

Efficient Technique for Automatic Extraction and Identification of Text from an Image using Friend Pattern Chain and Euclidean Distance

Mohan P Pradhan

Dept. of CSE, SMIT, SMU, Sikkim, India.

M K Ghose

Dept. of CSE, SMIT, SMU, Sikkim, India.

ABSTRACT

A reference map acts as the fundamental source of information pertaining to various morphological features in studies related to Geographic information system (GIS). The feature set includes rivers, contours, transportation-network, national and international demarcation to name a few. Each of these features is represented on the reference map with the help of different color code to ease the task of visual interpretation while digitizing them. So reference map can be considered as superimposed layers of information casted on the terrain. One of the key features associated with a reference map is text demarking the landmarks and the rivers etc. These texts play an important role in identifying the land marks and rivers while performing studies related to drainage network or demography.

These features are can possibly digitized in two ways either using traditional manual techniques which involves visually interpreting individual characters and then grouping it to form a name. This process takes demands increased effort, time and precision from the researchers responsible for digitizing the feature. Instead relying on the less effective traditional manual approach an efficient automatic extraction procedure can be developed that takes into consideration certain pre acquired knowledge for identifying the text.

This proposed work introduces a rotation invariant text identifying procedure that relies on the combination of friend pattern chain concept and distance.

Keywords

Friend pattern chain, Morphology, Demography, Geographic information system.

1. INTRODUCTION

Textual data may be associated with a reference map to qualitatively represent various morphological features such as rivers, landmarks and water bodies etc. These textual data are incorporated to the reference map just to elevate and enhance the representation as well as to create quick reference to the feature of interest. In order to assure uniformity in representation, textual data associated with a feature is presented in a specific font and specific size. Upon identifying the font and the size in which the textual data are associated with a feature, a knowledge basis can be formulated. This knowledge basis can be further used in order to identify characters in a name and create a repository out of names.

The effectiveness of the identification process greatly relies on how these knowledge bases are created and the factors that are taken consideration while creating these knowledge bases.

One fundamental problem involved in text recognition procedure has to be alignment of text in the reference map on in simpler terms how the text is placed in the reference map. Normally the orientation and placement of feature in a reference map dictates the orientation of the text. That implies that the procedure should be able to identify text aligned along any direction.

So in order to facilitate the above specified requirement an efficient rotation invariant process is to be developed that utilizes the knowledge basis for identifying the text.

The efficiency of the procedure thus depends both on the technique as well as the reference knowledge bases.

This proposed work in abstraction involves the following steps for identifying the text,

- color segmentation of the reference map in order to highlight only the feature with which the text associated is to be extracted
- application of morphological operation for representing feature using single pixel width
- acquisition of knowledge in terms of friend pattern chain
- determining uniqueness in terms of friend pattern chain
- identify conflicting friend pattern chain
- associate additional parameter such as distance with conflicting friend pattern chains
- association priority with the frequencies in the friend pattern chain
- determination of words
- scan the image for determining the match
- creating of repository that hosts the identified words

2. RELATED WORK

In past several research initiatives for text identification have been proposed, most of these research initiative heavily rely on the capability of Optical Character Recognition Module for initial data acquisition process. Besides its ability of easy and quick recognition OCR suffers from several inherent lags firstly the accuracy and preciseness of OCR depends heavily on the OCR application and the devices. Secondly OCRs performance degrades with degradation in the quality and tone of the image.

There are three main approaches in pattern recognition applications: statistical, structural and soft-computing approaches. In the statistical pattern recognition quantitative or numerical or frequency features are adopted. A. Gupta et al. [1] presented an algorithm based on matched wavelets and MRF model to automatically identify and extract the low contrast text regions from scanned manuscript images and enhance them using a histograms matching technique. Hamamoto and Uchimura et al. [2] proposed a Gabor filter-

based feature extraction method for handwritten numeral character recognition where they have shown that frequency features demonstrated the good behavior in character recognition.

Structural approaches emphasizes on qualitative features and their relationships it includes direction, skeleton, and topological features [3, 4]. Dubey and Sinthupinyo [5] proposed structural feature extraction technique. The structural features are defined automatically through K-Mean clustering, unlike traditional techniques that was manually defined. Poudroux [6] proposed an automatic approach to extract and recognize toponyms based on image segmentation and connected component processing. Pezeshk [7] developed a custom multi-font segmentation-free OCR that combines the outputs of two sets of Hidden Markov Models (HMMs), and bigram and character width probabilities to recognize the text. Chiang et al. [8] suggested an automatic text recognition approach which focuses on locating individual text labels in the map and detecting their orientations and leverages horizontal text recognition of commercial OCR software. Caprioli et al. [9] described a method for the semi-automated extraction and clustering of characters occurring in scanned topographic maps which takes into account the oriented strings that are frequent in topographic maps.

Soft-computing approaches include fuzzy, neural-net and genetic algorithms. Yang et al. [10] proposed character recognition method of license plate number based on parallel BP neural networks. The character feature is extracted by using skeleton and the character feature is put into the parallel neural networks for character is recognition. Velázquez et al. [11] applied OCR-based recognition with Artificial Neural Networks (ANN) to define the coordinates, size and orientation of alphanumeric character strings.

Yamamoto et al. [12] proposed a method where numerals and symbols are recognized by the multi-angled parallelism (MAP) matching method, while small dots and lines are extracted by the MAP operation method. These results are then used to determine the value, position, and attributes of the elevations marked on the topographic maps. Multi Angled Parallelism (MAP) provides an efficient tool to detect miscellaneous linear features. However, parts of lines that pass through characters are often misclassified. It is proposed in [13], an improvement over MAP to automatically extract complete line networks with arbitrary orientation and curvature even when they pass through characters with minimal impact on the text content. The resulting text can then be processed for text grouping, reorientation, and recognition. Pezeshk et al. [7] proposed automatic extraction of text based on new line representation technique and a set of directional morphological operations that are based on the MAP algorithm.

A custom OCR is then used to recognize the extracted street labels and major place names. Poudroux et al. Detected text area is then fed to OCR software for recognition. Luyang et al. [14] used connected components of black layer extracted from topo-sheets to separate line Art, text, and icons. Anegawa et al. [15] proposed a system for recognizing numeric strings from topographical maps, which is composed of extracting uncertain numeric strings using automatic recognition stage based on topographical map features only and corrected by the interactive recognition stage. Nakamura et al. [16] described a method for recognizing character strings from topographical maps which consists of a bottom-up process for extracting character candidates from a map and a top-down process in which these character candidates are grouped into strings using linguistic knowledge of strings.

3. METHODOLOGY

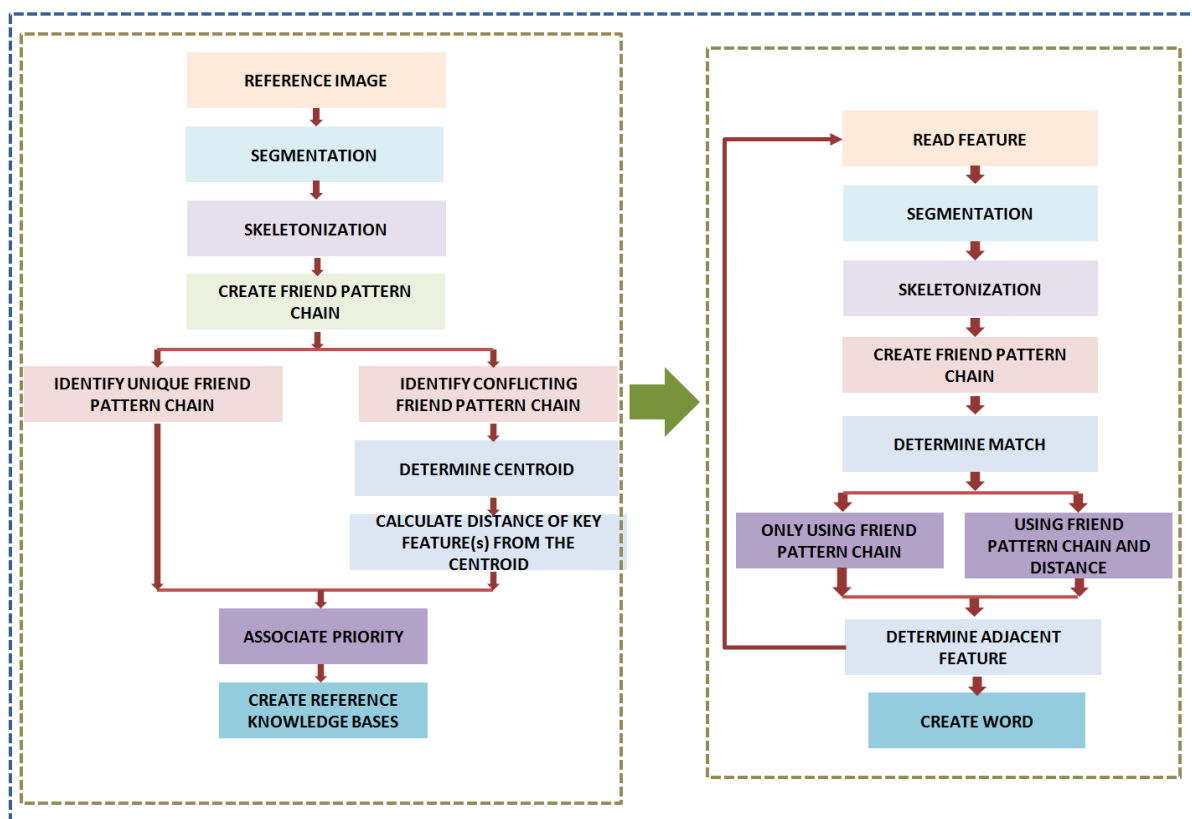


Fig 1:- Schema diagram for proposed work

In order to initiate any of the post seeding operation firstly the image is to be read and stored in a 3D array where in each layer contributes to blue green and red layers of the image.

The color reference map for the ease of processing is first reduced from multi layer image to single layer gray image. This process is called as conversing of RGB image to gray image. While converting a RGB image to gray image, intensity at a given coordinate in the result of the proportion contribution of the red, green and blue at the same coordinate and is implemented as sum of products.

So, intensity at a given coordinate x, y in a gray image is equal to $G(x, y) = \text{proportion contribution of blue} * \text{Blue}(x, y) + \text{proportion contribution of green} * \text{Green}(x, y) + \text{proportion contribution of red} * \text{Red}(x, y)$

Where, $\text{proportion contribution of blue} + \text{proportion contribution of green} + \text{proportion contribution of red} = 1$

3.1 Segmentation

A reference map hosts several features and their attribute details. All of the features represented in the reference may not be required while executing an inferential study related to one specific feature. So, it becomes essential to highlight only the required feature and suppress rest of the features. This can be done by image segmentation. Image segmentation is a process that tries to partition the image into set of subset based on dissimilarity of intensity value. Segmentation takes into consideration certain threshold value as basis for partitioning. In this work a simple thresholding technique is used in order to segment the image.

3.2 Skeletonization

Skeletonization in image processing is the process of representing the feature of interest using single pixel width. The advantage of Skeletonization process is that it eases computation procedure. Skeletonization or thinning is a morphological operator.

3.3 Creation of friend pattern chain

The friend pattern chain matches closely with the chain code scheme but in this case rather than creating sequence of movement along eight directions, we store the frequency of occurrence of pixels with same number of friends in a chain. In a window of 3X3 following are the location where friends may be encountered, let us consider $G(i, j)$ be the coordinates for which we are determining friends.

Coordinate	$g(i,j+1)$	$g(i+1,j+1)$	$g(i+1,j)$	$g(i+1,j-1)$
Friends				
	$g(i,j-1)$	$g(i-1,j-1)$	$g(i-1,j)$	$g(i-1,j+1)$

Following cases identifies the types of friends

Coordinates	0 friend	1 friend	
Friends			
	2 friend	3 friend	4 friend

The position of friends might be any of the eight neighbors.

Example, Friend Pattern Chain

Image	Skeletonized	Friend Pattern							
friend pattern chain									
Frequency with i friends									
Friends	0	1	2	3	4	5	6	7	8
Frequency	0	2	24	2	0	0	0	0	0

3.4 Creation of repository

As in case of 'A' explained above similar friend pattern chain is created for all the character including the upper case and lower case characters. Table 1 highlights the friend pattern chain for all uppercase and all lower case characters.

In this work Arial font with size 12 was taken into consideration for implementation.

Table 1:- Friend pattern chain and the count of significant pixels

CHAR	ID	0F	1F	2F	3F	4F	5F	6F	7F	8F	TOTAL
A	0	0	2	22	6	0	0	0	0	0	30
B	1	0	0	37	4	0	0	0	0	0	41
C	2	0	2	24	0	0	0	0	0	0	26
D	3	0	0	34	0	0	0	0	0	0	34
E	4	0	3	29	1	0	0	0	0	0	33
F	5	0	3	19	1	0	0	0	0	0	23
G	6	0	2	30	0	0	0	0	0	0	32
H	7	0	4	26	2	0	0	0	0	0	32
I	8	0	2	11	0	0	0	0	0	0	13
J	9	0	2	17	0	0	0	0	0	0	19
K	10	0	4	26	2	0	0	0	0	0	32
L	11	0	2	16	0	0	0	0	0	0	18
M	12	0	5	28	9	0	0	0	0	0	42
N	13	0	3	25	5	0	0	0	0	0	33
O	14	0	0	33	0	0	0	0	0	0	33
P	15	0	1	26	1	0	0	0	0	0	28
Q	16	0	2	35	2	1	0	0	0	0	40
R	17	0	2	33	2	0	0	0	0	0	37
S	18	0	2	27	0	0	0	0	0	0	29
T	19	0	3	15	3	0	0	0	0	0	21
U	20	0	2	27	0	0	0	0	0	0	29
V	21	0	2	23	0	0	0	0	0	0	25
W	22	0	4	35	4	0	0	0	0	0	43
X	23	0	4	19	6	0	0	0	0	0	29
Y	24	0	3	18	1	0	0	0	0	0	22
Z	25	0	2	21	0	0	0	0	0	0	23
a	26	0	1	19	3	0	0	0	0	0	23
b	27	0	1	26	1	0	0	0	0	0	28
c	28	0	2	16	0	0	0	0	0	0	18
d	29	0	1	23	3	0	0	0	0	0	27
e	30	0	1	22	1	0	0	0	0	0	24
f	31	0	4	10	0	5	0	0	0	0	19
g	32	0	1	30	3	0	0	0	0	0	34
h	33	0	3	21	1	0	0	0	0	0	25
i	34	0	4	7	0	0	0	0	0	0	11
j	35	0	4	13	0	0	0	0	0	0	17
k	36	0	4	18	4	0	0	0	0	0	26
l	37	0	2	11	0	0	0	0	0	0	13
m	38	0	3	30	1	0	0	0	0	0	34
n	39	0	2	19	0	0	0	0	0	0	21
o	40	0	0	22	0	0	0	0	0	0	22
p	41	0	1	28	1	0	0	0	0	0	30
q	42	0	1	23	3	0	0	0	0	0	27
r	43	0	2	10	0	0	0	0	0	0	12
s	44	0	2	18	0	0	0	0	0	0	20
t	45	0	4	11	0	5	0	0	0	0	20
u	46	0	2	18	0	0	0	0	0	0	20
v	47	0	2	15	0	0	0	0	0	0	17
w	48	0	5	19	9	0	0	0	0	0	33
x	49	0	4	12	4	0	0	0	0	0	20
y	50	0	3	16	3	0	0	0	0	0	22
z	51	0	2	13	4	1	0	0	0	0	20

On analyzing the friend pattern it is observed that certain characters have unique friend pattern with unique total count and certain characters have unique friend pattern with same total count of significant pixels. In either of the cases characters can be uniquely identified as they have unique friend pattern.

Whereas certain characters have same friend pattern and same total count of significant pixels, when such situation is encountered then friend pattern cannot be only taken as the basis for identifying characters. Table 2 highlights the

characters that have same count of significant pixels sorted in ascending order.

Table 2:- Friend pattern chain and the count of significant pixels along with unique values.

CHAR	ID	0F	1F	2F	3F	4F	5F	6F	7F	8F	TOTAL	UNIQUE VALUES
i	34	0	4	7	0	0	0	0	0	0	11	1
r	43	0	2	10	0	0	0	0	0	0	12	2
I	8	0	2	11	0	0	0	0	0	0	13	3
l	37	0	2	11	0	0	0	0	0	0	13	
j	35	0	4	13	0	0	0	0	0	0	17	4
v	47	0	2	15	0	0	0	0	0	0	17	
L	11	0	2	16	0	0	0	0	0	0	18	5
c	28	0	2	16	0	0	0	0	0	0	18	
J	9	0	2	17	0	0	0	0	0	0	19	6
f	31	0	4	10	0	5	0	0	0	0	19	
s	44	0	2	18	0	0	0	0	0	0	20	7
t	45	0	4	11	0	5	0	0	0	0	20	
u	46	0	2	18	0	0	0	0	0	0	20	
x	49	0	4	12	4	0	0	0	0	0	20	
z	51	0	2	13	4	1	0	0	0	0	20	
T	19	0	3	15	3	0	0	0	0	0	21	8
n	39	0	2	19	0	0	0	0	0	0	21	
Y	24	0	3	18	1	0	0	0	0	0	22	9
o	40	0	0	22	0	0	0	0	0	0	22	
y	50	0	3	16	3	0	0	0	0	0	22	
F	5	0	3	19	1	0	0	0	0	0	23	10
Z	25	0	2	21	0	0	0	0	0	0	23	
a	26	0	1	19	3	0	0	0	0	0	23	
e	30	0	1	22	1	0	0	0	0	0	24	11
V	21	0	2	23	0	0	0	0	0	0	25	12
h	33	0	3	21	1	0	0	0	0	0	25	
C	2	0	2	24	0	0	0	0	0	0	26	13
k	36	0	4	18	4	0	0	0	0	0	26	
d	29	0	1	23	3	0	0	0	0	0	27	14
q	42	0	1	23	3	0	0	0	0	0	27	
P	15	0	1	26	1	0	0	0	0	0	28	15
b	27	0	1	26	1	0	0	0	0	0	28	
S	18	0	2	27	0	0	0	0	0	0	29	16
U	20	0	2	27	0	0	0	0	0	0	29	
X	23	0	4	19	6	0	0	0	0	0	29	
A	0	0	2	22	6	0	0	0	0	0	30	17
p	41	0	1	28	1	0	0	0	0	0	30	
G	6	0	2	30	0	0	0	0	0	0	32	18
H	7	0	4	26	2	0	0	0	0	0	32	
K	10	0	4	26	2	0	0	0	0	0	32	
E	4	0	3	29	1	0	0	0	0	0	33	19
N	13	0	3	25	5	0	0	0	0	0	33	
O	14	0	0	33	0	0	0	0	0	0	33	
w	48	0	5	19	9	0	0	0	0	0	33	
D	3	0	0	34	0	0	0	0	0	0	34	20
g	32	0	1	30	3	0	0	0	0	0	34	
m	38	0	3	30	1	0	0	0	0	0	34	
R	17	0	2	33	2	0	0	0	0	0	37	21
Q	16	0	2	35	2	1	0	0	0	0	40	22
B	1	0	0	37	4	0	0	0	0	0	41	23
M	12	0	5	28	9	0	0	0	0	0	42	23
W	22	0	4	35	4	0	0	0	0	0	43	25

On analyzing Table 2 it is observed that certain characters even after having same count of significant pixels have different

friend pattern chain which can be taken as the basis for uniquely identifying them. As incase the case of ‘F’ and ‘j’.

Whereas, certain characters have same total count of significant pixels, as well as same friend pattern chain. As incase the case of ‘L’ and ‘j’. Table 3 highlights those characters that have unique friend pattern chain and those with same friend pattern chain.

Table 3:- Friend pattern chain and the count of significant pixels along with unique values and commonalities.

CHAR	ID	0F	1F	2F	3F	4F	5F	6F	7F	8F	TOTAL	UNIQUE VALUES	COMMONALITIES
i	34	0	4	7	0	0	0	0	0	0	11	1	
r	43	0	2	10	0	0	0	0	0	0	12	2	
I	8	0	2	11	0	0	0	0	0	0	13	3	1
l	37	0	2	11	0	0	0	0	0	0	13		
j	35	0	4	13	0	0	0	0	0	0	17	4	
v	47	0	2	15	0	0	0	0	0	0	17		
L	11	0	2	16	0	0	0	0	0	0	18	5	2
c	28	0	2	16	0	0	0	0	0	0	18		
J	9	0	2	17	0	0	0	0	0	0	19	6	
f	31	0	4	10	0	5	0	0	0	0	19		
s	44	0	2	18	0	0	0	0	0	0	20		3
t	45	0	4	11	0	5	0	0	0	0	20		
u	46	0	2	18	0	0	0	0	0	0	20	7	3
x	49	0	4	12	4	0	0	0	0	0	20		
z	51	0	2	13	4	1	0	0	0	0	20		
T	19	0	3	15	3	0	0	0	0	0	21	8	
n	39	0	2	19	0	0	0	0	0	0	21		
Y	24	0	3	18	1	0	0	0	0	0	22		
o	40	0	0	22	0	0	0	0	0	0	22	9	
y	50	0	3	16	3	0	0	0	0	0	22		
F	5	0	3	19	1	0	0	0	0	0	23	10	
Z	25	0	2	21	0	0	0	0	0	0	23		
a	26	0	1	19	3	0	0	0	0	0	23		
e	30	0	1	22	1	0	0	0	0	0	24	11	
V	21	0	2	23	0	0	0	0	0	0	25	12	
h	33	0	3	21	1	0	0	0	0	0	25		
C	2	0	2	24	0	0	0	0	0	0	26	13	
k	36	0	4	18	4	0	0	0	0	0	26		
d	29	0	1	23	3	0	0	0	0	0	27	14	4
q	42	0	1	23	3	0	0	0	0	0	27		
P	15	0	1	26	1	0	0	0	0	0	28	15	5
b	27	0	1	26	1	0	0	0	0	0	28		
S	18	0	2	27	0	0	0	0	0	0	29	16	6
U	20	0	2	27	0	0	0	0	0	0	29		
X	23	0	4	19	6	0	0	0	0	0	29		
A	0	0	2	22	6	0	0	0	0	0	30	17	
p	41	0	1	28	1	0	0	0	0	0	30		
G	6	0	2	30	0	0	0	0	0	0	32		
H	7	0	4	26	2	0	0	0	0	0	32	18	7
K	10	0	4	26	2	0	0	0	0	0	32		
E	4	0	3	29	1	0	0	0	0	0	33		
N	13	0	3	25	5	0	0	0	0	0	33	19	
O	14	0	0	33	0	0	0	0	0	0	33		
w	48	0	5	19	9	0	0	0	0	0	33		
D	3	0	0	34	0	0	0	0	0	0	34		
g	32	0	1	30	3	0	0	0	0	0	34	20	
m	38	0	3	30	1	0	0	0	0	0	34		
R	17	0	2	33	2	0	0	0	0	0	37	21	
Q	16	0	2	35	2	1	0	0	0	0	40	22	
B	1	0	0	37	4	0	0	0	0	0	41	23	
M	12	0	5	28	9	0	0	0	0	0	42	23	

W	22	0	4	35	4	0	0	0	0	0	43	25
---	----	---	---	----	---	---	---	---	---	---	----	----

For those character that share same total count of significant pixels, as well as same friend pattern chain additional feature set can be associated. In this work distance between the least friends in the friend pattern chain and the Centroid of the character is taken into consideration for further association of parameters.

Centroid of a character can be determined by using the Centroid formulae. Let us consider a character represented by G, then the Centroid i.e. G(x, y) where in x and y are the coordinate of the Centroid, is given by

$$x = (\sum x_i) / \text{no of instances where } x_i \text{ represents } x \text{ coordinate of significant value.}$$

$$y = (\sum y_i) / \text{no of instances where } y_i \text{ represents } y \text{ coordinate of significant value.}$$

Character												x _i 's	y _i 's	x _i 's	y _i 's
												1	6	8	9
												2	5	9	3
												2	7	9	5
												3	4	9	6
												3	8	9	7
												4	4	9	8
												4	8	9	10
												5	4	10	2
												5	8	10	10
												6	3	11	2
												6	9	11	10
												7	3	12	1
												7	9	12	11
												8	4	12	12
												Summation		203	178
												Centroid		7	6

Table 4:- highlights the friend pattern chain and the Centroid associated with the friend pattern chain.

CHAR	ID	0F	1F	2F	3F	4F	5F	6F	7F	8F	TOTAL	UNIQUE VALUES	COMMONALITIES	CENTROID(X)	CENTROID(Y)
i	34	0	4	7	0	0	0	0	0	0	11	1		7	3
r	43	0	2	10	0	0	0	0	0	0	12	2		4	2
I	8	0	2	11	0	0	0	0	0	0	13	3	1	7	4
l	37	0	2	11	0	0	0	0	0	0	13			8	5
j	35	0	4	13	0	0	0	0	0	0	17	4		10	4
v	47	0	2	15	0	0	0	0	0	0	17		5	5	
L	11	0	2	16	0	0	0	0	0	0	18	5	2	8	3
c	28	0	2	16	0	0	0	0	0	0	18			4	4
J	9	0	2	17	0	0	0	0	0	0	19	6		8	6
f	31	0	4	10	0	5	0	0	0	0	19		6	4	
s	44	0	2	18	0	0	0	0	0	0	20	7	3	5	5
t	45	0	4	11	0	5	0	0	0	0	20			7	4
u	46	0	2	18	0	0	0	0	0	0	20		3	5	5
x	49	0	4	12	4	0	0	0	0	0	20			5	6
z	51	0	2	13	4	1	0	0	0	0	20		6	7	
T	19	0	3	15	3	0	0	0	0	0	21	8		4	6
n	39	0	2	19	0	0	0	0	0	0	21		4	5	
Y	24	0	3	18	1	0	0	0	0	0	22	9		6	7
o	40	0	0	22	0	0	0	0	0	0	22		5	5	
y	50	0	3	16	3	0	0	0	0	0	22		6	5	
F	5	0	3	19	1	0	0	0	0	0	23	10		6	4
Z	25	0	2	21	0	0	0	0	0	0	23		7	5	
a	26	0	1	19	3	0	0	0	0	0	23		5	5	
e	30	0	1	22	1	0	0	0	0	0	24	11		5	5
V	21	0	2	23	0	0	0	0	0	0	25	12		6	6
h	33	0	3	21	1	0	0	0	0	0	25		7	4	
C	2	0	2	24	0	0	0	0	0	0	26	13		6	6
k	36	0	4	18	4	0	0	0	0	0	26		7	4	
d	29	0	1	23	3	0	0	0	0	0	27	14	4	7	6
q	42	0	1	23	3	0	0	0	0	0	27			6	6
P	15	0	1	26	1	0	0	0	0	0	28	15	5	5	4
b	27	0	1	26	1	0	0	0	0	0	28			8	4
S	18	0	2	27	0	0	0	0	0	0	29	16	6	7	6
U	20	0	2	27	0	0	0	0	0	0	29			7	6
X	23	0	4	19	6	0	0	0	0	0	29		8	6	

b	27	0	1	26	1	0	0	0	0	0	28			8	4				
S	18	0	2	27	0	0	0	0	0	0	29	16	6	7	6				
U	20	0	2	27	0	0	0	0	0	0	29			7	6				
X	23	0	4	19	6	0	0	0	0	0	29			8	6				
A	0	0	2	22	6	0	0	0	0	0	30	17		9	7				
p	41	0	1	28	1	0	0	0	0	0	30		5	5					
G	6	0	2	30	0	0	0	0	0	0	32	18	7	7	6				
H	7	0	4	26	2	0	0	0	0	0	32			7	6				
K	10	0	4	26	2	0	0	0	0	0	32			8	6				
E	4	0	3	29	1	0	0	0	0	0	33	19		7	5				
N	13	0	3	25	5	0	0	0	0	0	33		7	6					
O	14	0	0	33	0	0	0	0	0	0	33		6	7					
w	48	0	5	19	9	0	0	0	0	0	33	20		5	7				
D	3	0	0	34	0	0	0	0	0	0	34		7	5					
g	32	0	1	30	3	0	0	0	0	0	34		7	6					
m	38	0	3	30	1	0	0	0	0	0	34		4	9					
R	17	0	2	33	2	0	0	0	0	0	37	21		6	6				
Q	16	0	2	35	2	1	0	0	0	0	40	22		7	7				
B	1	0	0	37	4	0	0	0	0	0	41	23		7	6				
M	12	0	5	28	9	0	0	0	0	0	42	23		7	7				
W	22	0	4	35	4	0	0	0	0	0	43	25		6	9				

3.5 Association of priority with the friends in the pattern

1. Association of priority with the friend pattern based on greater frequency value

Once the friend pattern chain for different characters are created and stored in a data structure, an appropriate comparison schema is to be developed that efficiently identifies the alternatives and appropriate character for an identified friend pattern chain. The comparison schema should effectively partition the set of friend pattern into probable set and non probable set in a particular iteration by selecting one frequency value from the friend pattern chain and the process is iterated by further selecting frequency values until match is not found. The effectiveness of the partitioning process highly relies on the way how frequency values are selected. So as to ease the

selection of frequency value a priority should be associated based on its significance to the friend pattern chains. Priority should be assigned in a manner that the frequency value that has the maximum instance is awarded the highest priority followed by instance that have lowest number of instances. To determine the frequency that needs to be assigned highest and subsequent priorities we need to see the distribution of the frequency values. Table 6 highlights the frequency and distribution of unique values in those frequencies. For example, in case of frequency value 1 out of 52 characters there are 4 characters with 0 one friend, 9 characters with 1 one friend, 15 characters each with 2 and 3 one friends, 8 characters with 4 one friends and 1 characters with 5 one friends

Table 6:- Determination of number of unique values in a particular frequency.

Frequency	0		1		2		3		4		5		6		7		8	
Distributions	0	52	0	4	7	1	0	23	0	48	0	52	0	52	0	52	0	52
	x	x	1	8	10	2	1	9	1	2	x	x	x	x	x	x	x	x
	x	x	2	20	11	3	2	4	5	2	x	x	x	x	x	x	x	x
	x	x	3	8	12	1	3	6	x	x	x	x	x	x	x	x	x	x
	x	x	4	10	13	2	4	5	x	x	x	x	x	x	x	x	x	x
	x	x	5	2	15	2	5	1	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	16	3	6	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	17	1	9	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	18	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	19	5	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	21	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	22	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	23	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	24	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	25	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	26	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	27	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	28	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	29	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	30	3	x	x	x	x	x	x	x	x	x	x	x	x
x	x	x	x	33	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	34	1	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	35	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	37	1	x	x	x	x	x	x	x	x	x	x	x	x	

(Assumption Priority 1 > 2 > 3 > 4 > 5 ...)

Fig 2:- Plot based on greater number of unique values

The above Table gives the distributing of unique values present in the friend pattern chain, frequency 2 has the highest number of unique values followed by 3, then 1 and then 4 where as all others have only one unique value. The above can be explained by creating a graph in which against each friend number the number of unique values is also highlighted.

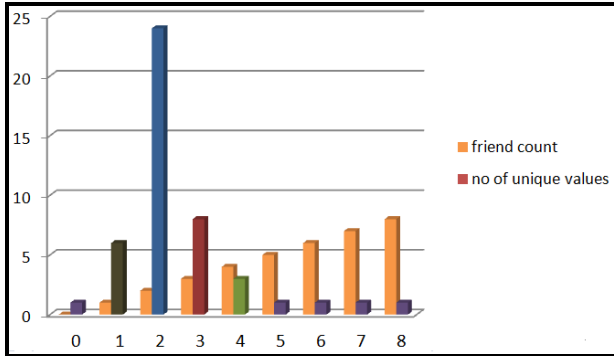


Table 7:- Assignment of priority based on greater number of significant unique values in the frequency.

Frequency	0		1		2		3		4		5		6		7		8	
Distributions	0	52	0	4	7	1	0	23	0	48	0	52	0	52	0	52	0	52
	x	x	1	8	10	2	1	9	1	2	x	x	x	x	x	x	x	x
	x	x	2	20	11	3	2	4	5	2	x	x	x	x	x	x	x	x
	x	x	3	8	12	1	3	6	x	x	x	x	x	x	x	x	x	x
	x	x	4	10	13	2	4	5	x	x	x	x	x	x	x	x	x	x
	x	x	5	2	15	2	5	1	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	16	3	6	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	17	1	9	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	18	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	19	5	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	21	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	22	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	23	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	24	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	25	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	26	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	27	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	28	2	x	x	x	x	x	x	x	x	x	x	x	x
x	x	x	x	29	1	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	30	3	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	33	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	34	1	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	35	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	37	1	x	x	x	x	x	x	x	x	x	x	x	x	
Priority	5		3		1		2		4		5		5		5		5	

Now the priority incorporated is to be associated with the base Table that has the friend pattern chain along with the distance from the Centroid. So while initiating the comparison procedure the friend pattern chain of the identified character is to be determined first and then the friend pattern chain is to be compared with the friend pattern chain in the database, if one instance of same friend pattern chain is encountered then the distance factor need not be taken into account where as if multiple instance of same friend pattern is encountered then further the distance from the Centroid can be taken into account in sequence to determine a suitable match.

Whenever priorities are assigned to the values in the friend pattern chain based on frequency the process of text detection highly restricts itself to the detecting text written in the same font and same size.

In the graph in Figure 2 blue represents highest priority, followed by brown, black, light green and purple.

2. Association of priority with the friend pattern based on smaller frequency value.

Completely contradicting the assumption taken as basis in the previous model what if the priorities are assigned in the reverse order i.e. the friend with maximum frequency count is assigned the maximum priority and the friend with minimum frequency count is assigned the maximum priority. Now, let us consider the case of character 'A' the friend pattern chain is as

	0	1	2	3	4	5	6	7	8	TOTAL
A	0	2	24	2	0	0	0	0	0	28

In order to ensure that the identified character is 'A' it has to have 2 pixels with three friends and 2 pixels with one friend and if the text is in same size 24 pixels with 2 friends. Now, what if the size of the text is different consider big or small. In such case it is observed that the not all the frequency values are

affected i.e. the unique values remains same where as the most frequent value increases.

So, taking into consideration the assumption made in the 2nd process the 1st process may be improvised to work with text in same font but different size.

Table 8:- Assignment of priority based on lesser number of significant unique values in the frequency.

Frequency	0		1		2		3		4		5		6		7		8	
	0	52	0	4	7	1	0	23	0	48	0	52	0	52	0	52	0	52
Distributions	x	x	1	8	10	2	1	9	1	2	x	x	x	x	x	x	x	x
	x	x	2	20	11	3	2	4	5	2	x	x	x	x	x	x	x	x
	x	x	3	8	12	1	3	6	x	x	x	x	x	x	x	x	x	x
	x	x	4	10	13	2	4	5	x	x	x	x	x	x	x	x	x	x
	x	x	5	2	15	2	5	1	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	16	3	6	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	17	1	9	2	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	18	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	19	5	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	21	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	22	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	23	3	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	24	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	25	1	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	26	4	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	27	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	28	2	x	x	x	x	x	x	x	x	x	x	x	x
	x	x	x	x	29	1	x	x	x	x	x	x	x	x	x	x	x	x
x	x	x	x	30	3	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	33	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	34	1	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	35	2	x	x	x	x	x	x	x	x	x	x	x	x	
x	x	x	x	37	1	x	x	x	x	x	x	x	x	x	x	x	x	
Priority	5		2		4		3		1		5		5		5		5	

(Assumption Priority 1 > 2 > 3 > 4 > 5 ...)

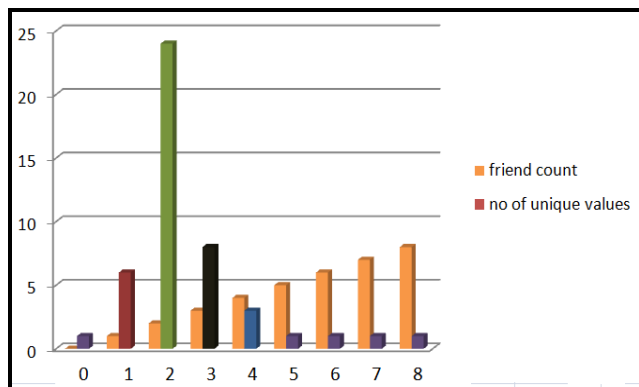


Fig 3:- Plot based on smaller number of unique values

In the graph in Figure 3 blue represents highest priority, followed by brown, black, light green and purple.

Determination of words

After completion of the reference data creation procedure, in order to determine the ability of the reference dataset in aiding identification of characters in a word, a test sample is to be read and then compared with the character definitions. While determining the words from the sample it is important for us to first determine the first character of the word, instead of the last or any intermediate characters. It is required because the text may be oriented in different directions. So, in

this case the traditional row column approach proves less effective.

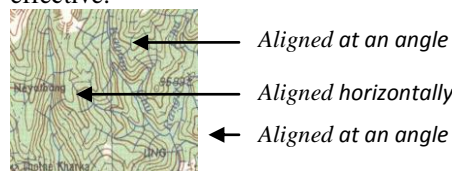





Fig 4:- Alignment of Text

Instead in this case rather than the traditional row column approach, a column row approach may be initiated that reads the sample data set taking into consideration column first. So for any text in the sample first the pattern of the first character will be encountered.

Further, in-order to determine the successive character in a word an odd order mask of minimum of the inter character separation distance is taken into account and then checked whether any significant value from another character fall in that mask or not. If yes then the process is continued else the word is considered to be terminated.

While, the image is initially segmented the features in the image that are represented using the same color code as that of the text are also identified as the part of the text to be interpreted. Therefore it is very much essential for us to eradicate these unwanted features. In order to clean the unwanted feature, following attributes such as length and friend pattern can be taken into account i.e. total number of friend pattern can be considered as length and the similarity in frequency chain. If any feature with non matching count and non matching chain is encountered then the feature may be discarded.

Test cases and results

Serial No	Input Image	Thinned Image	Character Id	Word
1	Gangtok		6,26,39,32,45,40,36	Gangtok
2	Sikkim		18,34,36,36,34,38	Sikkim
3	India		(8,37),39,29,34,26	India, (I,I i.e. capital I and small L has same definition as they are same in structure and orientation

4. CONCLUSION

This research initiative aims at introducing an alternative efficient method for identifying text from an image based on friend pattern chain. In this process first the friend pattern chain for various characters that may be used to compose the words was created and stored in the repository. In addition to the friend pattern chains, Centroid for individual characters was determined with respect to which distance of the individual significant coordinates was determined and stored in the repository followed by assignment of priority. Comparison was performed by determining the friend pattern chain of the identified character and matching it with the friend pattern chain stored in the repository. If there is a single candidate then the comparison stops else if there is more than one candidate then the comparison of the distance calculated from the Centroid was done in sequence with the priority assigned to find a suitable match. Matched characters were then grouped to form a word.

This process efficiently identifies the characters with reduced time and space requirement but demands proper initial segmentation of the reference image while identifying text. Advancement to the algorithm is also discussed to identify character in the same font but with different size.

5. REFERENCES

- [1] Gupta, A., Kumar, S., Gupta, R., Chaudhury, S. and Joshi, S. D., Enhancement of Old Manuscript Images, Proc. of ICDAR, Vol. 2, 744 – 748 pp., 2007.
- [2] Hamamoto, Y., Uchimura, S., Watanabe, M., Yasuda, T., Mitani, Y., and Tomita, S., A Gabor filter-based method for recognizing handwritten numerals, Pattern Recognition, Vol. 31, No. 4, 395-400 pp., 1998.
- [3] Wang, L. and Pavlidis, T., Direct gray-scale extraction of features for character recognition, IEEE Trans. PAMI, Vol. 15, No.10, 1053–1066 pp., 1993.
- [4] Lee, S.W. and Kim, Y.J., Direct Extraction of Topographic Features for Gray Scale Character Recognition, IEEE Trans. PAMI, Vol.17, No. 7, 724-729 pp., 1995.
- [5] Dubey, P., and Sinthupinyo, W., New Approach on Structural Feature Extraction for Character Recognition, Proc. of International Symposium on Communications and Information Technologies (ISCIT), 946-949 pp. 2010.
- [6] Poudroux, J., Toponym Recognition in Scanned Color Topographic Maps, Proc. of 9th International Conference on Document Analysis and Recognition (ICDAR), 531-535 pp., 2007.
- [7] Pezeshk, A., Automatic Feature Extraction and Text Recognition From Scanned Topographic Maps, IEEE Transactions on Geoscience and Remote Sensing, Vol. 49, Issue 12, 5047 – 5063 pp. 2011.
- [8] Chiang, Y.Y. and Knoblock, C. A., An Approach for Recognizing Text Labels in Raster Maps, Proc. of 20th International Conference on Pattern Recognition, 3199-3202 pp., 2010.
- [9] Caprioli, M., Gamba, P., Detecting and grouping words in topographic maps by means of perceptual concepts, Proc. of European Conference of Signal Processing (EUSPICON), 889-892 pp., 2000.
- [10] Feng Yang and Fan Yang et al., Character Recognition Using Parallel BP Neural Network, Proc. of International Conference on Audio, Language and Image Processing (ICALIP), 1595-1599 pp., 2008.
- [11] Velázquez, A., Levachkine, S., Text/Graphics Separation and Recognition in Raster-Scanned Color Cartographic Maps, Graphics Recognition-Recent Advances and Perspectives, Lecture Notes in Computer Science, Volume 3088, 63-74 pp., 2004.
- [12] Yamamoto, K., Yamada, H., Muraki, S., Symbol recognition and surface reconstruction from topographic map by parallel method, Proc. of the Second International Conference on Document Analysis and Recognition, 1993, pp 914-917.
- [13] Pezeshk, A., Improved Multi Angled Parallelism for separation of text from intersecting linear features in scanned topographic maps, proc. of IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), March 2010, pp. 1078 – 1081.
- [14] Luyang L, Nagy, G., Samal, A., Seth, S. , Yihong Xu, Cooperative text and line-art extraction from a topographic map, Proc. of the Fifth International Conference on Document Analysis and Recognition, 1999, PP. 467 – 470.
- [15] Anegawa, M., Shiku, O. , Nakamura, A. , Ohymat, T. ; Kuroda, H., A system for recognizing numeric strings from topographical maps, proc. of the Third International Conference on Document Analysis and Recognition, 1995, pp. 940-943.
- [16] Nakamura, A., Shiku, O. , Anegawa, M. , Nakamura, C. ; Kuroda, H., A method for recognizing character strings from maps using linguistic knowledge, Proc. of the Second International Conference on Document Analysis and Recognition, 1993., pp. 561 – 564.