A Survey on Offline-Methods of Character Segmentation

Thakkar Nirav B
Kandarp Pandya
Juhi Kaneria
Ruchita Tailor
Cpmputer Science and Engineering
Shri S'ad Vidya Mandal Institute of Technology
Bharuch, India

Kruti J Dangarwala(Faculty)
Computer Science and Engineering
Shri S'ad Vidya Mandal Institute Of Technology
Bharuch, India

Abstract— Character segmentation is the critical area of the Optical Character Recognition process. The higher recognition rates for isolated characters as compared to those obtained for words and connected character strings illustrate this fact.

This paper provides a review of various techniques of character segmentation, which are classified mainly into four classes. In classical approach the input image is partitioned into sub images, which are then classified. The operation of attempting to decompose the image into classifiable units is called "dissection". In the second class of method, the dissection method is avoided and the image is segmented either explicitly by classification of pre specified windows, or implicitly by classification of subsets of spatial features collected from the image. The third strategy is hybrid of the first two, employing dissection together with recombination rules to define potential segments, but using classification to select from the range of admissible segmentation possibilities offered by these sub images. Finally, holistic approaches avoid segmentation by recognizing entire character strings as units.

Keywords—segmentation, contextual method, graphemes, Hidden Markov Models, holistic recognition, Optical character recognition, recognition-based segmentation and survey

I. INTRODUCTION

A. Role of segmentation in the OCR system

The optical character recognition system consists of two main processing units – a character separator and an isolated character classifier.

Character separation also known as segmentation can work in two modes:

- Fixed (constrained) spacing mode: In this mode character size is known in advance and therefore segmentation can be very robust.
- Variable (arbitrary) spacing mode: In this mode no priori information can be assumed.

In character segmentation phase, an image of sequence of characters is decomposed into sub images of individual symbol or character. A character is a pattern that resembles one of the symbols the system is designed to recognize. To determine such a resemblance the pattern must be segmented from the document image. In this paper we present a survey of character segmentation phase of the optical character recognition system. It provides the basic information of the methods used for segmentation.[1]

B. Methods surveyed for segmentation

A major problem in discussing segmentation is how to classify methods. Based on the interaction between segmentation and classification, there are three types of character segmentation methods (Fig. 1)[2]:

- Classical/Dissection approach: a single partitioning of the image into sub images based on "character-like" properties followed by the classification of the sub images.
- Recognition-based approach: segmentation where the image is iteratively searched for components that most closely match the classes in the alphabet.
- Holistic approach: segment and recognize words as single unit.

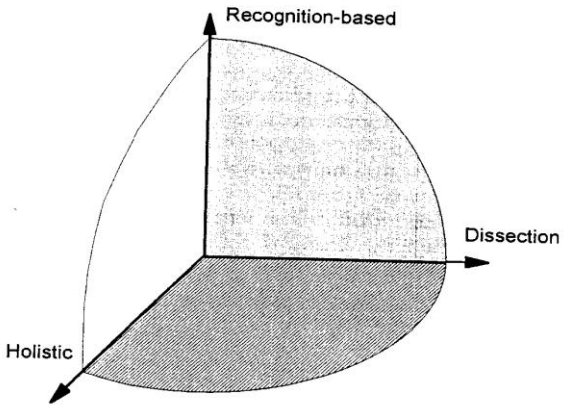

Figure 1. The three fundamental strategies of segmentation occupy orthogonal axis, and the hybrid methods can be represented as weighted combinations these, lying at points in intervening space.

The segmentation methods can also be represented in a hierarchical way as shown in the Fig. 2 on the next page.[1].

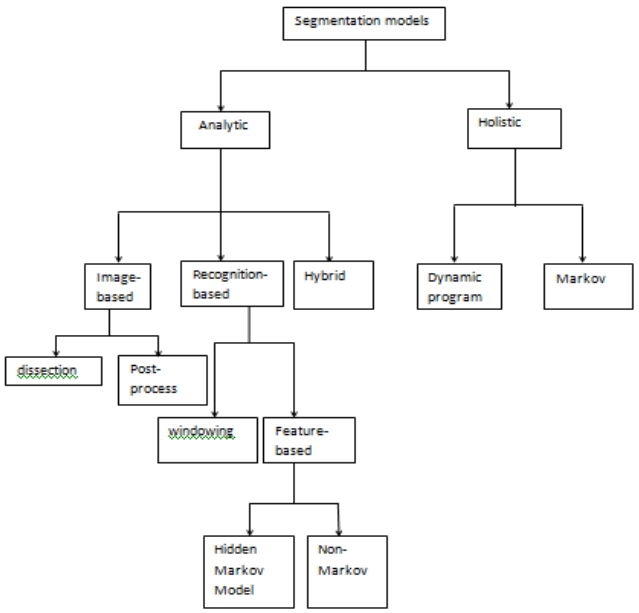

Figure 2. The hierarchical representation of the segmentation methods

II. DISSECTION APPROACH

Dissection is an operation to decompose the image into sequence of sub images using general features (Fig. 3). It is an intelligent process in that an analysis of the image is carried out, however classification into symbols is not involved at this point[1]. The criterion for good segmentation using the dissection approach is the agreement of character properties in the segmented sub image and the expected symbol. The character properties include height, width, separation from neighbouring components, disposition along the baseline, etc. Interaction with the classifier is limited to reprocessing of ambiguous recognition results. For example, if the classifier can't make any decision at all, the segment may need to be split again into new segments [2].

A. *Dissection directly into characters*

1) White space and Pitch: The simplest and earliest dissection approach relies on the vertical whitespace between successive characters. The number of characters per unit of horizontal distance is defined as "pitch" and it can be used for estimating segmentation points. For the sake of convenience the segmentation approach used a fixed pitch.

In machine printing, vertical white space can be used to separate successive characters, and to use the same strategy in handwritten characters, separate boxes can be provided for each symbol.

Another variant of this strategy uses two scans of print line; in the first scan (from left to right) the pitch distance D is measured, while actual segmentation is done in the second scan (from right to left). In this strategy double white columns triggered the segmentation boundary, in case within a distance D if no such boundary is obtained then segmentation was forced.

Hoffman and McCullough [3] designed a system that could aid in segmentation when a fixed pitch could not be enforced. The system consisted of three steps:

1. Detection of the start of a character based on an a priori pitch measurement.
2. A decision to begin testing for the end of the character.
3. Detection of the end of a character.

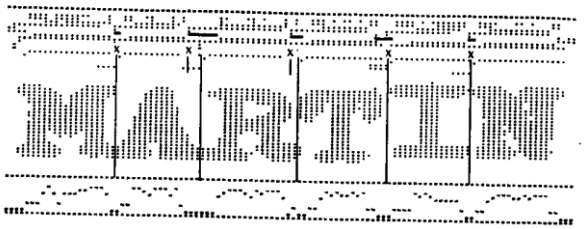

Figure 3. The dissection method of Hoffman and Mc Cullough. An evaluation function based on a running count of horizontal black-white and white black transitions is plotted below the image. The horizontal black bars above the image indicate the activation region for the function. Vertical lines indicate human estimates of optimal segmentation columns, while Xs indicate columns chosen by the algorithm.

In this method the step 2 is critical, which is based on weighted analysis of horizontal black runs completed versus runs still incomplete. An estimate of the character pitch is the parameter of the process. Once the sectioning algorithm indicated the region of permissible segmentation, segmentation can be achieved by using the rules of increased bit density or by using special features to detect end-of-character.

The authors reported 97% accuracy, but the results were heavily dependent on the quality of the input image. The input data consisted of machine-printed lines of 10-11- and 12-pitch serif-type multi font characters.

2) *Projection Analysis:* In the projection profile methods, the horizontal and vertical profiles are computed. Projection profile is the histogram of the image. When the projection profiles are plotted we can see peaks and valleys in the plot. The zero valued valleys are identified to separate the lines, words and characters. The horizontal profile is used for line segmentation and vertical profile is used for words and character segmentation. This method is suitable for segmenting image documents that are well spaced without overlapping and touching. The vertical projection of a printed line consists of a simple running count of the black pixels in each column. It can serve for detection of white space between successive letters. Thus analysis of the projection

of a line has been used as a basis for segmentation of non-cursive writing.

When printed characters touch or overlap horizontally, projection often contains minimum at proper segmentation column. In another method the projection is first obtained then the ratio of the second derivative of this curve to its height is used for choosing separating columns (Fig. 4b). [1]

In another variant of this method a peak-to-valley function was designed to improve the above method. A minimum of the projection is located and the value is noted. The sum of differences between this minimum value and the peaks on each side is calculated. The ratio of the sum to the minimum value itself is the discriminator used to select segmentation boundaries.

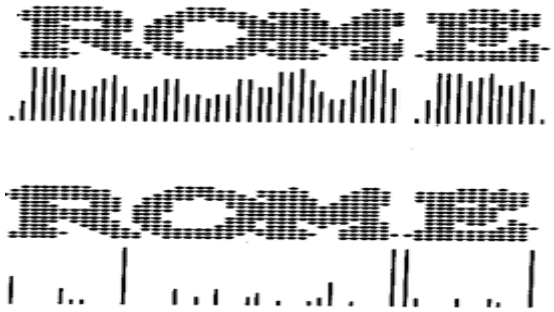

Figure 4: Dissection based on projection: (a) Vertical projection of an image. (b) Differencing measure for column splitting

3) *Connected Component Processing:* The above method, one based on pitch and the other based on projection analysis, fails to give satisfactory results when width of characters is variable and the characters are slanted.

Segmentation of handprint or kerned machine printing requires two dimensional analyses, which is generally based on determining the connected black regions ("connected components") Fig(5) and might require further processing.

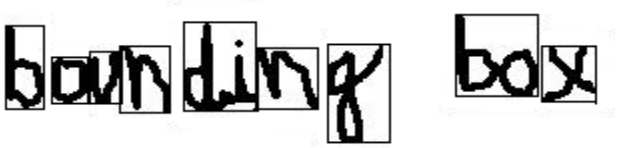

Figure 5: Connected components. This example illustrates characters that consist of two components (e.g., the "u" in "you"), as well as components consisting of more than one character (e.g., the "ry" in "very"). The bounding box of each component is also shown. The latter is often used as a basis for dissection methods

Mainly there are two types of follow up processing:

1. Based on bounding box
2. Based on the detailed analysis of connected components.

a) Bounding Box Analysis: Bounding box method is generally used for segmentation of non cursive characters. By testing the adjacency relationships or the size and aspect ratio the characters can be merged or split. Connected components have also served to provide a basis for the segmentation of scanned handwriting into words. Here it was assumed that words do not touch, but may be fragmented. Thus the problem is to group fragments (connected components) into word images.

b) Splitting of Connected Components: In order to separate joined characters reliably, more detailed processing is necessary. Intersection of two characters can give rise to special image features. Thus dissection methods are developed to find these features from the image and these can be used to split the image into sub images. Mainly these methods work as follow-on process after the bounding box analysis. Only image components failing certain dimensional tests are subjected to detailed examination. One of the concerns in separation along straight-line path is that accuracy is very low for slanted characters as well as for overlapping characters. Accurate segmentation requires analysis of the shape of the pattern to be split, as well as the determination of an appropriate segmentation path.

One of the algorithm designed, uses contour analysis for the detection of likely segmentation points, which uses local vertical minima encountered in following the bottom contour as "landmark points." Successive minima detected in the same connected components are assumed to belong to different characters, which are to be separated.

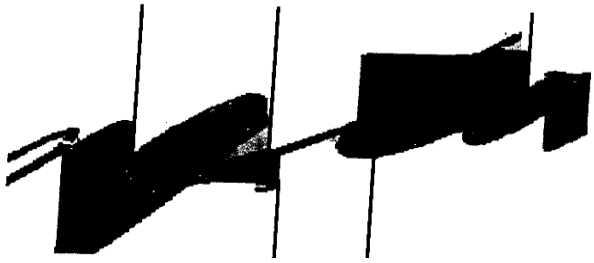

Figure 6: Dissection based on search and deflect. The initial path of the cut trajectory is along a column corresponding to a minimum in the upper profile of the image. When black is encountered the path is modified recursively, seeking better positions at which to enforce a cut. In this way, multiple cuts can be made at positions which are in the shadow region of the image.

Another variant of the algorithm not only detects likely segmentation points, but also computes an appropriate segmentation path. In this algorithm first a vertical scan is performed in the middle of the image assumed to contain possibly connected characters. If the number of black to white transitions is 0 or 2, then the digits are either not connected or else simply connected, respectively, and can therefore be separated easily by means of a vertical cut. If the number of transitions found during this scan exceeds 2, the writing is probably slanted, which requires the use of a special algorithm based on "Hit and defect strategy." This

algorithm computes the curved segmentation path by iteratively moving a scanning point, which starts from the maximum peak in the bottom profile of the lower half of the image and moves upwards by avoiding cutting until no further movement is possible.[1]

B. *Dissection with Contextual Post processing: Graphemes*

The segmentation obtained by dissection can be later subjected to evaluation based on linguistic context [1]. Markov model represents splitting, merging as well as misclassification in a recognition process. The system seeks to correct errors by minimizing an edit distance between recognition output, and words in a given lexicon and thus it merely tries to correct poorly made segmentation.

Another approach divides the input image into sub images that are not necessarily individual characters. The dissection is performed at stable image features that may occur within or between characters. The preliminary shapes, called graphemes are intended to fall into readily identifiable classes. A contextual mapping function from grapheme classes to symbols can then complete the recognition process. The mapping function may combine or split grapheme classes.

Techniques for dissecting cursive scripts are based on heuristic rules, there is no "magic" rule and it is not feasible to segment all handwritten words into perfectly separated characters in absence of recognition. Thus word unit resulting from segmentation are not only expected to be entire characters but also parts or combination of characters (graphemes).Moreover the relationship between characters and graphemes must remain simple enough to allow definition of an efficient post-processing stage- which means that a single character decomposes into at most two graphemes and conversely single grapheme represents at most two or three character sequence.

The line segments that form connection between characters in cursive script are known as "ligatures." Thus techniques which use "lower ligatures" connections near the baseline that link most lower case characters for segmenting can be used.

III. Recognition-based Segmentation Approach

Recognition based segmentation also segment words into individual units. Recognition based segmentation in effect bypass the requirement to discretely segment the word. No complex dissection algorithms is designed or implemented. This method directly interacts with the classifier [2]. In this method, whatever may be the content, a mobile variable width window is used which just divides the image into many overlapping part/pieces and the correct segmentation result is selected based on the recognition confidence. Therefore, the criterion for good segmentation is the recognition confidence given by the recognizer of the sub-image including syntactic or semantic correctness. This approach has also been called "segmentation-free" recognition.

In this method the recognition can be done either serially in which one by one every possibility is checked, while in parallel method features are compared to each letter. In serial case, recognition is done iteratively in a left to right scan of words, searching for a satisfactory recognition result. The parallel method proceeds in a global way. It generates a lattice of all possible feature-to-letter combinations. The final decision is found by choosing an optimal path through the lattice.

In this approach, two different methods can be employed:

1. Methods that make some search process
2. The method that segment a feature representation of the image.

A. *Methods that Search the Image*

Recognition based segmentation systems generally work as follows: First, the windowing is performed on the image to generate segmentation hypothesis. After this, the best hypothesis (guess) as determined by the classifier is chosen during verification step.

1) Recursive Segmentation: As explained above, traditional methods use windowing techniques that classify the character, based on a prototype character. The system exhaustively searches all possible cut points in the image until all characters are matched against a prototype library within a given threshold.

In recursive segmentation a sort of permutation combination is used for choosing the admissible/acceptable boundaries. The algorithm checks for all boundaries taking into consideration all the cut points. (Fig. 7). Thus in this method several segmentation are obtained by different combinations, but the acceptable segmentation is the one in which every segmented pattern matches a library prototype within a pre specified distance tolerance.

2) Shortest Path Segmentation: This method combines dynamic programming and neural net recognition for finding the best segmentation from the many obtained for the given word. As the name suggests, a graph is used in which nodes corresponding to acceptable segmentations exist. Moreover a path from one node to the other exists only if the two nodes i.e. the two segmentations are compatible. Thus we can say that these paths define the acceptable segmentations of a given word. Next the neural net is used to assign "distance" to node. The shortest path through the graph thus corresponds to the best recognition and segmentation of the word.

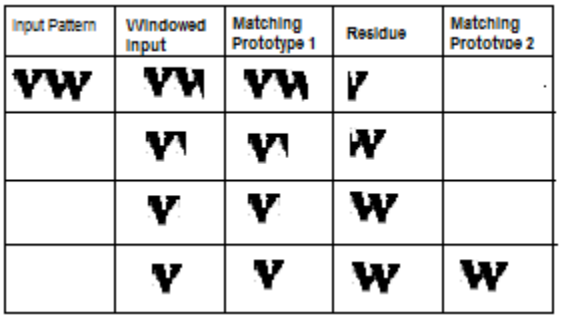

Figure 7: Recursive Segmentation. The example shows the results of applying windows of decreasing width to the left side of an input image. When the sub image in the window is recognized, then the procedure is recursively applied to the residue image. Recognition (and segmentation) is accomplished if a complete series of matching windows is found. In the top three rows, no match is obtained for the residue image, but successful segmentation is finally obtained as shown at the bottom.

3) Selective Attention Segmentation: The method of "selective attention" takes neural networks even further in handling of segmentation problems. In this approach (Fig 8), neural network seeks recognizable patterns in an image input, but is inhibited (reserved) automatically after recognition in order to ignore the region of the recognized character and search for new character images in the neighboring regions. Thus only a part of the image is used for recognition purpose, in case a match is found for that part, it is not considered in the next step and only the remaining image is segmented further and if that part is not recognizable then a part of the already recognized character is used.

Figure 8: Selective attention . (a) An input pattern. (b) The recognizer gradually reinforces pixels that corresponds to objects in its template library, and inhabits those that do not, yielding a partial recognition. (c) After a delay, attention is switched to unmatched regions, and another match to the library is found (after Fukishima).

B. Methods that Segment a Feature Representation of the Image

This method segments the image implicitly by classification of subsets of spatial features collected from the image as a whole. This method can be divided into two categories: Hidden Markov Model based approach and Non-Markov based approaches.

1) *Hidden Markov Models:* Hidden Markov model is widely used technique and can be considered a ubiquitous component in the current systems developed for recognizing machine print, online, offline handwritten data.

It is a stochastic model which characterizes the segmentation uncertainty, shape ambiguity and character transition information in a good way. Hidden Markov Model represents the variations in printing or cursive writing as probabilistic structure which is not directly observable. This structure consists of a set of states and the transition probabilities. The sequence of states is Markov chain because the choice of the next state to occupy depends on the identity of the current state. However this state sequence is not observable, only the symbol sequence generated by hidden states is observed. In addition, the observations that the system makes on an image are represented as random variables whose distribution depends on the state. This observation constitutes a sequential feature representation of the input image.

The Markov Models can be distinguished based on type of feature extraction. One of these Markov Models represents the state-to-state transition within a character, and these transitions provide a sequence of observations on the character. In this method, features are obtained in the left-to-right direction, and thus the words can be represented as combination/concatenation of character models. In this method segmentation is implicitly done while the model is matched against a given set of feature values gathered from a word image. Thus it decides where one character model leaves and the next begins, in the series of features analyzed.

For example, Fig. 9 shows a sample feature vector produced from the word "cat". This sequence can be segmented into three letters in many different ways, of which two are shown. The probability that a particular segmentation resulted from the word "cat" is the product of the probabilities of segment 1 resulting from "c", segment 2 from "a", etc. The probability of a different lexicon word can likewise be calculated. To choose the most likely word from a set of alternatives the designer of the system may select either the composite model that gives the segmentation having greatest probability, or else that model which maximizes the a posteriori probability of the observations, i.e., the sum over all segmentations. [1]

Perfect letter dissection is difficult to obtain as the letters do not always have a distinct/clear boundaries. But this problem can be compensated by HMM's as they able to learn by observing letter segmentation behavior on a training set.

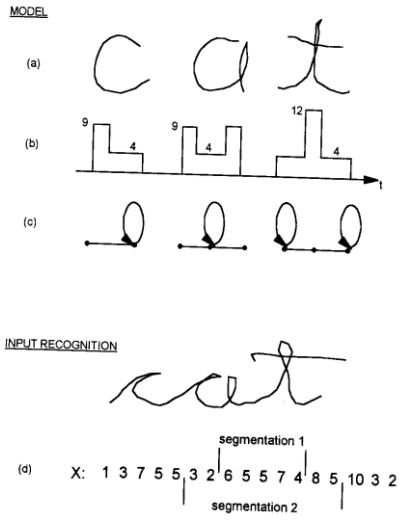

Figure 9: Hidden Markov Models example. (a) Training letters and (b) typical sequences of feature values obtained from (a). (c) The Markov models underlying (b), indicating the state sequence, and showing that certain states may be re-entered. Each state outputs a value from a feature distribution whose mean is indicated in the diagram above. The model for a word is the concatenation of such letter models. (d) A sequence of feature values obtained from a word, indicating several different segmentations. The HMM solution is found by evaluating many possible segmentation, of which two are shown.

These simple HMMs describing letters can be combined to form a single-path discriminant model or several model discriminant word HMMs.

In the single path-model, only one global model is constructed and the most likely path through this model gives the recognized word. The paths thus correspond to sequence of letters. These can handle large number of words, but their accuracy is quite less, which can be increased by incorporating lexical comparison modules.

In the model-discriminant HMMs, for every word one model is constructed, and checked to determine which is most likely to have produced a given set of observations. But this has limitation that large database is required and only the words existing in the system can be recognized.

2) Non-Markov approaches: Non Markov approaches stems from concepts used in machine vision for recognition of occluded objects. This family of recognition based approaches use probabilistic relaxation, the concept of regularities and singularities and similarities and backward matching [6].

In this method, in addition to features, their position of occurrence are recorded and used for segmentation purpose. Each feature indicates existence of one or more character at the position of occurrence. The positions are quantized into bins such that the evidence for each character indicated in a bin can be summed to give a score for classification. These scores are subjected to contextual processing using a predefined lexicon in order to recognize words. The method is being applied to text printed in a known proportional font. A method that recognizes word feature graphs is presented in [11]. This system attempts to match sub-graphs of features with predefined character prototypes. Different alternatives are represented by a directed network whose nodes correspond to the matched sub graphs. Word recognition is performed by searching for the path that gives the best interpretation of the word features. The characters are detected in the order defined by the matching quality. These can overlap or can be broken or underlined. A major drawback of this technique is that it requires intensive computation.

A top-down directed word verification method called "backward matching" (Fig. 10) is proposed in [14]. In cursive word recognition some characters can be easily recognized while some are quite difficult to guess. Thus all the letters have different discriminating power. So in this method unlike usual way of recognition, a "meaningfu order of scan is used/performed. This order mainly depends on visual and lexical significance of letters and follows edge-toward-center movement just like human vision.[13] Matching between symbolic and physical descriptions can be performed at the letter, feature and even sub-feature levels. As the system knows in advance what it is searching for, it can make use of high-level contextual knowledge to improve recognition, even at low-level stages. This system is an attempt to provide a general framework allowing efficient cooperation between low-level and high-level recognition processes.

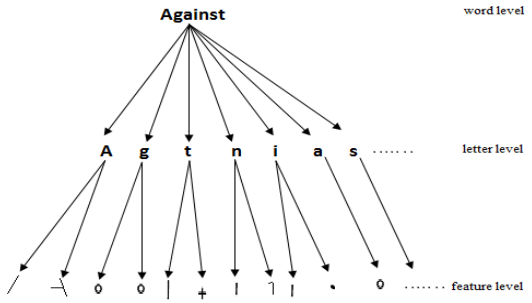

Figure 10: Backward matching. Recognition is performed by matching an image against various candidate words, the most distinctive letter being matched first. In this example, the algorithm first seeks the letter "T" in the image, then "t", "h", and so on. These letters are more informative visually because they contain ascenders, and lexically because they are consonants.

IV. HYBRID STATEGIES:"OVERSEGMENTING"

This segmenting strategy combines dissection and search methods in a hybrid way. In this method, dissection algorithm is applied to the image, but the objective here is not to get a single character or specific features but to "over segment", i.e. to cut the image in sufficiently many places so that the correct segmentation boundaries are included among the cuts made(Fig. 11). Next the subset of segment obtained can be evaluated to get the optimal one. Now each of this subset is nothing but segmentation hypothesis which can be further evaluated by classifier to get most promising segmentation.[6]. Thus some set of potential cuts is generated first from which segmentation hypothesis are obtained and evaluated with the help of classifier [7]. When

the number of characters in the image to be dissected is not know a priori, or if there are many touching characters, e.g., cursive writing, then it is usual to generate the various hypotheses in two steps.

In the first step, a set of likely cutting paths is determined, and the input image is divided into elementary components by separating along each path. In the second step, segmentation hypotheses are generated by forming combinations of the components. All combinations meeting certain acceptability constraints (such as size, position etc) are produced and scored by classification confidence. An optimization algorithm, typically implemented on dynamic programming principles and possibly making use of contextual knowledge, does the actual selection.

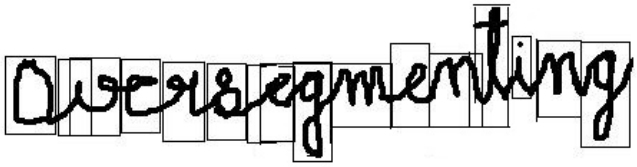

Figure 11: Oversegmenting. Note that letters that contain valleys have been dissected into multiple parts. However, no merged characters remain, so that a correct segmentation can be produced by recombination of some of the segments.

It is also possible to carry out an over-segmenting procedure sequentially by evaluating trial separation boundaries. In this work a neural net was trained to detect likely cutting columns for machine printed characters using neighborhood characteristics. Using these as a base, the optimization algorithm recursively explored a tree of possible segmentation hypotheses. The left column was fixed at each step and various right columns were evaluated using recognition confidence. Recursion is used to vary the left column as well, but pruning rules are employed to avoid testing all possible combinations.

V. HOLISTIC APPROACH

Holistic approach in handwritten word recognition treats the word as a single, indivisible entity and attempts to recognize it using features of the word as a whole. In this method, the entire word is considered to be a single entity and the total features of a word are used for recognition purpose. No attempt to analyze the letter/part of the whole word is made for recognition of words.

These techniques are derived for recognition of the entire word without attempting to analyze the letter content of the word. here recognition is based just on extraction and comparison of a collection of simple features extracts which describe the entire word such as strokes, holes, arcs ascenders, descenders etc against a lexicon of codes representing possible shapes. [12].

This approach has limited usage, i.e. it can be applied under certain pre-specified conditions only: 1. The number of words to be recognized is small and are pre-specified. 2. When this method is to be used only to reduce the lexicon size by eliminating obvious mismatches, thereby facilitating more accurate but computationally more intensive technique to be used for final word recognition.

Though holistic approach is different compared to classical approach, it follows typically the same scheme: First It performs feature extraction which is based on determination of middle zone of the words and next ascenders and descenders are found by considering the part of writing exceeding this zone to create representation of the word. Next global recognition is performed by matching the representation of word with a representation of the world. [2]

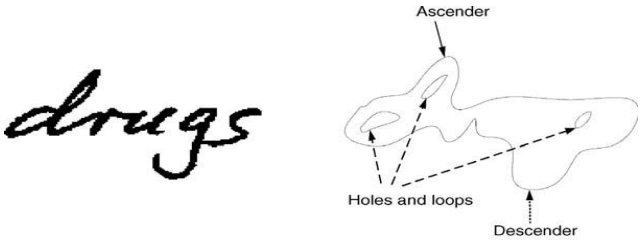

Figure 12. A word image (a) and its shape features (b) including “length”, “ascenders”, ”descenders”, ”loops” and so on.

ACKNOWLEDGMENT

We are highly indebted guide Mrs Kruti J Dangarwala or their guidance and constant supervision as well as for providing necessary information. An earlier a survey on online as well as offline methods of segmentation was presented by Richard G Casey and Eric Lecolnet which has been of great use for this survey.

REFERENCES

[1] Richard G. Casey and Erics Lecolinet, “A rvey of Methods and Strategies in Character Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, July 1996.

[2] Kurt Alfred Kluever, “Independent Study Report Character Segmentation and Classification,” Department of Computer Science, Golisano College Of Computer and Information Sciences, Rochester Institute of Technology, Februray 28, 2008.

[3] Richard L. Hoffman and J. Warren McCullough, ”Segmentation Methods for Recognition of Machine-printed Characters,” In edManics of Materials, IBM Corporation, March 1971.

[4] M Swamy Das ,Dr CRK Reddy, Dr A Govardhan and G Saikrishna. “Segmentation of Overlapping Text Lines, Characters in Printed Telugu Text Document Images.” *International Journal of Engineering Science and Technology*, vol.2(11), 2010, 6606-6610.

[5] R.G. Casey and G. Nagy, ”Recursive Segmentation and Classification of Composite Patterns”, Proc. Sixth Int’l Conf. Pattern Recognition ,p. 1,023, 1982.

[6] Robert Burduk, Marek Kurzynski, Michal Wozniak, Andrzej Zolnierek, *Computer Recognition Systems 4*.

[7] Nfiz Arica, “An Offline Character Segmentation System for Free Style Handwriting.”

[8] Kurt Alfred Kluever, “Independent Study Report Character Segmenation and Classification”, Golisano

College of Computing and Information Sciences Rochester Institue of Technology, Feburary 28,2008

[9] Zaidi Razak, Khansa Zulkiflee, Noorzaily Mohamed Noor, Rosli Salleh, Mashkuri Yaacob, "Off-line Handwritten Jawi Character Segmentation using Histogram Normalization and Sliding Window Approach for Hardware Implementation," University of Malaya, 50603 Kuala Lumpur, Malaysia.

[10] A. M. Zeki, "The Segmentation Problem in Arabic Character Recognition The State Of The Art", in First International Conference on Information and Communication Technologies, ICICT, 27-28 Aug. 2005, pp. 11-26.

[11] J. Rocha and T.Pavlidis, "New Method for Word Recognition Without Segmentation," *Proc. SPIE charcter Recognition Technologies*, vol. 1,906,pp.74-80,1993.

[12] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Cheng –Lin Liu, Ching Y. Suen, *Character Recognition Sytem, A guide for students and practitioners*, Wiley.

[13] I. Taylor and M. Taylor, "Psychology of Reading", Acedamic Press, 1983.

[14] E. Lecolinet," A New Model for Context Driven Word Recognittion ,"*Proc. Symp. Document Analysis and Information Rerival*, Las Vegas, p.135, Apr. 1993.