A Survey on Touching Character Segmentation

Alok Kumar, Madhuri Yadav, Tushar Patnaik, Bhupendra Kumar

Abstract*—Character segmentation is an important step of Optical Character Recognition (OCR) system. The segmentation of touching symbols is one of the key factors which affect the performance of recognition system. Therefore, to make OCR systems more effective and accurate, segmentation of touching characters is an important task. This paper explains the concepts of touching characters and presents the survey of various approaches for touching character segmentation.

Keywords*— Optical Character Recognition systems, Recognition rate, Segmentation, Touching Characters.

I. INTRODUCTION

A Touching Character is the connected component which is produced when the adjacent characters touch each other. The segmentation of touching characters means separation of the connected components into individual characters which can be recognized by OCR system.

The touching characters can be produced due to presence of noise in images, people writing habits and other factors [1]. The touching characters can be found in rich text documents, old books due to low ink quality and aging, old newspapers, and other literature works of ancient times.

An OCR system is developed to convert the scanned documents into editable form. For this conversion, the text contained in the scanned document is recognized by segmenting it into individual line, words, and characters. After that, individual characters are recognized by extracting their features from segmented image [2].

Touching character patterns emerges when some parts of character are connected horizontally/left-right or vertically/ up-down. Particularly, for Devnagari, Bangla and numerals, the touching components are generated when two adjacent symbols are connected horizontally/ left-right. Vertically/ up-down touching characters are found in Chinese language.

The following figure shows the example of various touching Characters found in different scripts [7] [3] [6] [2]:

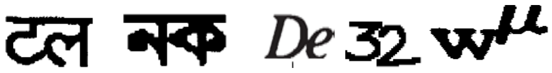

Figure1: Example of touching characters

Manuscript published on 28 February 2013.
***** Correspondence Author (s)
Alok Kumar, M.Tech(CS),School Of Information and Technology, GGSIPU, CDAC Noida, India.
Madhuri Yadav, M.Tech(IT), School Of Information and Technology, GGSIPU, CDAC Noida, India.
Tushar Patnaik, Sr. Lecturer/Sr. Project Engineer, School Of Information and Technology, GGSIPU, CDAC Noida, India.
Bhupendra Kumar, Senior Technical Officer School Of Information and Technology, GGSIPU, CDAC Noida, India.

The Touching Character Segmentation approaches can be classified basically into three parts namely, Recognition-free, Recognition-based and hybrid approaches.

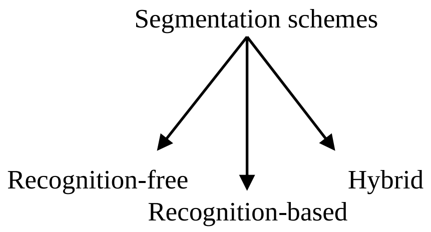

Figure2: Segmentation schemes

In Recognition-free segmentation a touching component can be divided into segments by rules without recognition.

In Recognition-based segmentation the candidate segmentation points are verified with recognizer. Hybrid approach is the combination of above two approaches.

This paper discusses various techniques based on the above segmentation schemes and tries to provide the review of the work accomplished by various researchers working in this field.

The rest of the paper is structured as follows. Section 2 gives the description of various approaches undertaken by researchers. Section 3 compares the results of different approaches in tabular format. The last section derives the conclusion and references.

II. APPROACHES FOR SEGMENTATION OF TOUCHING SYMBOLS

Dong-Yu Zhang, Xue-Dong Tian, and Xin-Fu Li [3], they proposed an approach for segmentation of touching symbols in printed mathematical expressions. Their approach follows three basic steps which can be listed as follows:

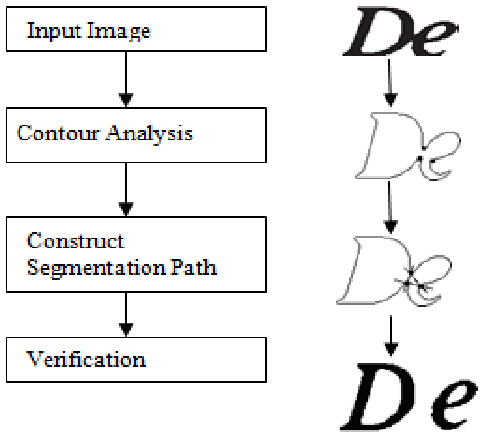

Figure3: Flowchart of proposed approach along with output of each stage

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

(a) Detection of touching symbols:
This stage detects those symbols whose recognition distance is larger than a predefined threshold and these symbols are called as touching symbols.
(b) segmentation :
This stage first performs Counter Analysis and then constructs the segmentation path. During counter analysis, concave corner points on the outer counter of touching symbols are detected. After the detection of all concave corner points, connection line joining points on outer counter are considered as candidate segmentation paths.
(c) Verification:
This stage segments the touching symbol image by each segmentation path.

After segmentation, if the recognition distances of all connected components are less than the threshold, the recognition result is adopted otherwise the original image is considered again and worked upon with different segmentation paths.

Salman Amin Khan [6], proposed an approach which works on segmentation of handwritten numeral strings. The author worked on the concept of "drop fall algorithms". Drop falling algorithms build segmentation path by mimicking an object falling or rolling in between two connected characters.

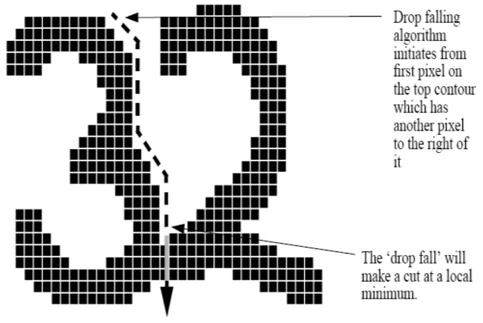
Drop falling algorithm initiates from first pixel on the top contour which has another pixel to the right of it
The 'drop fall' will make a cut at a local minimum.

Figure4: Drop-fall cut

There are four variants of drop-fall algorithms:

- Top-left (Left Descending)
- Top-right (Right-Descending)
- Bottom-left (Left-Ascending)
- Bottom-right (Left-Descending)

To detect the starting point of drop-fall algorithm, the pixels are horizontally scanned until a black boundary pixel with another black boundary pixel to right of it, is detected.
The figure shows the initial point for the above mentioned process:

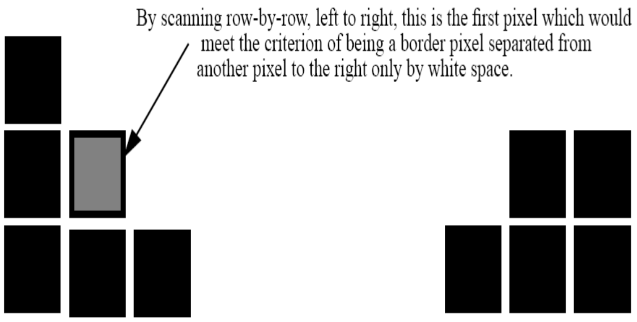
By scanning row-by-row, left to right, this is the first pixel which would meet the criterion of being a border pixel separated from another pixel to the right only by white space.

Figure5: Initial pixel position

After determination of initial pixel, the next step is to begin the actual drop fall. The drop-falling algorithm is designed to mimic falling, so it will always move downwards, diagonally downwards, to the right, or to the left. The directions that the algorithm will 'move' in according to the current pixel position and its surroundings are illustrated through figure:

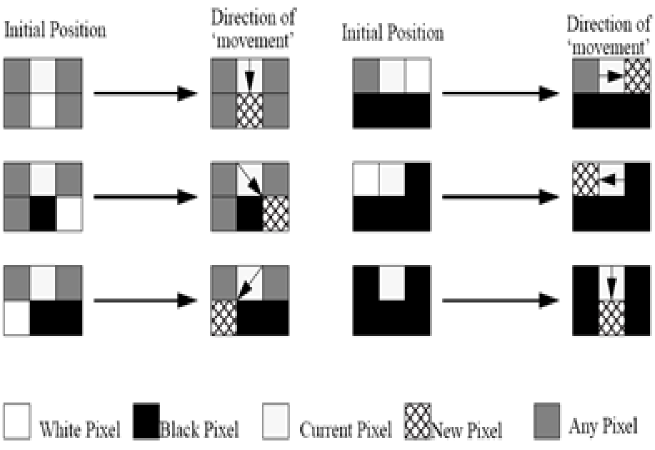

Figure6: Movement of drop-fall algorithm

The top-left variant of algorithm failed for the samples where a gradual and flat minimum was found. The figure illustrates the success and failure examples of top-left algorithm:

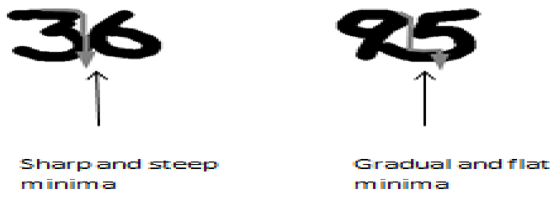

Figure7: Example of top-left drop-fall success and failure

Thus, for successful segmentation bottom-right drop fall was applied.

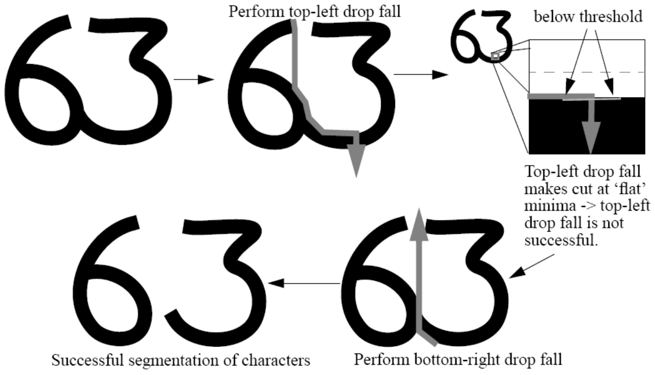
Perform top-left drop fall
below threshold
Top-left drop fall makes cut at 'flat' minima -> top-left drop fall is not successful.
Successful segmentation of characters
Perform bottom-right drop fall

Figure7: Performing bottom-right drop fall when top-left drop-fall is recognized to fail

U. Pa1, A. Belai and C. Choisy [8], they introduced an approach based on concept of Water Reservoir, for segmentation of unconstrained handwritten connected numerals.

When two numerals touch each other they create a large space between numerals, if water is poured from top of a connected numeral than water will be stored in this large space. This water filled large space is called Reservoir. Two types of reservoirs can be obtained: top and bottom reservoir.

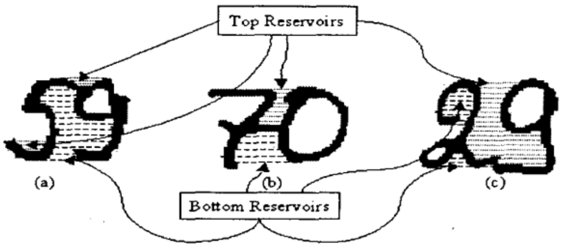

Figure8: Type of reservoirs

Numerous reservoirs may be formed but only those reservoirs will be considered whose heights are greater than threshold T_1. The value of threshold is 1/8 of the height of the numeral.

In the first stage, the input numerals are identified as connected or isolated numerals. The classification is done based on the number of reservoirs, their size and positions and number of close loop and their location. If the numerals are isolated then the process is aborted otherwise, for connected numerals the position where they touch each other is determined.

For touching position detection the touching component is enclosed in a bounding box and the bounding box area is divided into three regions horizontally as well as vertically. The three regions are top, middle, bottom covering areas 25% of bounding box (bb), 50% of bb and 25% of bb respectively. Similarly in vertical direction with the same coverage areas.

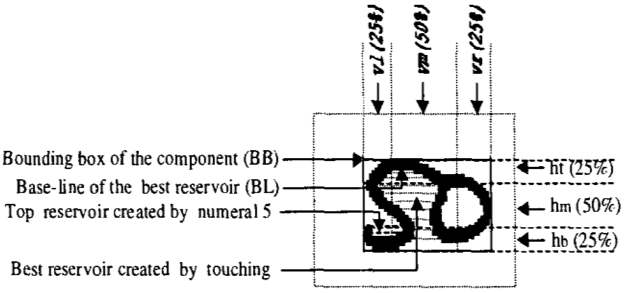
vl (25%)
vm (50%)
vr (25%)
Bounding box of the component (BB)
Base-line of the best reservoir (BL)
Top reservoir created by numeral 5
Best reservoir created by touching
ht (25%)
hm (50%)
hb (25%)

Figure10: Feature Detection

The above figure shows the various regions of the touching component. At first, the largest reservoir of the component whose center of gravity lies in v_m region is considered. This reservoir is called as the *best reservoir* for touching. The *base-line* of the best reservoir is detected. Depending upon the region in which this base-line lies the position of touching is determined i.e. whether the base-line is in h_t, h_m, or hb.

According to touching position the reservoirs are selected, and their feature points are extracted.

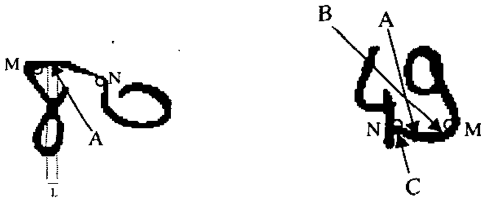

Figure11: Initial feature points (A, B, C, D) and node points (N, M). A is the best feature point. 'L is length of base-line.

The leftmost and rightmost points of the base-line of considered reservoirs are the feature points. These points are initial feature points. If the distance (L) between the leftmost and rightmost points of a base-line is less than 2R then instead of two feature points (leftmost and rightmost points of base line) the author considered the midpoint of leftmost and rightmost points as an initial feature point. The confidence value is determined for each feature point and the feature with highest value is selected for further processing. Further, the segmentation path is determined depending whether the components have two side by side and touching close loops or other touching components. For the first case, segmentation is done through the middle of their common touching area. For the top and bottom touching components vertical segmentation is done.

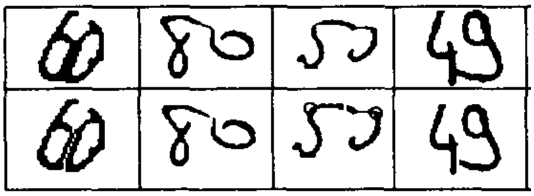

Figure12: Segmentation Results

The segmentation scheme was tested on 978 images of numeral string of French bank check courtesy amount. The results were verified manually and observed that 94.34% of the connected numerals were correctly segmented.

Utpal Garain and Bidyut B. Chaudhuri [7], purposed an approach for identification and segmentation of touching characters for Printed Devnagari and Bangla Scripts. The technique is based on fuzzy multifactorial analysis. A predictive algorithm is developed for effectively selecting possible cut columns for segmenting the touching characters. For identification of touching character following two factor are used:-

- Measure of dissimilarity.
- Aspect ratio.

Mathematically, *fmd is represented by:* $fmd=1-d_{off}/d$ where d is the minimum similarity distance for a target character against a set of stored prototypes, and d_{off} is the offset distance used by the character classifier (which is the character for which d less than d_{off} is accepted or otherwise rejected).

The second factor *far* is defined as *far =e^a/(1+e^a)* where *a = w/h*, w and h are the width and height of the minimum up-right rectangular bounding box of a character.

For identification of cut column following five factors are used:-

1. fic: inverse crossing-count = 1/c, where c is the vertical crossing count.
2. Fdm: degree of middleness =min (l1, l2)/max (l1, l2).
3. Fmt: measure of blob thickness = 1-t / T.
4. The fourth factor Fup=Wu+ [(1-wu)*Au] ^Bu.
5. the fifth factor flow= Wl+[(1-wl)*Al]^Bl

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

To determine the most favorable cut column, the column with highest (f) value is considered. The cut column is the column from where the characters are separated from each other. These separated characters are then sent to classifier and based on its decision either the algorithm is terminated or next cut-column with second highest (f) value is considered for the cut-position.
Author reported that by using 11577 and 16714 touching characters for Devnagari and Bangla script respectively, an overall 98.87% (for Devnagari) and 98.63% (for Bangla) accuracy in identifying the touching characters was achieved. The touching character segmentation accuracy was 98.92% and 98.47% accuracy Devnagari and Bangla scripts.

Neena Madan Davessar, Sunil Madan, Hardeep Singh [9], proposed an approach for machine printed Gurumukhi documents. The approach works for segmentation of pair as well as triplet of touching characters. The approach proceeds as follows:

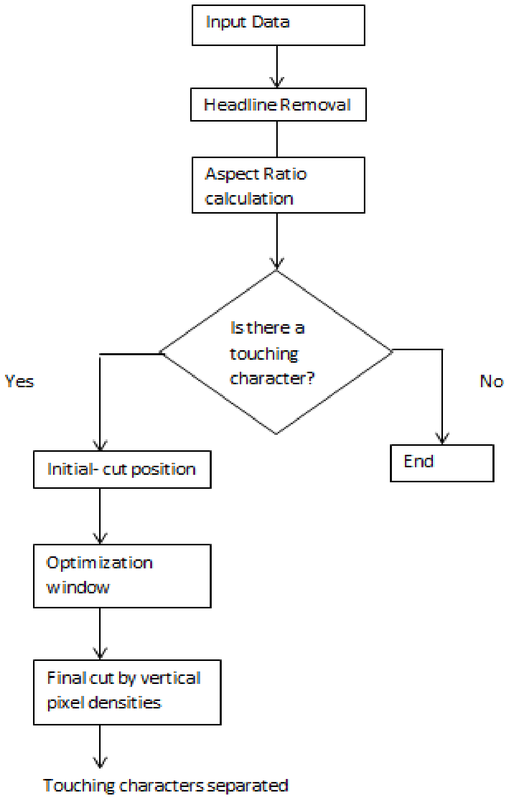

Figure13: Flowchart of proposed approach

The first step is headline removal. After headline removals touching characters are not separated as there are no vertical white spaces. Hence, aspect ratio technique is used and it is calculated for every character. If its value is greater, it means that the width of character is greater than the length of the character and thus that particular character is good candidate for touching character. The initial-cut is made in the half of touching character pair.

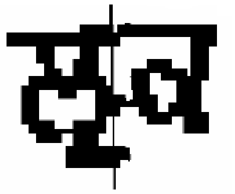

Figure14: Example of initial cut.

For computation of optimal-cut point an optimizing window is used. An optimizing window is the constant width window around approximate cut-point.

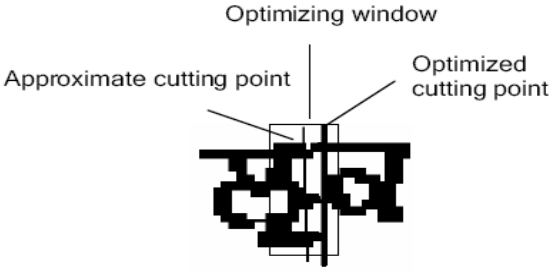

Figure15: Example of optimizing window for a pair

For triplets, the initial cut-position is computed in the middle of the touching character. Further, two optimized windows are used around approximate cut-point.

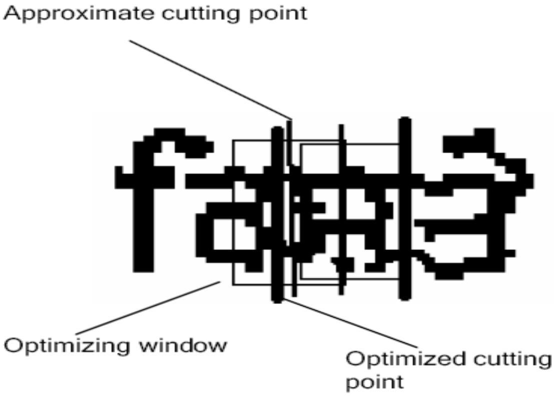

Figure16: Example of optimizing window for a triplet

After construction of optimization window vertical pixel densities are computed for each column in that window. The final cut is made at the position which has minimum vertical pixel density.
Utpal Garain and B. B. Chaudhuri [2], proposed an approach for mathematical expressions on multifactorial analysis was proposed. This approach segments the touching characters in two simple steps:
(a) Evaluates the four different factors.
(b) Confirmation of cut- column positions.
In first stage, four different factors are defined for four directions (vertical, horizontal, +45^0, -45^0). In each direction the factors are evaluated for each object (i.e. rows for horizontal, columns for vertical). The above calculated factors are combined for finding appropriate cut position in each direction. Thus, a single value (f) is obtained for multiple factors for each direction.
In the second stage, to reduce the computation time of classifiers they determine the favorable cut positions before character classifier is used to confirm it. To determine the most favorable cut column, the column with highest (f) value is considered. The cut column is the column from where the characters are separated from each other.

These separated characters are then sent to classifier and based on its decision either the algorithm is terminated or next cut-column with second highest (f) value is considered or the cut-position is determined in remaining directions.

This proposed technology has efficiency of 64.78% and also improved computational efficiency of classifier.

U.K.S. Jayarathna and G.E.M.D.C. Bandara [4], proposed a junction based segmentation algorithm for offline handwritten connected two digit strings. Their approach focuses on removal of unwanted connected segments from the characters. The various stages of the proposed approach are shown below:

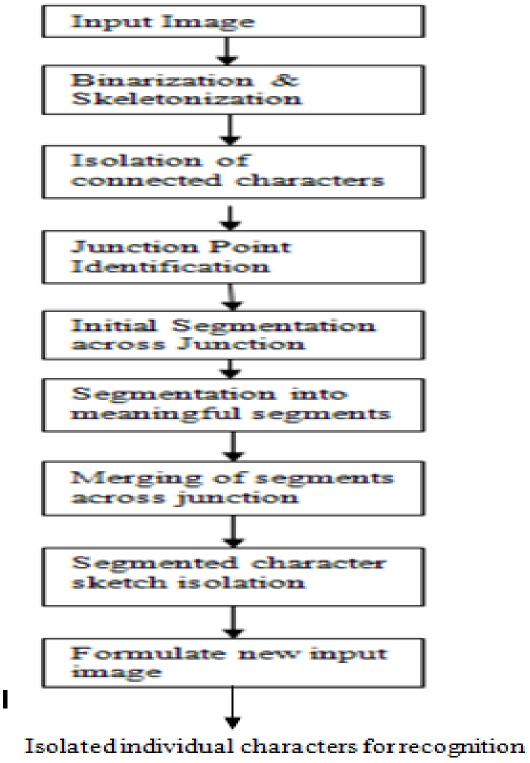
Input Image
Binarization & Skeletonization
Isolation of connected characters
Junction Point Identification
Initial Segmentation across Junction
Segmentation into meaningful segments
Merging of segments across junction
Segmented character sketch isolation
Formulate new input image
Isolated individual characters for recognition

Figure17: Flowchart of proposed approach

After Skeletonization, the connected character components are isolated in correlation area. A *correlation area* is a rectangular area in the image, which contains connected character skeleton.

Segmentation stage of the proposed work consists of two phases namely, the initial segmentation phase and the total segmentation phase. In the initial segmentation, Junction Points of the character image are identified and Junction Based Segmentation is done. A Junction Point is a pixel point in the Correlation area, having three or more neighboring pixels. After the identification of the all junction points in the correlation area, they are used to segment the connected character skeleton in to initial segments. In the total segmentation phase a separate segmentation algorithm is applied which is proposed in [5].

Merging of each segment across junction is proposed, to avoid unnecessary segmentation and to reconstruct meaningful segments.

In this research, each of the segments was named with respect to its participating junction point number. Junction point numbers are unique integers assigned to the junction points in correlation area. The segments having same starting or ending junction points are merged together in a group depending upon their fuzzy values computed using [5]. All successfully merged meaningful segments are placed in a group called major segment group and all the others are placed in a group called minor segment group. It was found that the minor segment group contained an unnecessary connection between individual characters.

In this work, each of the segments in the major segment group is used to formulate new input character image by selecting segments from the minor segment group. This method is used to create input character image with several correlation areas. For the recognition, the work proposed in [5], is used to identify each individual character by sending each correlation area to the isolated character identification process. In this work, the correlation area, which only produces successful two characters output, is considered as the separated characters output of the original connected character input.

100% accuracy has been reported for machine printed characters, which have unwanted connection. Although the method used in this work does not deal with the two stroke connections from two adjacent digits touching end to end.

III. COMPARISON OF VARIOUS APPROACHES

Sr. No	Author	Work done on	Concept used	Data Set(images)	Results (%)
1.	Dong-Yu Zhang[3]	Printed Mathematical Expressions	Concave corner points	-	-
2.	U.K.S. Jayarathna[4]	English Handwritten uppercase characters	Junction points	-	-
3.	Utpal Garain and B. B. Chaudhuri [2]	Scanned mathematical expressions	Multi-factorial analysis	200	64.78
4.	Salman Amin Khan [6]	Handwritten numeral strings	Drop-falling process	-	91
5.	U.pal[7]	Printed Devnagari and Bangla scripts	Fuzzy Multi-factorial analysis	11,577 (Devnagari) 16714 (Bangla)	98.87 (Devnagari) 98.63 (Bangla)
6.	U.pal[8]	Handwritten numerals	Water Reservoir	978	94.34

IV. CONCLUSION

The above survey concludes that remarkable work has been done for printed touching characters but more research work is required for hand-written touching characters. Finally, it is hoped that this comprehensive discussion will clarify the approaches and the various concepts involved in them, and will provide further guidelines to naïve researchers.

ACKNOWLEDGMENT

The authors are thankful to the referees for their critical comments. We are also thankful to Mr. Ranjeet Srivastva (M.Tech-IT at C -DAC Noida) for his helpful suggestions.

REFERENCES

[1] G. Congedo, G. Dimauro, S. Impedovo, G. Pirlo, “Segmentation of Numeric Strings”, 1995, IEEE, p.p- 1038-1041

[2] Utpal Garain and B. B. Chaudhuri , “Segmentation of Touching Symbols for OCR of Printed Mathematical Expressions: An Approach based on Multifactorial Analysis”, Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR’05), ,IEEE , 2005.

[3] Dong-Yu Zhang, Xue-Dong Tian, and Xin-Fu Li “An Improved Method for Segmentation of Touching Symbols in Printed Mathematical Expressions”. 2nd International Conference on Advanced Computer Control (ICACC), vol.2, pp. 251 - 253, 2010

[4] U.K.S. Jayarathna G.E.M.D.C. Bandara, “A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation”, International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06),IEEE, 2006.

[5] K. B. M. R. Batuwita, G.E.M.D.C. Bandara, “An online Adaptable fuzzy system for offline handwritten Character recognition”, Proceedings of 11th World Congress of International Fuzzy Systems Association (IFSA 2005), Beijing ,China, Springer-Tsinghua, 2005, Vol. II, p.1185-1190.

[6] Salman Amin Khan “Character Segmentation Heuristics for Check Amount Verification”

[7] U. Garain and B.B. Chaudhuri. “Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis.” IEEE Transaction on Systems, Man and Cybernetics,Vol. 32(4), 449-459, 2002.

[8] U. Pa1, A. Belai and C. Choisy, “Water Reservoir Based Approach for Touching Numeral Segmentation”.2001

[9] George Nagy, Thomas A. Nartker, Stephen V. Rice, Optical Character Recognition: “An illustrated guide to the frontier”, Procs. Document Recognition and Retrieval VII, SPIE Vol. 3967, 58-69.

Mr. Alok Kumar pursuing M.Tech from CDAC, Noida. He received his B.tech degree from C.R. state Engineering college Murthal. He is working on the project “Touching Character Segmentation”. His areas of interest are Digital image processing, Computer Network, Computer Security and Theory of Computation.

Ms.Madhuri Yadav pursuing M.Tech from CDAC, Noida. She received her B.tech degree from Maharishi Dayanand University, Rhotak. She is working on the project “Offline Signature Verification”. Her areas of interest are Digital image processing, Software Testing, Computer Network, Computer Security and Theory of Computation.

Mr. Tushar Patnaik (Sr. Lecturer/Sr. Project Engineer) joined CDAC in 1998. He has eleven years of teaching experience. His interest areas are Computer Graphics, Multimedia and Database Management System and Pattern Recognition. At present he is leading the consortium based project “Development of Robust Document Image Understanding System for Documents in Indian Scripts”.

Mr. Bhupendra Kumar (Senior Technical Officer) joined CDAC in 2005, he received his M.Tech. degree from IIIT Allahabad with the specialization in wireless communication and computing. His interest area are Advanced Image processing, pattern recognition, computer network, wireless network, MANETs. Currently he is involved in project “Development of Robust Document Image Understanding System for Documents in Indian Scripts”.