A Structural Feature based approach for script identification of Gurmukhi and Roman characters and words

Gurpreet Singh Lehal[1] , Chandan Singh[2] and Renu Dhir[3]

[1] Department of Computer Science and Engineering, Punjabi University, Patiala, India.
gslehal@mailcity.com
[2] Department of Computer Science and Engineering, Punjabi University, Patiala, India.
chandan@pbi.ernet.in
[3] Department of Computer Science and Engineering, Mational Institute of Technology, Jalandhar, India.
renu_dhir@yahoo.com

KEYWORDS : Gurmukhi, Roman, script identification, structural features

INTRODUCTION

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. This gives rise to the need for developing an automatic script identification system for words in Gurmukhi and Roman scripts.

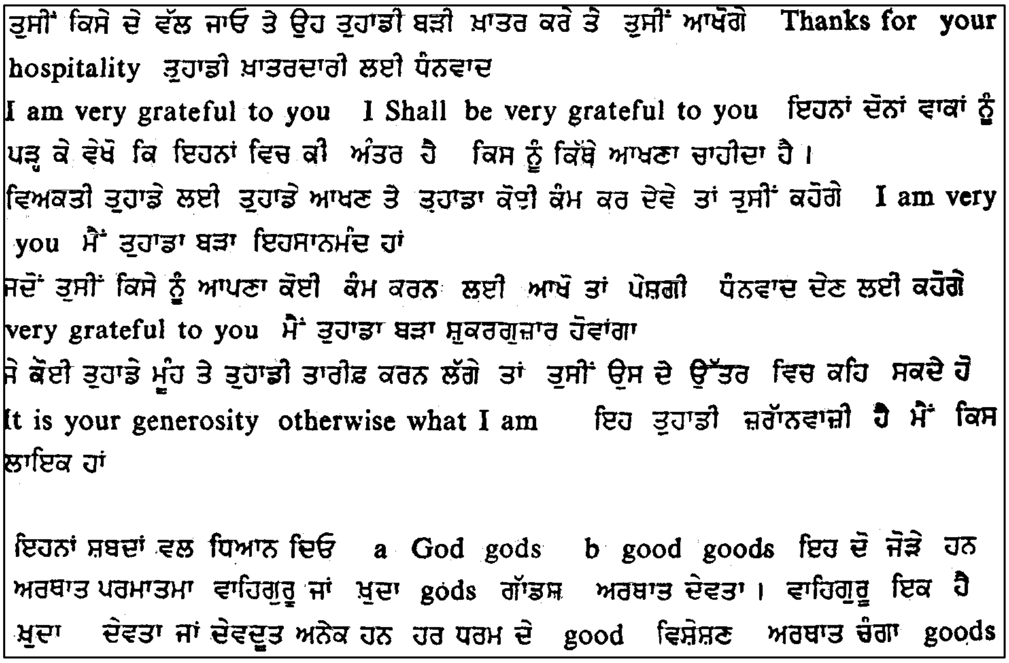
Many researchers have developed character recognizers tuned to specific applications, but multilingual capability has not received much attention. The capability of recognizing multilingual documents is both novel and useful. Sptiz has done some pioneering work in multilingual document recogntion[1-2]. Two types of techniques are usually adopted in language differentiation[3] : token matching[4] and statistical analysis[1]. Usually the first method is used to identify the specific language of a document, while the second method is used for gross classification. Most of these papers are concerned with classification of Oriental and European scripts. Pal and Chaudhury[5-6] have worked for classification of Roman and Indian language scripts. In [5], the authors have suggested a technique based on the headline property of Devanagri and Bangla script characters. This technique works well with majority of words but sometimes fails for italics or small sized words of one or two characters. Dhanya et al[7] have developed a script recognition system for Roman and Tamil scripts which performs script recognition at word level. Patil and Subbareddy[8] have described a neural network-based script identification system for Roman, Devanagri and Kannada language scripts.

All the reported studies perform script recognition either at the word level, line level or document level. To the best of our knowledge no one has yet successfully attempted to identify the script of a given character. At the word level too we have come across only two papers[5 and 7] which automatically determine the script of a word. 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 Fig.1, we have a sample text image containing both Gurmukhi and Roman script words and characters.

In this paper we have presented an automatic script identification system for Roman and Gurmukhi script words and characters. The features used in the system have been developed after a close study of structural features of Gurmukhi and Roman script characters and words.

SOME PROPERTIES OF GURMUKHI SCRIPT

Gurmukhi script is the official recognized script of Punjabi language in India. Gurmukhi script consists of 38 consonants called vianjans, 9 vowel symbols called laga or matras, 3 vowel carriers, 2 symbols for nasal sounds, one symbol for reduplication of sound of any consonant and three half characters (Fig. 2).



ਤੁਸੀਂ ਕਿਸੇ ਦੇ ਵੱਲ ਜਾਓ ਤੇ ਉਹ ਤੁਹਾਡੀ ਬੜੀ ਖ਼ਾਤਰ ਕਰੇ ਤੇ ਤੁਸੀਂ ਆਖੋਗੇ Thanks for your hospitality ਤੁਹਾਡੀ ਖ਼ਾਤਰਦਾਰੀ ਲਈ ਧੰਨਵਾਦ

I am very grateful to you I Shall be very grateful to you ਇਹਨਾਂ ਦੋਨਾਂ ਵਾਕਾਂ ਨੂੰ ਪੜ੍ਹ ਕੇ ਵੇਖੋ ਕਿ ਇਹਨਾਂ ਵਿਚ ਕੀ ਅੰਤਰ ਹੈ ਕਿਸ ਨੂੰ ਕਿੱਥੇ ਆਖਣਾ ਚਾਹੀਦਾ ਹੈ ।

ਵਿਅਕਤੀ ਤੁਹਾਡੇ ਲਈ ਤੁਹਾਡੇ ਆਖਣ ਤੇ ਤੁਹਾਡਾ ਕੋਈ ਕੰਮ ਕਰ ਦੇਵੇ ਤਾਂ ਤੁਸੀਂ ਕਹੋਗੇ I am very you ਮੈਂ ਤੁਹਾਡਾ ਬੜਾ ਇਹਸਾਨਮੰਦ ਹਾਂ

ਜਦੋਂ ਤੁਸੀਂ ਕਿਸੇ ਨੂੰ ਆਪਣਾ ਕੋਈ ਕੰਮ ਕਰਨ ਲਈ ਆਖੋ ਤਾਂ ਪੇਸ਼ਗੀ ਧੰਨਵਾਦ ਦੇਣ ਲਈ ਕਹੋਗੇ very grateful to you ਮੈਂ ਤੁਹਾਡਾ ਬੜਾ ਸ਼ੁਕਰਗੁਜ਼ਾਰ ਹੋਵਾਂਗਾ

ਜੇ ਕੋਈ ਤੁਹਾਡੇ ਮੂੰਹ ਤੇ ਤੁਹਾਡੀ ਤਾਰੀਫ਼ ਕਰਨ ਲੱਗੇ ਤਾਂ ਤੁਸੀਂ ਉਸ ਦੇ ਉੱਤਰ ਵਿਚ ਕਹਿ ਸਕਦੇ ਹੋ It is your generosity otherwise what I am ਇਹ ਤੁਹਾਡੀ ਜ਼ਰੱਨਵਾਜ਼ੀ ਹੈ ਮੈਂ ਕਿਸ ਲਾਇਕ ਹਾਂ

ਇਹਨਾਂ ਸ਼ਬਦਾਂ ਵਲ ਧਿਆਨ ਦਿਓ a God gods b good goods ਇਹ ਦੋ ਜੋੜੇ ਹਨ ਅਰਥਾਤ ਪਰਮਾਤਮਾ ਵਾਹਿਗੁਰੂ ਜਾਂ ਖ਼ੁਦਾ gods ਗੱਡਜ਼ ਅਰਥਾਤ ਦੇਵਤਾ । ਵਾਹਿਗੁਰੂ ਇਕ ਹੈ ਖ਼ੁਦਾ ਦੇਵਤਾ ਜਾਂ ਦੇਵਦੂਤ ਅਨੇਕ ਹਨ ਹਰ ਧਰਮ ਦੇ good ਵਿਸ਼ੇਸ਼ਣ ਅਰਥਾਤ ਚੰਗਾ goods

Fig. 1 : A Sample text containing English and Punjabi words

Some of the distinguishing characteristics of Gurmukhi script are:

- Most of the characters have a horizontal line at the upper part. The characters of words are connected mostly by this line called head line
- A word in Gurmukhi script can be partitioned into three horizontal zones (Fig 3). The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The middle zone is the busiest zone. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants.

Consonants

ੳ	ਅ	ੲ				Matra Vahak
		ਸ	ਹ			Mul Varag
ਕ	ਖ	ਗ	ਘ	ਙ		Kakvarg Toli
ਚ	ਛ	ਜ	ਝ	ਞ		Chach Varg Toli
ਟ	ਠ	ਡ	ਢ	ਣ		Ttatvarg Toli
ਤ	ਥ	ਦ	ਧ	ਨ		Tatvarg Toli
ਪ	ਫ	ਬ	ਭ	ਮ		Pavarg Toli
ਯ	ਰ	ਲ	ਵ	ੜ		Antim Toli
ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼	ਲ਼	Naveen Toli

Vowels

ਾ ਿ ੀ ੁ ੂ ੇ ੈ ੋ ੌ

Additional symbols

ਂ ੰ ੱ

Half Characters

੍ਰ ੍ਹ ੍ਵ

Fig 2 : Gurmukhi Character Set

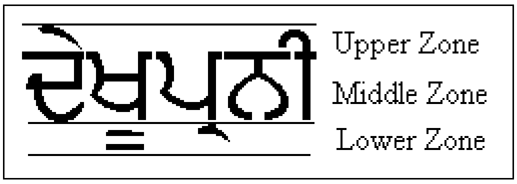


Fig 3 : Three zones of a word in Gurmukhi script

Proposed Script Recognition System

The proposed script recognition system for Gurmukhi and Roman script is composed of the following phases:

1. Digitization
2. Pre-processing
3. Segmentation
4. Word length identification
5. Feature extraction
6. Script Classification

Digitization

The first phase of this system is data acquisition, which is done by scanning the text at 300 dpi resolution.

Pre-processing

In the preprocessing stage skew correction and thinning is performed.

1. Skewness refers to the tilt in the bitmapped image of the scanned paper for OCR. It is usually caused if the document is not well aligned on the scanner, thus yielding a skewed (rotated) digital image. The segmentation and feature extraction algorithms developed by the authors for the script recognition are sensitive to the orientation (or skew) of the input document image making it necessary to develop algorithms to perform skew detection and correction automatically. We have used the projection profile technique for skew detection and correction.
2. Thinning is an essential pre-processing step whose main task is reducing patterns to their skeletons. It is an efficient method for expressing structural relationships in characters as it reduces space and processing time by simplifying data structures. For our present work we have used the thinning algorithm by Abdulla et al[9].

Segmentation

In our present work, the segmentation process is performed in two successive stages : line segmentation and word segmentation. For line and word segmentation horizontal and vertical projection profiles are respectively used. A text line is located between scan lines whose horizontal projection profile histogram values are greater than some threshold value. After a text line is detected, its vertical projection profile is determined and if a number of successive vertical projection profile histogram values are greater than some predefined threshold value, a word is considered to exist between these vertical lines. We do not go in for character segmentation, since we do not know in advance the script of the word image and different techniques have to be used for character segmentation of Gurmukhi and Roman script words. This is because in case of Roman script words there is inter character gap while no such gap exists for Gurmukhi script words.

Word Length Identification

As we shall see in later sections, we have used different features for script identification of words and characters. Here we define a word as a combination of more than one character. Thus it is necessary to know if the current word image represents a single character or more than one character. We used aspect ratio of a glyph to determine if the shape represents a single character or a word. It was observed that 94.35% and 93.6% of Roman and Gurmukhi script words respectively were correctly identified as words while 99.56% and 97.70% of single characters for Gurmukhi and Roman scripts respectively were correctly identified as characters.

FEATURE EXTRACTION

We have used structural features for script identification. 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 as follows:

Headline pixel count (F_1) : Headline is defined as the horizontal row in upper 40% region of a word with maximum number of black pixels. The headline is very prominently visible in Gurmukhi script words and not so prominent in Roman script words. This count of black pixels in the headline is used to distinguish the Roman words. If the count of black pixels is greater than a threshold T_1, then it is considered as Gurmukhi word otherwise it is a Roman word. The threshold T_1 is a function of word length. During the experiments the value of T_1 was found to be 60% of word length. It was found that about 93.36% of Roman script words have headline coverage lesser than 60% of total word width, while 97.17% of Gurmukhi script words have the headline coverage greater than 60%. This feature works very well for longer words especially for words with more than three characters in English or more than three characters in middle zone for Punjabi. It was found that there are 98.53% of long words in Roman script have the headline coverage lesser than 60% and 98.22% of long Gurmukhi script words have headline coverage greater than 60. In some cases of small words of two or three characters for Roman script it fails (e.g. tt or TIP) and similarly it does not work correctly if majority of the middle zone consonants in the Gurmukhi script did not have the headline such as the word ਸਾਪ. In case of single characters there are a few English letters such as [E F T Z] and Gurmukhi letters such as [ਮ ਘ ਪ ਖ਼ ਸ] for which this feature does not work.

Inter character gap (F_2) : Another distinct characteristic of Gurmukhi words is the absence of any inter character gap. The characters are glued along the headline. The second feature is based on the inter-character gaps in the word. If there is no gap then it is highly probable that the word is in Gurmukhi and as the number of inter-character gaps increase the probability of the word being in Roman script increases. From a statistical analysis of Gurmukhi and Roman script words it was found that 97.43% of Punjabi words do not have any gap between characters in middle zone, while 98.20% of English words have inter-character gap. The only exceptions for Roman script are small or italic words such as (fi *THE*).

Bottom Projection Profile (F_3) : The third feature is based on bottom projection profile. It is calculated by drawing vertical lines from each black pixel on the headline to bottom of the character. If the line reaches the bottom without encountering any black pixel then the count is incremented by one. Or in other words we look for the columns which contain exactly one black pixel and that black pixel is present in the row corresponding to the headline. If such a column is found then the count is incremented by one. The percentage ratio of count to word width is then evaluated and stored in F_3. From the experiments, it was found that 93.41% of Gurmukhi words have value of F_3 in the range [21..55] while 89.40% of Roman script words have value of F_3 outside that range (Fig. 4).

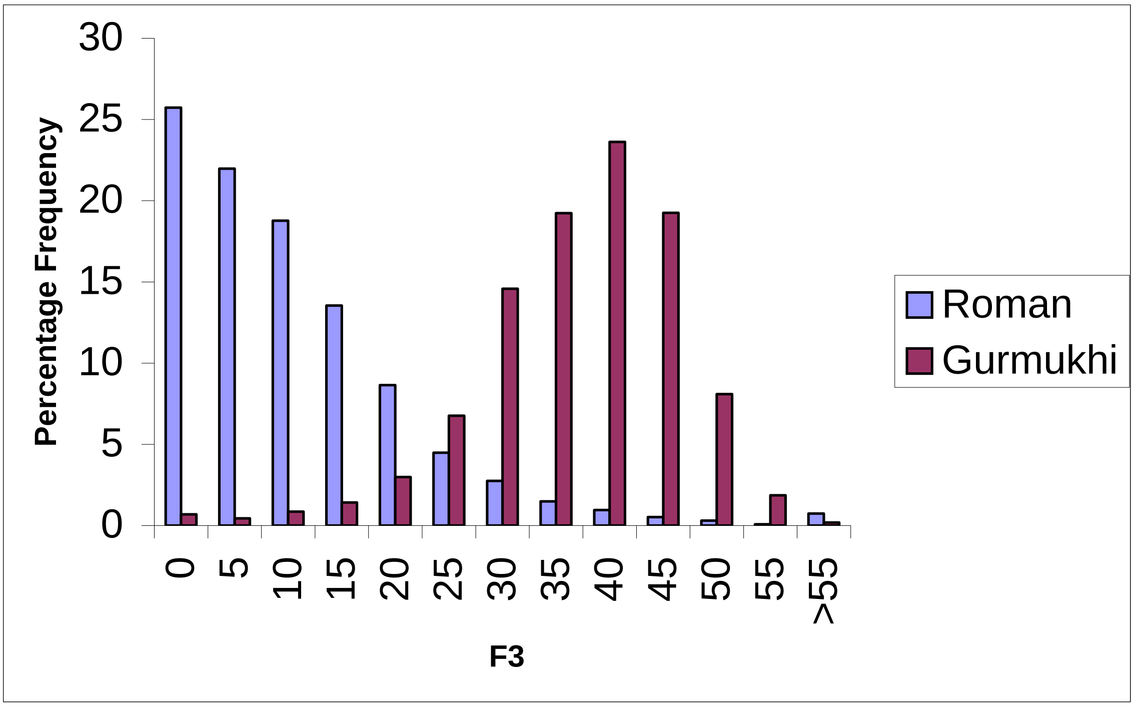


Fig 4 : Frequency distribution of Feature F_3

Protruding Regions Beyond Headline (F_4) : It is observed that usually there is no character in Gurmukhi which has any protruding regions beyond the headline on the left or right side of word, while many English characters such as (q u i o a d g h l b n m etc.) have this property. So if there is any such character present in the word, then it is highly probable that the word is in Roman script. The value of F_4 is 1 if this feature is present in any character of the word else it is 0. The experiments revealed that 89.05% of Roman script words have at least one character with protruding regions while 98.10% of Gurmukhi script words do not have any such character.

Right Vertical Bar (F_5) : A majority of the Gurmukhi characters have a vertical bar at the right extreme. In comparison about 20% of Roman alphabets have such bar. This feature searches for presence of a vertical bar on the right most portion of the word. It is normalized by dividing the length of any such vertical bar by the word height and multiplying with 100. The frequency distribution of values of F_5 for Gurmukhi and Roman script words is depicted in Table 1. As it can be observed, the value of F_5 is greater than 90 for 41.2% of Punjabi words and 14.64% of English words and on the other extreme lesser than 30 for 39.57% of English words and 3.67% of Punjabi. Thus we obtain two cut off points for F_5. The first cutoff point, which is for value lesser than 30, indicates that the word is in Roman script and the second cut off point, which is for value greater than 90, which indicates that the word could be in Gurmukhi script.

TABLE 1: Percentage Frequency of Occurrence of F_5 for Roman and Gurmukhi Scripts

F_5	Roman script		Gurmukhi script	
	Percentage Frequency of occurrence	Cumulative Frequency	Percentage Frequency of occurrence	Cumulative Frequency
0-10	20.12032	20.12032	0.984103	0.984103
11-20	11.96524	32.08556	2.611658	2.611658
21-30	7.486631	39.57219	1.059803	3.671461
31-40	6.818182	46.39037	3.974262	7.645723
41-50	13.1016	59.49198	16.162	23.80772
51-60	5.548128	65.04011	17.25965	41.06737
61-70	4.612299	69.65241	6.888721	47.95609
71-80	4.679144	74.33155	4.314913	52.27101
81-90	11.02941	85.36096	6.51022	58.78123
91-100	14.63904	100	41.21878	100

Loop in lower half (F_6) : This feature is based on one of the features developed by Lehal and Singh[10] for Gurmukhi character recognition. It was observed that 19 Gurmukhi symbols in middle zone (ਉ ਓ ਕ ਖ ਖ਼ ਗ ਗ਼ ਙ ਚ ਛ ਠ ਡ ਢ ਥ ਫ ਫ਼ ਬ ਭ ਰ) have a loop which is not touching the headline and in contrast only three Roman script characters (a g B) have such loops. There are many other Roman characters which have loops (a e o q etc.), but their loops touch the headline and so they are not considered. The value of F_6 is evaluated by dividing the number of such loops by (F_2+1). If the value of F_6 is greater than or equal to 0.5, then the probability of the word being in Gurmukhi script increases else the probability of the word being in Roman script increases. It was determined during the experiments that about 61.59% of Gurmukhi script words have value of F_6 greater than or equal to 0.50 and 95.22% of Roman script words have value of F_6 lesser than 0.50.

It was found that the above six features were sufficient to differentiate between the Gurmukhi and Roman script words with a very high accuracy but at the character level the script recognition accuracy was still around 91%. Thus we had to design some other structural features for character level recognition. These features are as follows:

Left Vertical bar(F_7) : This feature searches for a vertical bar on the left extreme such that the length of the bar is greater than or equal to any other vertical line in the character and the length is greater than 90% of character height. There are many Roman characters (b h k m n p r u B C D E F G H K L M N P R U) which have such vertical bars but there is no Gurmukhi character having such a bar at its left extreme. The feature has value one if such a vertical bar is present else it is zero.

C shape in lower half (F_8) : This feature is also developed specifically for script recognition at character level. It looks for a C shape like structure, which is not touching the headline. Some of the Gurmukhi characters which have such a structure are (ੲ ਵ ਟ ਦ ਣ ਢ ਫ ਫ਼ ਞ) while there is no Roman character with such a shape. The feature has value one if such shape is present in the character else it is zero.

U like shape (F_9) : This feature is true if a U shape like structure is present in the character. This structure is present if the following three conditions are true a) there exists a horizontal line at the bottom of the character b)the length of the line is at least 40% of the character width and c) there is no black pixel lying above the line. This structure is present in some of the Roman script characters such as (u U j J L), but there is no Gurmukhi character having this structure. The feature has value one if such shape is present in the character else it is zero.

As we are dealing with thinned images, so the straight lines will not be straight many times and they will have pixels shifted from their positions. So while looking for horizontal and vertical lines, we look for the existence of pixels in adjacent positions too.

CLASSIFICATION

For classification of words a democratic procedure involving voting at different levels is used. At the first level the most robust features i.e. F_1 and F_2 are used to decide the script. If they both reach at a consensus, then no further features are invoked and the script decided by the two is accepted else the classifier moves to the second level where it takes the assistance of features F_3 and F_4. If they both agree on same script, then that is accepted else the features F_5 and F_6, which are in next level, are used. If still the issue is not resolved then the feature F_2, which is the most stable feature, is given the veto power and the decision made by it is taken as final decision. The cut off points used for classification by the different features are displayed in Table 2.

For classification of single characters, since the feature F_2 is zero for both Roman and Gurmukhi scripts so this feature is ignored and instead features F_7, F_8 and F_9 are used. The classification process is modified as follow.

Since there is no other feature, besides the feature F_1 which is robust and dependable so each of the feature is allowed to cast its vote for a particular script and the script is decided by the majority. In case of tie, the decision made by feature F_1 is taken as final. It may also be noted that some of the features such as F_7, F_8 and F_9 give only a one way decision. That is, if the feature has value one then the character belongs to a particular script but if it is zero then we cannot say that the character belongs to the other script since there may be many characters of first script for which the feature has zero value. As for example, the feature F_7 has value zero for many Roman script characters such as (a d e f g etc.) and thus the absence of this feature does not mean that the character belongs to Gurmukhi script. The utility of these one-way decision features is that they increase the confidence of the decision obtained by the other binary decision features. While in some of the features such as F_7, the feature was devised specially for characters such as M and N which could not be classified correctly in some cases using the rest of the features. The cut off points used for classification of characters are displayed in Table 3.

Table 2: Values of Features F_1 to F_6 used for Classification of Roman and Gurmukhi Scripts words

Feature	Roman	Gurmukhi
F_1	Lesser than or equal to 60	Greater than 60
F_2	Greater than Zero	Zero
F_3	Lesser than or equal to 20 or greater than 55	Greater than 20 and lesser than equal to 55
F_4	One	Zero
F_5	Lesser than or equal to 30	Greater than 90
F_6	Lesser than 50	Greater than or equal to 50

Table 3: Values of Features F_1 to F_9 used for Classification of Roman and Gurmukhi Script characters

Feature	Roman	Gurmukhi
F_1	Lesser than or equal to 60	Greater than 60
F_2	Don’t Care	Don’t Care
F_3	Lesser than or equal to 20 or greater than 55	Greater than 20 and lesser than equal to 55
F_4	One	Zero
F_5	Lesser than or equal to 30	Greater than 90
F_6	Lesser than 50	Greater than or equal to 50
F_7	One	Don’t Care
F_8	Don’t Care	One
F_9	One	Don’t Care

As an example, we have some sample characters in Fig. 5. The values of the features are listed in Table 4. For illustration purpose the values are colour coded. Blue colour represents that the feature votes the script to be in Roman script while red colour represents Gurmukhi script, grey colour represents that the script is undecided. Thus, for example in case of the character ਪ, since the headline coverage is 40 so the feature F_1 votes the character to be in Roman script. The value of feature F_3 is 34 and this implies that the character is in Gurmukhi. To illustrate the calculation of F_3, beneath each character image we have drawn a colour coded line. Red pixel denotes that in that column there is one and only one black pixel present in the headline row of that particular column, while blue pixel represents that there could be either zero black pixel or one black pixel in non-headline row or more than one black pixel in that column. The sum of red pixels divided by the sum of red and blue pixels and multiplied by hundred yields the value of F_3. Since finally for character ਪ we have three features voting for Gurmukhi script and two features voting for Roman script and three features are undecided so by majority the character is decided to be in Gurmukhi script.

Some sample words for script recognition are displayed in Fig. 6. The values of the features and the level at which the script recognition is finalized are listed in Table 5. The value of F_1 for the first word is greater than 60 for first two words. The value of F_2 is zero for the first word and as both features F_1 and F_2 judge the word’s script to be Gurmukhi, the classification process stops at first level. The value of F_2 for the second word is two and thus according to it, the script of the word is Roman. Since features F_1 and F_2 differ in their opinion, so the features at the second level, F_3 and F_4, are used. Since values of F_3 and F_4 are 27 and 0 respectively, so both judge the word's script to be Gurmukhi and the classification process stops at second level. In case of the fourth word, the issue could not be decided at second level and so the help of features F_5 and F_6 is taken. The value of F_5 is 100 and F_6 is 0. So feature F_5 indicates that the word is in Gurmukhi while feature F_6 votes for Roman script. Thus the issue is resolved at the fourth level by looking at the value of F_2. Since F_2 is zero, so the script is classified as Gurmukhi.

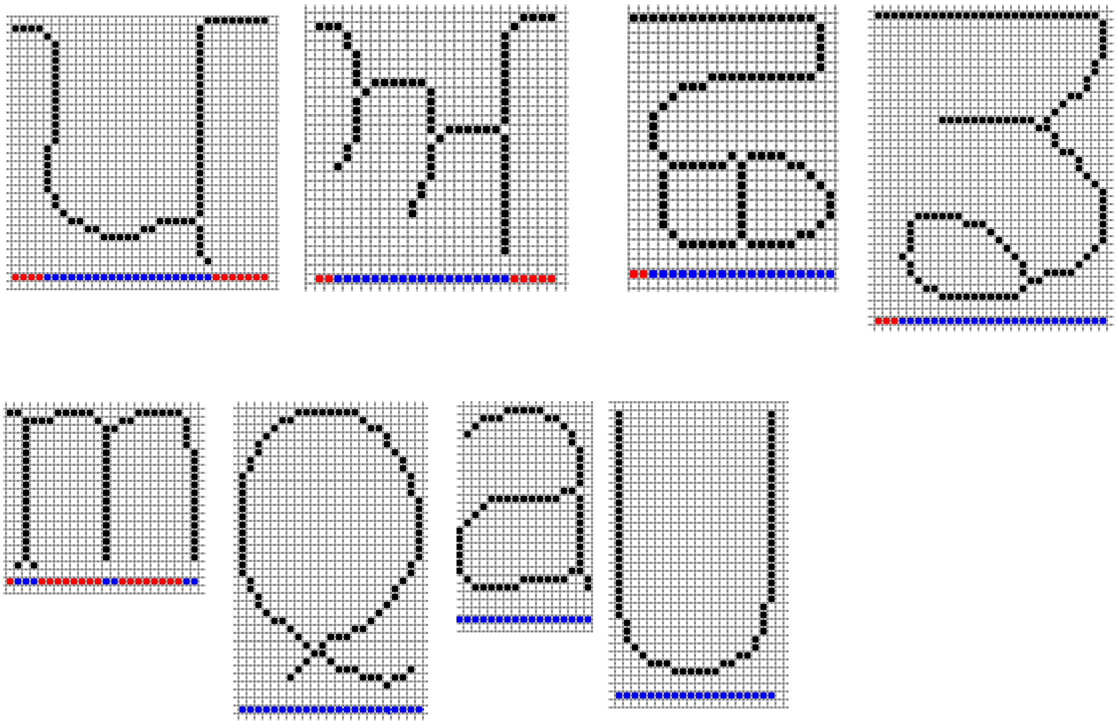


Fig. 5 : Some sample character images

Table 4 : Feature values for the character images

Character → Feature↓	ਪ	ਅ	ਛ	ਤ	m	Q	a	U
F_1	40	30	95	100	87	50	58	10
F_3	34	26	9	10	70	0	0	0
F_4	0	0	1	0	1	1	1	0
F_5	100	96	56	57	100	62	90	90
F_6	0	0	2.0	1.0	0	0	1.0	0
F_7	0	0	0	0	1	0	0	1
F_8	0	0	1	0	0	0	0	0
F_9	0	0	0	0	0	0	0	1

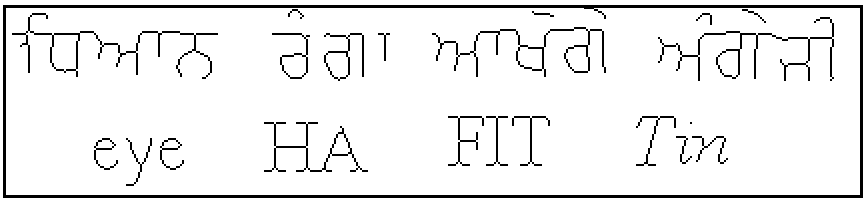


Fig. 6 : Some sample word images

Table 5 : Feature values for the word images

Word → Feature↓	ਧਿਆਨ	ਰੰਗਾ	ਆਖੋਗੇ	ਅੰਗੇਜੀ	eye	HA	FIT	*Tin*
F_1	79	72	54	43	51	44	92	43
F_2	0	2	0	0	2	0	2	0
F_3	40	27	34	18	2	0	15	14
F_4	0	0	1	0	2	1	0	0
F_5	50	56	100	100	33	18	9	44
F_6	0	0.66	2.0	0	0	0	0	0
Recognition Level	First	Second	Third	Fourth	First	Second	Third	Third

EXPERIMENTAL RESULTS

We tested our system on more than 100 text images scanned from books, magazines and computer printouts. There are about 18000 Gurmukhi script words and 15000 Roman script words. The recognition accuracy of the system is 98.29% for Gurmukhi script words and 99.02% for Roman script words (Table 6). It can be observed from Table 6, that as the word size increases the script recognition accuracy also increases. In fact for words having more than five characters the script recognition accuracy is 100% for Roman script and 99.86% for Gurmukhi script. For single characters, the accuracy is 96.81% and 95.47% for Roman and Gurmukhi scripts respectively.

Table 6 : Script recognition accuracy according to the word size

Word Size	Roman Script	Gurmukhi Script
One	**96.81%**	**95.47%**
Two	97.58%	97.38%
Three	99.81%	99.20%
Four	99.92%	99.35%
Five	99.96%	99.75%
Greater than Five	100%	99.86%
All Words with more than character	**99.61%**	**99.06%**
Over All	**99.02%**	**98.29%**

As already discussed, we used aspect ratio of a glyph to determine if the shape represents a single character or a word. But it was observed that for 5.65% and 6.4% Roman and Gurmukhi script words respectively, the word was wrongly classified as character and thus the features for character were applied on it. It resulted in recognition accuracy of 98.82% and 94.32% for Roman and Gurmukhi script words which were identified as single characters, while in the rest of 94.35% and 93.6% of cases for Roman and Gurmukhi scripts respectively the recognition accuracy was 99.67% and 99.35%. The overall recognition accuracy is 99.61% and 99.06% for Roman and Gurmukhi scripts respectively.

For single characters in case of Gurmukhi and Roman scripts 99.56% and 97.70% characters are correctly identified as characters with script recognition accuracy of 98.26% and 95.43% respectively. The rest of the Gurmukhi and Roman scripts characters which are identified as words are recognized with 100% and 37.5% accuracy respectively. This is because of feature F_2. In case of Roman script character, if it is identified as a word, then since there is no inter character gap, so the feature F_2 will cast its vote that the word is in Gurmukhi script and thus the Roman script characters having headline coverage greater than 60% such as T E F etc. will be recognized as in Gurmukhi script.

Table 7 : Recognition level for script recognition of Words

	Correct Recognition		Incorrect Recognition	
Recognition Level	Roman	Gurmukhi	Roman	Gurmukhi
First	97.17%	95.92%	0.05%	0.16%
Second	1.85%	3.18%	0.20%	0.11%

Three	0.45%	0.11%	0.04%	0.27%
Four	0.20%	0.14%	0.04%	0.11%
Overall	99.67%	99.35%	0.33%	0.65%

In case of classification of script at word level, it can be observed from Table 7, that in 97.22% of cases for Roman script, the decision is made at the first level itself and there is no need to invoke the rest of the features. In only 0.05% of cases, is the decision wrongly made at the first level for Roman script. While in case of Gurmukhi script, both the features of first level correctly agree for the Gurmukhi script in 95.92% of cases and in 0.16% cases both the features arrive at wrong decision and overall the final decision is made in 96.08% of cases at first level itself.

Table 8 : Voting pattern for script recognition of characters

	Correct Recognition		Incorrect Recognition	
Difference	Roman	Gurmukhi	Roman	Gurmukhi
Tie	3.42%	5.20%	0.43%	0.95%
One	10.16%	11.86%	1.18%	1.33%
Two	8.95%	15.46%	0.10%	1.27%
Three	18.70%	20.29%	0.03%	0.38%
Four	20.28%	6.98%	0.0%	0.32%
Greater than 4	36.75%	35.64%	0.0%	0.32%
Overall	98.26%	95.43%	1.74%	4.57%

We have also studied the voting pattern for script classification at character level (Table 8). As it can be observed from Table 8, in case of Roman script if the glyph is correctly recognized as character, then the script recognition accuracy is 98.26% and it is 95.43% for Gurmukhi script. It is also important to note the votes polled by the features. In case the difference of votes is greater than three then in case of Roman script we can be sure that the script has been correctly recognized, while in case of Gurmukhi script in 99.36% of cases the script is correctly identified. In fact the votes polled for script recognition can be combined with the classification distance for character recognition of that word. Thus, for example, if we find that a particular character is identified as in Gurmukhi script and the difference of the votes polled is say one, then if the character is not recognized with very high confidence by the Gurmukhi OCR then it can fed to the Roman OCR and the results of OCR which recognizes the character with higher accuracy can be retained. While if the difference of votes polled is four and the script is recognized as Roman then even though the classification distance of the feature vector of the character and nearest matching prototype is high we still assume the script to be Roman.

CONCLUSION

We have proposed a method to differentiate between Gurmukhi and Roman script words based on a combined analysis of several discriminating features. This method has been implemented and tested on about 100 documents, and the experimental results indicate that this method is effective and reliable. This is the first time that such a script recognition system has been developed for Roman and an Indian language script, which works at word and character level.

REFERENCES

1. A.L.Spitz, ”Script and language determination from document images”, *Proceedings 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, pp. 229-235, 1994.
2. A.L.Spitz, ”Multilingual Document Recognition”, *Handbook of Character Recognition and Document Image Analysis*, Eds. H. Bunke and P.S.P. Wang, World Scientific Publishing Company, pp.259-284,1997.
3. J.Ding, L.Lam and C. Y. Suen, ”Classification of Oriental and European scripts by using characteristic features”, *Proceedings 4th International Conference on Document Analysis and Recognition,* Ulm, Germany, pp. 1023-1027, 1997.

4. J. Hochberg, L. Kerns, P. Kelly and T. Thomas, “Automatic script identification from images using cluster-based templates”, *Proceedings 3rd International Conference on Document Analysis and Recognition,* Montreal, Canada, pp. 378-381, 1995.
5. U. Pal and B.B. Chaudhuri, ”Automatic separation of words in multi-lingual multi-script Indian documents”, *Proceedings 4th International Conference on Document Analysis and Recognition,* Ulm, Germany, pp. 576-579, 1997.
6. U. Pal and B.B. Chaudhuri, ”Script line separation from Indian multi-script documents”, *Proceedings 5th International Conference on Document Analysis and Recognition,* Banglore, India, pp. 406-409, 1999.
7. D. Dhanya, A. G. Ramakrishnam and P. B. Pati, “Script identification in printed bilingual documents”, *Sandhana Academy Proceedings in Engineering Sciences*, Vol. 27, pp. 73-82, 2002.
8. N. V. S. Reddy and S. B. Patil, “Neural network based system for script identification in Indian documents”, *Sandhana Academy Proceedings in Engineering Sciences*, Vol. 27, pp. 83-98, 2002.
9. W. H. Abdulla, A. O. M. Saleh and A. H. Morad, “A preprocessing algorithm for handwritten character recognition”, *Pattern Recognition Letters,* Vol. 7, pp. 13-18, 1988.
10. G S Lehal and Chandan Singh, “A Gurmukhi script recognition system”, *Proceedings 15th International Conference on Pattern Recognition*, Barcelona, Spain, Vol. 2, pp. 557-560, 2000.