier work by the author has established: (i) that cursive script can be synthesized out of individual characters by using polynomial merging functions which satisfy boundary conditions of continuity of the displacement functions x(t) and y(t) for each character and their first and second derivatives; and (ii) that the procedure lends itself to a Bezier curve based formulation. This was done since cursive writing avoids discontinuities (of shape) between individual characters as well as discontinuities in pen movement.

We show here that even individual characters can be synthesized out of more primitive elements by using the same merging functions. We choose directed straight lines which we call shape vectors as basic elements for this. Script characters generally have shapes which are essentially straight segments alternating with 'bends' or regions of relatively high curvature. For a character with n bends, we need only n+1 shape vectors. Thus each script character needs only three to seven shape vectors, depending on its complexity.

The "character generation" shape vectors are derived from the original character by means of a simple procedure that identifies comparatively straight regions in it. These are then approximated to straight line by linear regression and positioned to be tangential to the original curve. The synthesized version of this character is obtained by 'merging' or concatenating these vectors. The close fit between the original and re-synthesized characters demonstrates that the shape vectors adequately characterize their identities and shapes. Data reduction ratios in the range of 15 to 25 are thus possible. This method thus shows good promise as a possible basis for script character recognition, and a recognition scheme based on it has yielded an accuracy of 94% for a vocabulary size of 67 words.

5.1.43 Shape Vectors For On-Line And Of-Line Recognition Of Cursive Script Words

P.V.S. Rao, *First Intl. Conference on Document Analysis and Recognition (ICDAR 91), Saint-Malo, France, Sept. 30-Oct. 2, 1991, pp. 568-575.*

Abstract

This paper describes a novel approach for cursive script recognition and uses simple and easily extractable features such as the direction of movement and curvature and the relative locations of regions where these suffer discontinuities. This is based on earlier work which showed that (1) cursive script can be synthesized out of individual characters by using polynomial merging functions which satisfy boundary conditions of continuity of the displacement functions x(t) and y(t) for each character and their first and second derivatives, and that (2) even individual characters could be synthesized out of straight lines (which we call shape vectors). The recognition scheme does not require elaborate training and yields recognition scores around 94%.

5.1.44 Telugu Script Recognition - A Feature Based Approach

P.V.S. Rao and T.M. Ajitha, *International Conf. on Document Analysis and Recognition (ICDAR-95), Montreal, Canada, August 14-16, 1995.*

Abstract

Telugu characters can be visualized as being composed of circular segments of different radii. Recognition consists in segmenting the characters into the component elements and identifying them. We choose a feature set to preserve the canonical shapes while filtering out as noise the shape deviations encountered in real life. Hence, this approach does not require extensive training. Instead, 'Feature Vector' parameters for individual 'basic' characters are extracted from single specimens written in isolation. These are suitably combined to construct 'Feature Vectors' for compound characters for the lexicon. These are compared with similar 'Feature Vectors' extracted from the test samples to be recognized. Recognition scores ranged from 78 to 90% across different subjects, (when the best match alone is taken) and from 91 to 95% for a single subject.

5.1.45 A Complete Multi-Font OCR Systems For Printed Telugu Text

C.Vasantha Lakshmi, C.Patvardhan, *Language Engineering Conference, University of Hyderabad, India, Dec. 2002.*

Abstract

This work describes the design and development of a Telugu Optical Character Recognition system for Printed text (TOSP). Pre-processing tasks considered in this paper are: Conversion of gray scale image to a binary image, image rectification, skew detection and removal, segmentation of text into lines, words and basic symbols. Basic symbols are identified as the fundamental unit of segmentation in this paper which is recognized by the recognizer. The combinations of these basic symbols that together form characters and compound characters of Telugu are also determined to complete the recognition process. The special feature of TOSP is that it is designed to handle multiple sizes and multiple fonts. Further, the output produced by TOSP can directly be opened in any Indian language software that supports transliteration facility into Telugu script and edited. Several such soft wares are popular and available.

c) Speech Recognition

5.1.46 Relational Studies Between Phoneme And Grapheme Statistics In Current Bangla

B.B. Chaudhuri & U. Pal, *Journal of Acoustical Society of India, vol.-23, pp. 67-77, 1995.*

Abstract

This paper deals with the study of phonemic and graphemic character occurrences in words of Bangla (Bengali) language. The occurrence statistics of characters are presented for words collected from newspaper and popular magazines. Some of the computed grapheme, their position wise occurrences, percentage of compound character, word length versus frequency of occurrence, bi-gram etc. This study can be applied in Speech recognition, Speech analysis, Character recognition, Keyboard setting, Linguistics, Data communication, Cryptography and Error correction.

5.1.47 Animating Expressive Faces To Speak In Indian Languages

Tanveer A. Faruquie, Chalapathy Neti, Nitendra Rajput, L. Venkata Subramaniam, Ashish Verma, *NCC, Mumbai, Jan 26-27, 2002.*

Abstract

This paper describes a morphing based automated audio driven facial animation system. A novel scheme to implement a language independent system for audio-driven facial animation given a speech recognition system for just one language, in our case, English, is presented.. New viseme and expression combinations are synthesized to be able to generate animations with new facial expressions.

5.1.48 Spoken Word Recognition: Lexical Vs Sublexical

Amit Gupta and Sandeep M, Workshop on Spoken Language Processing, TIFR, Mumbai, January 9-11, 2003 .

Abstract

Spoken word has become a primary object of scientific inquiry with a focus on understanding how our speech perception capacities are used in segmenting and recognizing words in fluent speech. The present study investigated the nature of spoken word representation. Ten normal native Kannada speakers in the age range of 15-25yrs participated in the study. A word-spotting technique was used. Eighty Kannada words and non-words with 5 words and 5 non-words appearing twice were audio presented. The subjects were instructed to press the button when they heard the same word/ non-word for the second time and responses were audiorecorded which were then analyzed for the reaction time. The results of the present study indicated that words are spotted better than non-words supporting a lexical representation of words.

5.1.49 Data-guided Processing of Speech

Hynek H ermansky, *Workshop on Spoken Language Processing, TIFR, Mumbai, January 9-11, 2003 .*