A Study of Different Kinds of Degradation in Printed Gurmukhi Script

M. K. Jindal
Department of Computer Applications
Panjab University Regional Centre, Muktsar Punjab India.
manishphd@rediffmail.com

R. K. Sharma
School of Mathematics and Computer Applications
Thapar Institute of Engineering and Technology, Patiala, Punjab, India.
rksharma@tiet.ac.in

G. S. Lehal
Department of Computer Science and Engineering, Punjabi University, Patiala Punjab, India.
gslehal@gmail.com

Abstract

The performance of any OCR system heavily depends upon printing quality of the input document. Many OCRs have been designed which correctly identify fine printed documents both in Indian and foreign scripts. But little reported work has been found on the recognition of the degraded documents. The performance of standard machine printed OCR system working for fine printed documents decreases, if it is tested on degraded documents. The degradation in any document can be of many types. In this paper, we have identified different kinds of degradation available in printed Gurmukhi script. After identifying the different kinds of degradation, problems associated with each kind of degradation have been discussed; some possible solutions have also been discussed. This paper is extremely useful for researchers engaged in recognizing the degraded documents in any script, because same kinds of degradation can be found in most of the scripts of the world.

1. Introduction

System for recognizing high quality machine-printed text can recognize words at a high level of accuracy [1]. However, given a degraded text page, performance usually drops significantly. For an OCR, which generates too many errors in recognizing a degraded document, it will be more efficient to type the contents of the entire page into the computer than to correct the errors in the OCR output. It is surprising that after more than 45 years of research, OCR systems are not close to matching human performance. Current accuracy of the machine-printed characters is easily higher than 99.90%, which is sufficient for many real applications. However, there are several applications where the image quality is poor and standard OCR is not able to reach sufficient recognition capability. This paper outlines areas in which today's systems can be improved. In this paper we have identified different kinds of degradation in printed Gurmukhi script.

A number of algorithms have been proposed in the past to recognize the touching characters. Segmentation of touching characters is major problem during recognizing touching characters. Lee *et al.* [2] have segmented the touching characters using projection profiles and topographic features extracted from the gray scale images. Bose and Kuo [3] used Hidden Markov Model to recognize the touching and degraded text. Tsujimoto and Asada [4] constructed a decision tree for resolving ambiguity in segmenting touching characters. Casey and Nagy [5] proposed a recursive segmentation algorithm for segmenting touching characters. Hong [6] has utilized visual inter-word constraint available in a text image to split word images into pieces for segmenting degraded Roman script characters. Kahan *et al.* [7] have proposed a very useful double differential function to segment the touching characters. Lu [8] proposed a generalized differential technique for the purpose by mapping vertical projection on to second projection called peak-to-valley ratio.

Few algorithms have been investigated on segmenting the touching characters in Indian scripts [9-14]. Bansal and Sinha [9] have segmented the conjuncts (one kind of touching patterns) in Devanagari script using the structural properties of the script. Garain and Chaudhuri [10] have used a technique based on fuzzy multifactorial analysis to segment the touching characters in Devanagari and Bangla scripts. Chaudhuri *et al.* [11] have used the principle of water overflow from a reservoir to segment the touching characters in Oriya script. Jindal

et al. [12] has used the structural properties for segmenting the touching characters in middle and upper zone of printed Gurmukhi script. Lehal and Singh [13, 14] have also tried to segment the touching characters in upper zone of Gurmukhi script.

Not very much work has been reported to be done to recognize the broken characters. Lu [8] has discussed two methods of segmenting the broken characters, first by employing a merging procedure based on the estimated character width and intervals, and second is to combine character components based on the recognition results. Lu *et al.* [15] proposed an algorithm based on estimation procedure, a sequential merging procedure, as grouping procedure based on the estimated character width, and a decision procedure. Nakamura *et al.* [16] and Okamoto *et al.* [17] proposed a character segmentation algorithm based on propagation and shrinking in vertical and horizontal directions. Droettboom [18] used a technique based on graph combinatorics to rejoin the appropriate connected components. Yanikoglu [19] has estimated the pitch of the text and guided the segmentation of broken characters by the location of the pitch window, defined by the estimated pitch and the offset.

Oguro *et al.* [20] have proposed three step solutions for restoring faxed document by producing gray level images. Natarajan *et al.* [21] have used Hidden Markov Model for recognizing fax degraded documents.

Cannon *et al.* [22, 23] have suggested a method for automatically improving the quality of degraded images in a typewritten archive. Rodríguez *et al.* [24] have developed a new cost function to segment degraded typewritten digits.

To the best of our knowledge, there is no reported work done in recognition of broken characters, faxed documents and typewritten documents for any Indian script. So there is wide scope to work in this area. Properties of Gurmukhi script characters have been discussed in [12].

2. Different Kinds of Degradations

We have identified different kinds of degradation available in printed Gurmukhi script. We scanned about 250 documents from different sources. The sources of each kind of degradation, the problems associated with each kind of degraded text, comparison of each kind of degraded text in Gurmukhi with corresponding degraded text in Roman script, some possible solution for identifying each kind of degraded text have been discussed in this section.

2.1. Touching Characters

This is the most commonly found degradation in printed Gurmukhi script. In this category of degraded text, two neighboring characters touch each other. The biggest issue involved in recognition of touching characters is to segment them correctly, i.e. identifying the position at which the touching pair of characters must be segmented. Every OCR must perform well to the sensitive task of separating the touching characters. The accuracy of any OCR depends heavily upon the accuracy of segmentation process. The sources of touching character documents are magazines with heavy printing, newspapers printed on low quality paper, very old books whose pages turn to be yellow due to aging, Photostatted documents copied on low quality machine etc..

Figure 1. Words containing touching characters in printed Gurmukhi script.

Occurrence of the touching characters in any document drastically decreases the performance of an OCR. On statistical analysis of the touching characters, we made following observations [10]:

1. Touching characters are found in all the three zones of the document line, i.e. upper, middle and lower zone.
2. The touching characters touch each other mostly at the center of the middle zone, less frequently at top of the middle zone and very less at the bottom of the middle zone.
3. Most of the time touching characters have greater aspect ratio than that of single individual characters.
4. Generally in a single word two characters touch each other. The possibility of more than two touching characters in single word is less.
5. Generally the vertical thickness of the black blob at the touching position is small as compared with the thickness of the stroke width. But in some cases thickness may equal or greater than the stroke width.
6. Indian scripts characters contain sidebars in the characters at the right end of the character, *e.g.*, in Gurmukhi script 12 consonants have sidebars at the right end of the character. The possibility of touching is increased at this position.

Before recognition, segmentation of the touching characters is most challenging task. There are two key issues involved in this problem. The first issue is to

find the candidate of segmentation, i.e. the segment of the complete word, which may contain the touching characters. Second issue is to find the break location within the candidate of segmentation, i.e. the column, which will correctly segment the two touching characters into isolated characters.

The problem of segmenting the touching characters in Gurmukhi script is very much different from the Roman script in many aspects:

1. In Gurmukhi script touching characters can be found in upper, middle and lower zone. Further the touching characters can be divided into 5 categories:
   (a) Upper zone characters touching with each other (as shown in figure 2(a)).
   (b) Upper zone characters touching with middle zone characters (as shown in figure 2(b)).
   (c) Middle zone characters touching with each other (as shown in figure 2(c)).
   (d) Middle zone characters touching with lower zone characters (as shown in figure 2(d)).
   (e) The lower zone characters touching with each other (as shown in figure 2(e)).
   But in Roman script there is no concept of upper, middle and lower zone.
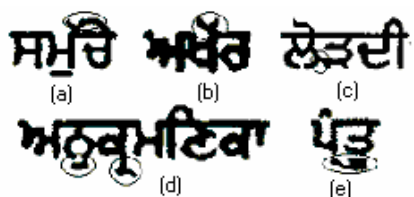
Figure 2. Touching characters in three zones (touching characters encircled): (a) upper zone characters touching each other, (b) upper zone characters touching with middle zone characters, (c) middle zone characters touching with each other, (d) middle zone characters touching with lower zone characters, (e) lower zone characters touching with each other.

2. Generally the shape of the two touching characters in Gurmukhi script is very much different from any basic character, in contrary to Roman script where the combined shape of "r" and "n" makes "m".

3. In Gurmukhi script the tendency of touching of middle zone characters is more in middle of the characters and less at upper and bottom of the characters, but in Roman script it is more at upper and bottom of the characters and less in the middle of the characters.

4. Due to presence of the headline, characters are always connected with neighboring characters in Gurmukhi script as shown in figure 1, which does not happens in Roman script since the concept of headline is not present in Roman script.

5. As shown in figure 3, in Gurmukhi script the problem of multiple horizontally overlapping lines exists. Also existence of small sized strips containing only upper/lower zones makes the problem of line segmentation more complicated [12]. This problem is rarely found in Roman script.
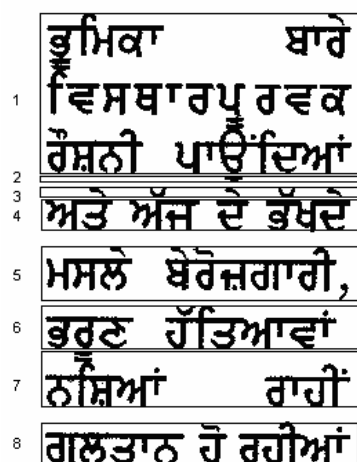

Figure 3. Different strip lines in printed Gurmukhi script.

For segmenting the touching characters in Gurmukhi script the authors have proposed a solution [12]. The algorithms are proposed for segmenting the touching characters in all three zones, i.e. upper, middle and lower zone. These algorithms have shown reasonable improvement in segmenting the touching characters.
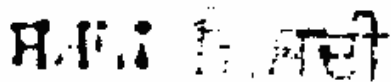
2.2. Broken characters

In this kind of degraded text the single character has been broken into more than one component. It is also observed that fragmented characters cause more errors than touching or heavy printed characters. This may be a natural consequence of the fact that there are generally more white pixels, even in text areas of the page, than black pixels. Therefore, converting a black pixel to a white pixel loses more information than vice versa [25]. Figure 4 shows some words of Gurmukhi script containing broken characters.
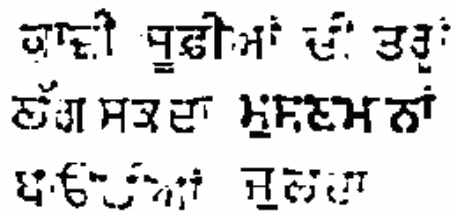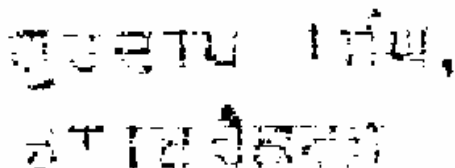
Figure 4. Broken Characters in Gurmukhi script.

The main reasons of the occurrence of the fragmented or broken characters in the document are inadequate scanning threshold, tired printer or copier cartridges, worn ribbons, light printed magazines or documents, misadjusted impact printers, degraded historical documents, faxed documents, dot matrix text etc. In extreme cases, only a few pixels of a character remain, not even enough for a human to identify the character in isolation as shown in figure 5.

Figure 5. Extremely Broken Characters in Gurmukhi script.

Excessive fragmentation may destroy an entire phrase as shown in figure 6, making even a person to identify with much difficulty.

Figure 6. Excessive fragmentation destroying entire phrase.

Due to presence of the broken characters the performance of any OCR may further decrease. Most of the work on recognition of headline based Indian scripts (Gurmukhi, Devanagari and Bangla) is based on the recognition and position of the headline. Due to broken characters as shown in figure 6, if the headline is destroyed, it will further make the problem more complicated.

On statistical analysis of the broken characters, we made following observations:
1. One character may be broken either horizontally or vertically in more than one fragment. The percentage of horizontally fragmented characters is more than of vertically fragmented characters. This is due to that, generally the headlines preserves it from breaking

which causes less fragmented characters in vertical direction. Diagonally broken characters are also found in printed Gurmukhi script.
2. If spacing between the characters is less, it becomes difficult to determine which fragment belongs to which character.
3. Generally each fragment of the broken character will have aspect ratio less than of a single isolated character.
4. Broken characters are generally found in middle zone, less in upper zone and very less in lower zone.
5. The fragment of a character is generally not similar in shape of some other individual character.

The recognition of broken characters in Gurmukhi script is somewhat different form the Roman script.
1. In Gurmukhi script mainly the broken characters are found in middle zone and less in upper and middle zone, but there is no concept of zoning in Roman Script.
2. Generally there is less information loss in case of Gurmukhi broken characters as the headline can be preserved even if we have broken characters. Restoration of headline may cause characters to lose less information as compared to Roman script characters as the concept of headline is not present in Roman script.
3. In Roman script diagonally broken characters are also found along with horizontal and vertical broken characters. But the chances of founding the diagonally broken characters in Gurmukhi script are less.
4. One broken character fragment in Roman script may be same in shape of some other character but in Gurmukhi script it happens rarely.

Restoration of the text is the most important task in case of recognition of the broken characters. Repeatedly application of dilation and erosion operation as preprocessing task can help to restore the image in some extent. Headline can be restored easily with the detection of its position. For segmentation of the broken characters, an approach to simultaneously segmentation and recognizing process can be useful. In this technique after selecting each fragment, it can be processed for recognition as a character. If it is recognized take the next segment otherwise merge with this segment the next segment and try to recognize both the segments as the fragments of a single character. Keep on adding the fragments until a character has been identified.

2.3. Heavy print

Sometimes even if the characters that are easily isolated, heavy print can distort their shapes, making them unidentifiable. It is very difficult to recognize a heavily printed character. The source of this king of degradation is the same as that of first category, i.e. touching characters. Figure 7 consists of some of the heavy printed characters in Gurmukhi script.

Figure 7. Heavy printed characters in Gurmukhi script.

The following observations have been made on the statistical analysis of the heavy print Gurmukhi characters:
1. The aspect ratio of the heavily printed characters is almost same as that of the original character.
2. It is very difficult to extract the features of such character, as it is just like a blob of black pixels of the height and width of the original character, with no ascenders or descenders to help distinguish them.
3. Generally heavy printed characters are also touching with neighboring characters, i.e. also falling in touching character category.
4. Most of the heavy printed characters have loop in their structure.
5. Heavy printed characters can be found in middle zone as well as lower and upper zone also. Even in clean documents characters in lower and upper zone are heavily printed.
6. Most of the times the shape of a heavy printed character may look like some other character.

Since the reason of production of heavily printed characters are same as that of touching characters, so most of time problem of heavily printed characters is merged with touching characters. Leading OCRs of Roman scripts fails to recognize heavy printed characters [25]. Nothing specific has been done in Indian scripts to deal with the problem of heavily printed characters. .

Not any special work has been reported to be done in this special category as every time this category of degraded text is treated same as the touching characters. The best solution to recognize these characters is to bypass the recognition process until post-processing stage is encountered. Here on the basis of dictionary look up, if the word containing the heavy printed characters is not a valid word, the dictionary lookup post processing work will correct it automatically.

2.4. Faxed Documents

Faxed documents are also treated as degraded documents as recognition of faxed documents creates its own kind of problems. Faxed documents are very light printed documents in general producing a large number of broken characters, few touching characters, and sometimes only few pixels remains of entire word. Faxed documents contains both salt and pepper noise. Sometimes it becomes difficult to recognize the faxed documents text even for human being. The following observations have been noted in the faxed document text:
1. The width of the stroke is not constant over the document.
2. Entire document contains varieties of the text, i.e. broken characters, touching characters in all three zones, broken and merged characters etc.
3. The quality the fax document also depends upon the quality of fax machine.

Some work has been reported to done to enhance the faxed documents so that a general OCR can understand them [20,21].
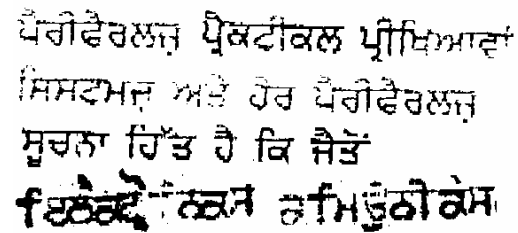
Figure 8. Words taken from faxed Gurmukhi script documents.

2.5. Typewritten Documents

Typewritten documents are another kind of degraded documents. Typewriters are widely used in government offices in India. Recognition of typewritten documents is itself a challenge as a typewritten document contains many problems

Figure 9. Words taken from typewritten Gurmukhi script documents.

On statistical analysis of the typewritten characters we made following observations:

1. There are many touching characters in middle zone of the text.
2. Lower zone characters touch the upper zone characters almost every time. Hence segmentation of lower zone from middle zone is very difficult task.
3. Unequal spacing between lines, words and even in the characters is observed.
4. There is a significant change in the shape of upper zone characters.
5. The headline of one complete word is usually broken in many parts and there are many ups and downs in the headline. Also, as shown in figure 9, most of the times characters in a single word are separate from each other and their headline do not touch. It makes difficult to recognize headline and baseline. As most of the algorithms are designed on the basis of headline and baseline, it will lead to decrease in recognition accuracy.
6. The darkness of the character depends upon that with how mush force the stroke key of typewriter is pressed. Due to hard pressing generally characters are heavily printed and the shape of characters is distorted as happens in heavy print category.
7. The typewriters can be of fixed width grid or variable width grid. Any typewriter with fixed width grid produces characters occupying same amount of horizontal space, whatever be the actual shape of the character.

3. Discussions

We have identified different kinds of degradation in printed Gurmukhi script in this paper. Touching characters, broken characters, heavy printed characters, faxed documents and typewritten documents are the types of degradations identified. There is wide scope of work to be done for enhancing the documents containing these kinds of degradation, and subsequently recognizing them. We have discussed few solutions for each kind of degradations also. There

are various other kinds of degradation also e.g. stray marks, curved baselines, blurred images, punctuations, and typographic degradations.

References

[1] R. G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", *IEEE Transactions on PAMI*, Vol. 18, No. 7, July 1996, pp. 690-706.

[2] Seong-Whan Lee, Dong-June Lee, and Hee-Seon Park, "A New Methodology for Gray-Scale Character Segmentation and Recognition", *IEEE Transactions on PAMI*, Vol. 18, No. 10, 1996, pp. 1045-1050.

[3] Chinmoy B. Bose and Shyh-Shiaw Kuo, "Connected and degraded text recognition using Hidden Markov Model", *Pattern Recognition*, Vol. 27, No. 10, 1994, pp. 1345-1363.

[4] S. Tsujimoto and H. Asada, "Resolving Ambiguity in Segmenting Touching Characters", Ist Int. Conference on Document Analysis and Recognition, Saint-Malo, France, Oct. 1991, pp. 701-709.

[5] R. G. Casey and G. Nagy, "Recursive Segmentation and Classification of Composite Character Patterns", Proc. 6th Int. Conf. on Pattern Recognition, Munich, Germany, 1982, pp. 1023-1026.

[6] T. Hong, **"**Degraded text recognition using visual and linguistic context**"**, Ph.D. Thesis, Computer Science Department of SUNY at Buffalo, 1995.

[7] S. Kahan, T. Pavlidis, and H. S. Baird, "On the recognition of printed characters of any font and size", *IEEE Transactions on PAMI*, Vol. 9, No. 2, March 1987, pp. 274-288.

[8] Y. Lu, *"*Machine printed character segmentation – An overview*", Pattern Recognition,* Vol. 28, No. 1, 1995, pp. 67-80.

[9] Veena Bansal and R.M.K. Sinha, "Segmentation of touching and fused Devanagari characters", *Pattern Recognition*, Vol. 35, No. 4, 2002, pp. 875-893.

[10] U. Garain and B.B. Chaudhuri, "Segmentation of touching characters in printed Devanagari and Bangla scripts using fuzzy multifactorial analysis", *IEEE Trans. Systems Man Cybern.,* Part C, Vol. 32, 2002, pp. 449–459.

[11] B.B. Chaudhuri, U. Pal, and M. Mitra, "Automatic Recognition of Printed Oriya Script", *ICDAR,* 2001, pp.795-799.

[12] M. K. Jindal, G. S. Lehal, and R. K. Sharma, "Segmentation problems and solutions in printed Degraded Gurmukhi Script", *International Journal of Signal Processing*, Vol. 2, No. 4, 2005, pp. 258-267.

[13] G. S. Lehal and Chandan Singh, "Text segmentation of machine printed Gurmukhi script", Document Recognition and Retrieval VIII, Proceedings SPIE, USA, Vol. 4307, 2001, pp. 223-231.

[14] G. S. Lehal and Chandan Singh, "A technique for segmentation of Gurmukhi script", Computer Analysis of Images and Patterns, Proceedings CAIP 2001, Warsaw, Poland, Lecture Notes in Computer Science, Springer-Verlag, Vol. 2127, 2001, pp. 191-200.

IEEE COMPUTER SOCIETY

[15] Y. Lu, B. Haist, L Harmon, J. Trenkle, and R Vogi, "An Accurate and Efficient System for Segmenting Machine-printed Text", U.S. Postal Service 5th Advanced Technology Conference, Washington D.C., Vol. 3, November 1992, pp. A 93 – A 105.

[16] O. Nakamura, M.Ujiie, N.Okamoto, and T. Minami, "A character segmentation algorithm for mixed-mode communication", IEICE, (D) 167-D, 11, 1984, pp. 1277-1285.

[17] N. Okamoto, O. Nakamura, and T. Minami, "Character Segmentation for Mixed-Mode Communication", IFIP'83, 1983, pp. 681-685.

[18] Michael Droettboom, "Correcting broken characters in the recognition of historical printed documents", Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries ((JCDL), Houston, Texas, USA, 27-31 May 2003, pp. 364 – 366.

[19] Berrin A. Yanikoglu, "Pitch-based segmentation and recognition of dot-matrix text", in *International Journal of Document Analysis and Recognition (IJDAR)*, Vol 3, 2000, pp. 34–39.

[20] M. Oguro, T. Akiyama, K. Ogura, "Faxed document image restoration using gray level representation", Proceedings of 4th International Conference on Document Analysis and Recognition, Vol. 2, 18-20 August 1997, pp. 679-683.

[21] Premkumar Natarajan, Issam Bazzi, Zhidong Lu, John Makhoul, and Richard Scwhartz, "Robust OCR of Degraded Documents", ICDAR'99, 1999, pp. 357-361.

[22] M. Cannon, J. Hochberg and P. Kelly, "QUARC: A Remarkably Effective Method for Increasing the OCR Accuracy of Degraded Typewritten Documents", Proceedings of the 1999 Symposium on Document Image Understanding Technology (SDIUT'99), Annapolis, MD, May 1999, pp. 154-158.

[23] M. Cannon, J. Hochberg and P. Kelly, "Quality assessment and restoration of typewritten document images", *IJDAR*, Vol. 2 , No. 2-3, 1999, pp. 80-89.

[24] C. Rodríguez, J. Muguerza, M. Navarro, A. Zárate, J.I. Martín, J.M. Pérez, "Segmentation of Low-Quality Typewritten Digits", Proceedings of 14th International Conference on Pattern Recognition, Vol. 2, 16-20 August 1998 pp. 1106 – 1109.

[25] Stephen V. Rice, George Nagy, and Thomas A. Nartker, *Optical Character Recognition*, Kluwer Academic Publishers, USA, 1999.