

## Skew Detection & Correction in

# Scanned Document Images

## Avinash Chandra Kishan

Varun Sharda



Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela-769 008, Orissa, India

# Skew Detection & Correction in Scanned Document Images

Thesis submitted in partial fulfillment of the requirements for the degree of

### **Bachelor of Technology**

in

### **Computer Science and Engineering**

by

### Avinash Chandra Kishan

(Roll: 10506001)

### Varun Sharda

(Roll: 10506055)



Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela-769 008, Orissa, India May 2009



Department of Computer Science and Engineering National Institute of Technology Rourkela Rourkela-769 008, Orissa, India.

### Certificate

This is to certify that the work in the thesis entitled *Skew Detection & Correction In Scanned Document Images* by *Avinash Chandra Kishan & Varun Sharda* is a record of an original research work carried out by them under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering during the session 2005–2009 in the department of Computer Science and Engineering, National Institute of Technology Rourkela. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

Date: 11 May 2009

Pankaj Kumar Sa Lecturer CSE department of NIT Rourkela

### Acknowledgment

We would like to gratefully acknowledge the enthusiastic supervision and guidance of Prof. Pankaj Kumar Sa during this work. His suggestions and constant encouragement proved to be an immense source of motivation.

We are very much indebted to Prof. Banshidhar Majhi, Head-CSE, for alloting us this project and believing in us. Also for the resources and facilities that were available to us whenever we needed, that proved to be a vital part of the success of this work. Our sincere thanks to Prof. S.K. Rath, Prof. S.K. Jena, Prof. B. D. Sahoo, Prof. A. K. Turuk, Prof. D. P. Mohapatra, Prof. S. Chinara, Prof. P. M. Khilar, Prof. R. Baliarsingh, and Prof. K. S. Babu for being our knowledge resource. Their help can never be penned in words.

We would like to thank all our friends, especially Shikhar for helping us. We would also like to thank all those who have directly or indirectly contributed in the success of our work.

Last but not the least, NIT Rourkela for providing us such a platform where we learned a lot and gained so much experience.

Avinash Chandra Kishan

Varun Sharda

### Abstract

During document scanning, skew is inevitably introduced into the incoming document image. Skew detection is one the first operations to be applied to scanned documents when converting data to a digital format. Its aim is to align an image before processing because text segmentation and recognition methods require properly aligned next lines.

Different algorithms of skew detection are implemented. The first one is Scan line based skew detection. In this method the image is projected at several angles and the variance in the number of black pixels per projected scan line is determined. The angle at which the maximum variance occurs is the angle of skew.

The second one is based on the Hough transform. Hough transform is performed on the scanned document image and the variance in  $\rho$  values is calculated for each value of  $\theta$ . The angle that gives the maximum variance is the skew angle.

The third approach is based on the base-point method. Here a concept of basepoint is introduced. After the successive base-points in every text line within a suitable sub-region were selected as samples for the straight-line fitting. The average of these baseline directions is computed, which corresponds to the degree of skew of the whole document image.

All the above mentioned algorithm have been implemented and the results of each have been compared for accuracy.

# Contents

Ce	ertifie	cate	ii
A	cknov	wledgement	iii
A	bstra	let	iv
Li	st of	Figures	vii
Li	st of	Tables	viii
1	$\operatorname{Intr}$	oduction	1
	1.1	Image Processing	1
	1.2	Document Image Processing	3
	1.3	Problem Definition	5
	1.4	Motivation	5
	1.5	Thesis Organisation	5
<b>2</b>	Scar	n Line based Skew Detection	7
	2.1	Scan Line	7
		2.1.1 Algorithm	7
	2.2	Time Complexity	8
	2.3	Hough Transform	8
		2.3.1 Implementation of Hough Transform	14
	2.4	Time Complexity	15
3	Ske	w Detection based on Base-Point	20
	3.1	Selection of a sub-region and objects	20
		3.1.1 Sub-region selection	20
		3.1.2 Objects choosing	21

	3.2	Base-points clustering	22
		3.2.1 Definitions	22
		3.2.2 Base-points clustering	22
	3.3	Time Complexity	23
4	Con	clusions	28
	4.1	Achievements and Limitations of the work	28
	4.2	Further Development	29
Bi	bliog	graphy	30

# List of Figures

Deskewed document image of scanned book page with scan line scheme	9
Deskewed document image of scanned exam paper with scan line scheme	10
Deskewed document image of scanned telephone directory with scan	
line scheme	11
Deskewed images applying the Scan line scheme	12
Normal Representation of a line	13
Deskewed document image of scanned book page with the Hough trans-	
form scheme	16
Deskewed document image of scanned exam paper with Hough trans-	
form scheme	17
Deskewed document image of scanned telephone directory with Hough	
transform scheme	18
Deskewed images applying the Hough transform scheme	19
Representation of Bounding-box and Base-point for a scanned character	22
Deskewed document image of scanned book page with Base-point scheme	24
Deskewed document image of scanned exam paper with Base-point	
scheme	25
Deskewed document image of scanned telephone directory with Base-	
point scheme	26
Deskewed images applying the Base-point scheme	27
	Deskewed document image of scanned exam paper with scan line scheme Deskewed document image of scanned telephone directory with scan line scheme

# List of Tables

2.1	Comparative Results in Skew of different scanned document images as	
	calculated and original angles (as measured) using Scanline based skew	
	correction	8
2.2	Comparative Results in Skew of different scanned document images as	
	calculated and original angles (as measured) using Hough Transform	
	Method	15
3.1	Comparative Results in Skew of different scanned document images as	
	calculated and original angles (as measured) using Base-Point Method.	23

# Chapter 1 Introduction

### 1.1 Image Processing

The sense of vision has been one of the most vital senses for human survival and evolution. Humans use the visual system to see or acquire visual information, perceive, i .e. process and understand it and then deduce inferences from the perceived information. The field of image processing focuses on automating the process of gathering and processing visual information. The process of receiving and analyzing visual information by digital computer is called *digital image processing*.

An image may be described as a two-dimensional function I.

$$I = f(x, y) \tag{1.1}$$

where x and y are spatial coordinates. Amplitude of f at any pair of coordinates (x, y) is called intensity I or gray value of the image. When spatial coordinates and amplitude values are all finite, discrete quantities, the image is called digital image [1].

Digital image processing may be classified into various subbranches based on methods whose: [1]

- input and output are images and
- inputs may be images where as outputs are attributes extracted from those images.

Following is the list of different image processing functions based on the above two classes.

- Image Acquisition
- Image Enhancement
- Image Restoration
- Color Image Processing
- Multi-resolution Processing
- Compression
- Morphological Processing
- Segmentation
- Representation and Description
- Object Recognition

For the first seven functions the inputs and outputs are images where as for the rest three the outputs are attributes from the input images. With the exception of image acquisition and display most image processing functions are implemented in software. Image processing is characterized by specific solutions, hence the technique that works well in one area can be inadequate in another. The actual solution of a specific problem still requires a significant research and development [2].

Out of the ten sub-branches of digital image processing, cited above, this thesis deals with image restoration. In thesis various restoration methodology are used and various inputs are restored using these methods.

This is chapter is organized as follows. Document Image Processing is discussed in Section 1.2. The problem definition is described in Section 1.3. Motivation behind carrying out the work is stated in Section 1.4. Organization of the thesis is outlined in Section 1.5.

### 1.2 Document Image Processing

Methods for the creation and persistent storage of text have existed since the Mesopotamian clay tablets, the Chinese writings on bamboo and silk as well as the Egyptian writings on papyrus. For search and retrieval, methods for systematic archiving of complete documents in a library were developed by monks and by the clerks of emperors and kings in several cultures. However, the technology of editing an existing document by local addition and correction of text elements has a much younger history. Traditional copying and improvement of text was a painstakingly slow process, sometimes involving many man years for one single document of importance. The invention of the pencil and eraser in 1858 was one of the signs of things to come. The advent of the typing machine by Sholes in 1860 allowed for faster copying and a simultaneous on-the-fly editing of text. The computer, finally, allowed for a very convenient processing of text in digital form. However, even today, methods for generating a new document are still more advanced and mature than are the methods for processing an existing document.

This observation may sound unlikely to the fervent user of a particular common word-processing system, since creation and correction of documents seems to pose little problems. However, such a user has forgotten that his or her favorite wordprocessor software will only deal with a finite number of digital text formats. The transformation of the image of an existing paper document - without loss of content or layout - into a digital format which can be textually processed is mostly difficult and often impossible. Our user may try to circumvent the problem by using some available software pack- age for optical-character recognition (OCR). Current OCR software packages will do a reasonable job in aiding the user to convert the image into a document format which can be handled by a regular word-processing system, provided that there are optimal conditions with respect to:

- image quality
- separability of the text from its background image
- presence of standard character-font types

- absence of connected-cursive handwritten script and
- simplicity of page layout

Indeed, in those cases where strict constraints on content, character shape and layout do exist, current methods will even do quite a decent job in faithfully converting the character images to their corresponding strings of digital character codes in ASCII or Unicode. Examples of such applications are postal address reading or digit recognition on bank checks.

On the other hand, if the user wants to digitally process the hand-written diary of a grandparent or a newspaper snippet from the eighteenth century, the chances of success are still dim. Librarians and humanities researchers worldwide will still prefer to manually type ancient texts into their computer while copying from paper rather than entrusting their material to current text-recognition algorithms. Not only is the word processing of arbitrary-origin text images a considerable problem. Even if the goal can be reduced to a mere search and retrieval of relevant text from a large digital archive of hetero- geneous text images there are many stumbling blocks. Furthermore, surprisingly, not only the ancient texts are posing problems.

Even the processing of modern, digitally created text in various formats such as web pages with their mixed encoded and image-based textual content will require reverse engineering before such a digital document can be loaded into the word processor of the recipient.Indeed, classification of text within an image is so difficult that the presence of human users of a web site is often gauged by presenting them with a text fragment in a distorted rendering which is easy on the human reading system but an insurmountable stumbling block for current OCR systems. This weakness of the artificial reading system thus can be put to good use. The principle is known as CAPTCHA: Completely Automated Public Turing Tests to Tell Computers and Humans Apart . During recent years, yet another exciting challenge has become apparent in pattern-recognition research. The reading of text from natural scenes as recorded by a camera poses many problems, unless we are dealing with a heavily constrained application such as, e.g., the automatic recognition of letters and digits in snapshots of automobile license plates. Whereas license-plate recognition has become a mere technical problem, the camera-based reading of text in man-made environments, e.g., within support systems for blind persons, is only starting to show preliminary results. [3]

### **1.3** Problem Definition

Document Image Processing has many different tasks and methods to accomplish those tasks. During document scanning, skew is inevitably introduced into the incoming document image. Skew is any deviation of the image from that of the original document, which is not parallel to the horizontal or vertical. Skew Correction remains one of the vital part in Document Processing.

### 1.4 Motivation

A literature survey of the existing solutions to the problem of skew detection leads to the following conclusions:

- Solutions that provide accurate skew angles are slow.
- Solutions that reduce the required time result in lesser accuracy in skew angle determination.

So a trade-off between accuracy and time complexity is the motivation for the work.

### 1.5 Thesis Organisation

The rest of the thesis is organized as follows:

**Chapter 2** proposes two methods of skew detection. One is based on the *scanline method*. And the other one based on the *Hough transform*. This method outperforms its counterparts in terms of accuracy.

**Chapter 3** proposes a method based on the *base-point method*. But this one gives more emphasis on the connected component and is more accurate when the document has clear connected components.

Finally **Chapter 4** presents the concluding remarks, with scope for further research work.

# Chapter 2 Scan Line based Skew Detection

There are various methods to detect skew in an scanned document image. But here we focus on the methods based on scanline i.e. on the scanline method and on the Hough transform. Finally the result obtained from our implementation is added.

### 2.1 Scan Line

This method projects the image at several angles and determines the variation in the number of black pixels per projected scan line. The angle at which the maximum variance occurs is the angle of skew. [4]

#### 2.1.1 Algorithm

- Calculate the coordinates, in the image plane, for each of the parallel scan lines, that lie at a slope tan(θ) in the image plane. The coordinates are calculated using the Bresenham's Line Drawing Algorithm.
- For each scan line, count the number of non-background pixels that lie on the line.
- 3. Calculate the variance v in the number of black pixels that lie on each scan line for a given angle  $\theta$ .
- 4. The angle of skew  $\theta$  is given by the angle at which the maximum variance  $v_{max}$  is found.

$Angles \Rightarrow$	Calculated Skew $(\theta_s)$	Original Skew $(\theta)$
Scanned Document Images $\Downarrow$	Calculated Shew (0 <sub>s</sub> )	Original blocw (0)
fig 2.6(a)	3	2
fig 2.6(c)	2	2
fig $3.3(a)$	-8	-8
fig $3.3(c)$	6	6
fig 3.4(a)	-7	-7
fig $3.4(c)$	15	15
fig $3.5(a)$	16	16
fig $3.5(c)$	-18	-18

Table 2.1: Comparative Results in Skew of different scanned document images as calculated and original angles (as measured) using Scanline based skew correction.

The following are the results of the correction figures of the scanned document images using Scanline based detection : fig. 2.1 , fig. 2.2, fig. 2.3, fig. 2.4

### 2.2 Time Complexity

Let N be the number of pixels in the scanned document image. Then, the time complexity is O(N).

### 2.3 Hough Transform

The Hough transform is an edge linking and boundary detection technique used in image analysis. The purpose of the technique is to find instances of objects within a certain class of shapes by a voting procedure. This voting procedure is carried out in a parameter space, from which object candidates are obtained as local maxima in a so-called accumulator space that is explicitly constructed by the algorithm for computing the Hough transform.

Consider a point  $(x_i, y_i)$  in the xy-plane and the general equation of a straight line in slope-intercept form,  $y_i = ax_i + b$ . Infinitely many lines pass through  $(x_i, y_i)$ , but they all satisfy the equation for  $y_i = ax_i + b$  for varying values of a and b. However writing this equation as  $b = -x_i a + y_i$  and considering the ab-plane (also THE SOCIAL EMERGENCY

Set THE SOCIAL RAHIFUENCY laundrysgick, for estample, who, after a  $(a_{p'_1})$ work of ten hours at the unachines, go at tight to their bounding-bouses where they was dishes to elso aut a Kving. — then this program would not be unlike the address of a physica who tells a poor man with tuberculasis that is unset go to the country for a year and live an error and ense.

cream and eggs. Even in the case of wage-earning girls who adopt loose ways to satisfy extravagant design. eream and eges. Were in the case of wags-earning girls we adopt loose ways to satisfy extravagant design their tasks are established by women of the weathy and mildle classes. The Island the ters, who in factories and mills make the dol, propage food stuffs, and do all sorts of task that formarily kept women of the upper class at home. Through the instinct of initiation which will be able to the well to do determine the ambition of many a working-sint. Other factors are industrial arrangements which segregate men in construction and has been employed the they ear, and the which segregate men in construction and has been employed the their further employ-ment, turn them loose into cities where our saloons welcome them and each their checks

THE SOCIAL EMERGENCY

and yiels, for example, when due a day, work of ten hours at the machines, so at high to their boarding - houses where they wai, dishes to ske out a living - then this present, would not be unlike the advice of a playing who tells a poor man with inherealosis that is, must go to the country for a year and live as erram and eggs.

must go to the country for a year and live as cream and eggs. Even in the case of wage-enring give also adopt loace ways to antidy extravagant itesing, their basiss are established by woman of the weakly and middle classes. The leave of these women is due to their wage-enring si-ters, who in factories and mills make the ded, prepare food-stuffs, and do all sorts of tab-that the formerly kept women of the upper classe at home. Through the instinct of ministica, combined with the American feeling of denor-racy, the habits of the well-to-do determine the ambition of many a working-girl.

ricy, the labels of the well-to-do detamin the ambition of many a working-girld. Other factors are industrial arrangements which segregate men in construction and hav-ber camps for a part of the year, and the without providing for their further employ-ment, turn them loose into clicks where oil subcons welcome them and cash their chefds

#### (a) First scanned image from book

60

#### (b) Corrected image of (a)

#### THE SOCIAL EMERGENCY

should be here made of this report, because 2,3 total effect is to minimize economic usage of prostitution, placing the responsibility else, where than on industrial conditions. It is to be noted, however, that it does emphasize the indirect effects of paverty, and dose speak of the moral danger lurking in certain occupation, and of the had effects of the lack of industria-

and of the had effects of the lack of ministry education. More definite responsibility for vice is as ericed to low wages in the reports of vice exp-missions. The Chicago Report says that d one group of 119 immeral worner, 18 cause from department storm, and 35 sold that they had department stores, and 38 and that they had taken up the career for the need of meany. The portland Report presents 22 works as "Cases in which Low Wage and Vice se closely associated."<sup>1</sup> The Report continees:-

In presenting the forspoing table and statement from girls, this commission does not take the pair too that the low ways of self-supporting girls is it side continuing cause of their d-kingmenty, calls ing that there are through of girls who wold of dure the utmost handbings before yielding there wives to those who are ready to schere them. To 1 Report of Forthand Vice Commission, p. 175.

THE SOCIAL EMERGENCY

e0 THE SOCIAL EMERGENCY should be here made of this report, hereause is, total effect is to maintaine economic ensu-of prestitution, plasing the responsibility eks-where than on industrial conditions. It is to be noted, herewere, that it does emphasize the indirect effects of poverty, and does speak at the moval darager lexism in certain comparison, and of the had effects of the lack of industrial education. More definite responsibility for vice is as enhed to low wages in the reports of vice res-related to low wages in the reports of vice res-related to low wages in the reports of vice res-relation.

More definite responsibility for vice is as critical to low wages in the reports of vice each missions. The Chicago Report says that of one group of 113 instead works, 113 critics fraw department stores, and 38 wild that they had taken up the caseer for the need of mone. The Portlard Report presents 22 works as "Chases in which Low Wage and Vice are closely associated."<sup>2</sup> The Report continues:-

In presenting the foregoing table and statements from goin, this mean issue of each state the pos-tion that the law expected self emporting gives be adde contributions cause of their definity-may, main ing that there are thousands of girls who work in-duce the utmost hatdhigh before yielding there selves to those who are ready to achieve them. The <sup>1</sup> Argert of Fortland Vice Commission, p. 178.

(c) Second scanned image from book

(d) Corrected image of (c)

Figure 2.1: Deskewed document image of scanned book page with scan line scheme





- (a) First scanned exam paper image
- (b) Corrected image of (a)

Q NATIONAL INSTITUTE OF TECHNOLOGY

EXAN

I wate Separate of RideN

528813181 1001-20.03

/1-5a)-1

CR BOOK

These & Algorithms

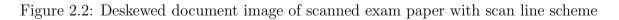
A Nº 03002

NATIONAL INS	TITUTE OF TECH ROURKELA	A Nº 030026
		Exercitation
To be taked in by the Standard.	Servestor	Mit Serrester
Academic TEE	CT Autumn	CrEnt Semester
2002-2007	Cit taring lind	Budgiementary
and Kennerghon	CT Summer	L.J. Com
29,104,00.07	LI	
Course Title	Shuicherts & All	Poli No 10 S MIND 3 6
De	BASIO	Last .
Name of Student	BASIO M.Sc	TERP
Name of Success Programme : 2 B. Tech.	BASIO	TIPLE
Nama of Buddent Programme : D B. Tech.		Ph.D. Publicie Rol
Nama of Buddent Programme : D B. Tech.	BASIS MSC	Ph.D. Publication Raf Signature of Studen Signature follow
Name of Bulloon Programme	E-DS10	Ph.D. Publication Raf Signature of Studen Signature follow
Name of Bulloon Programme	E-DS10	Ph.D. Publication Raf Signature of Studen Signature follow
Name of Stations - Programme - D B. Tech.	Internation meteration of internation MCA MCA MCA	PR. D. Pr. D. Bagnature of Blueben a barbantor of theoremation total by the exempter of the pager for response for
Name of Boolen. Programme () B. Tech. M. Tech. To be titled in by the Invalident. In use of server Invalued on Proceedings	MERSIS MCA MCA MCA MCA MCA	Ph.D. Publication Raf Signature of Studen Signature follow

(c) Second scanned exam paper image



terrer to be de 55 million and for all of a little of any in same of main inclusion of the external of the strengt two weeks model on all the inclusion for all one form the filled any by the samples of the same of the Leasting and 1990.



Camputer Spience & Engl Office	isaoving:	4208	Protes	A Bera, Sri G. Beheraj E Brastachuryva, Bri P. K. M	Arturna A. Sab
HOD Phile Mary Fix: Facing Binesing R. Dinas S. Ana S. Ana S. Mary B. Mary B.	22 23 23 235 235 235 235 2357 2352 2358 2359 2359 2359 2359 2359	2380 2351 059 67 23 2 2 30 30 30 238 238 238 3380	134 (3, 7 3355 3359 3367 3367 3359 3359 3359 3359 3359 3359 3359 335	Ayest, Sr.K. Ayyas) hool-ca robingh suchanang Gima shipsa anga baaja sakajbaa sakajbaa baany Maruu A	8-21 D-64 TF-13 A-8 D-17 D-6 O-29 A-13 TF-14
Image Processing & Cluster computing Lab Information Security & Communication Lab	4352 4353 4354	Prof. B	Math, St K. Tumat	iA.K. Day ; Sri M. K. Patru M.K. Patruskj Sr J. P. H. Nacj	nkj

Computer Science & Engineering	4256	Prof. S. Bhartach	anyya, Sei P. K. Mahanty	d
Ther	2350	(Sri G. Nayak: Sri	K Naiski	
400 (Prof. B. Majhi)	2351	3365	hod-ci	
1AX	2351			
Faculty				
Balantingh R.	2555	3359	rbeingh	8-21
Drimana 5.	2361	3367	acchients.	0.84
ana S	2183	3353	tioné	77-13
ima S.K.	2352	3352	skiona	A+0
Other #3d	2354	3354	prektilse	D-17
dapte IL.	2355	3355	bright	D-6
Achapatra O.P.	2356	3356	durps -	D-29
Rich S.K.	2397	3257	avat-	A-15
la Pirikij Kumar	2382	3362	pankajkoa	TF-14
Satury B.D.	2388	-	bduahu	
TURDIK AUK	2360	3360	attunk	0.22
oboratories & Facilities			103	
Software Laboratory ( My X. Trust multer	4351	Prof. B. Majhi, Sri	A K Des ; SI M K P	theres
Advanced Database Engg. Lab	4552	[Pht.S.K.Jeta]		
mage Processing & Cluster computing Lab	4353	Prof. B. Majh, Sri	M.K. Patranj.	
rfurnation Security & Communication Lals	4354	[Prof. A. K. Tunik]		
Fardware Lilo.	4555	Prof. R. B. Singh.	StiJ.P.H. Negl	
		-		

#### (a) First scanned telephone directory image

#### (b) Corrected image of (a)

	Tomponiture Laboratory	4005			
	jetna Analysis Laboratory	42010			
	darkalarip	4700	(Prof. 3 W. even		
Laboration 430 prof 5.1 mm 5.1 became Laboration 430 prof. J beca 5.1 beca 5.1 beca 5.1 beca	A Pory Laboratory	1307	(Prol. J. Berg, Ski G. Baller 3)		
Lander 400 pet 5. Constant Manual Lander 400 pet 3. Constant Annual Constant Annual Constant Constant Annual Constant Co	Exp-Ceremic Laboratory	4366	(Pret & Rhamonarya, Sr P	( Matarity	
Lasonan dan jold S. Francisco and an anti- stantan dan jold J. Ban Sold advant. National dan jold J. Ban Sold advant.	Computer Science & Engine	cinci (C3)			
	Offin	2150	[81 G. Nayal, Sri K. Naya#]		
John South State 100 (20) (10 C Legal St K Mar Land	DO PARE MAN	2151	3355 ho:	(Z)	
and the second termination of the second sec	FAX	2.01			
205 8.21	Faculty				1
phot B Maphi 2351 escret 0.14 updemain 17-13	Delersingh R.	2349	3899 (12)	rgn B-21	
ipid- TE-53	Ostara S	Z301	336*	kienits D-84	
200 101 100 4.0	Jose S.	2353	3153 #4	u (F-13	
2341 1353 pages 0.11	Jane G.X.	2353		B-A Ers	
	Khia Phi	2054	7.54 58	ther D.T	
	Mohi B.	2355	30.75 0.00	ally D+6	
ea Sha well with the la	Mohapatra D.P.	2356	3050	01 E-29	
	Forb 8 K	2357	3337 st	uth A+15	
100 PM 205 105 10.22	Ca Penilos Roman	738.7	3312	egina TF-14	
1121 B 2224 1288	Jahu B.D.	251		ichu	
unit in the second seco	Turus A.K	236;	3050 avd	UNA 17+28	
	abcrateries & Facilities				
Rub Sh. Su Peral Kimel 730 300 La Malt Sr A. K. Dai Sela K.	Schware Laboratory ( MI-Y 11KA	n=171 4351	(Free 5 Might SelA K Dar ;	Sri M. (Patrick)	
San BD. San BD. Tank K. R. Socialities Junited (100) Tank K. R. Socialities Junited (100) San B. S. Januari, J. San Januari, Januari	Aniversited Enteringene Encouncies	1352	Prof. 3 K Jane		
10110 ·	image Percending & Cluster comp.		(Prof. 8 Majhé Sri N. K. Pelne		
	Internation Security & Communicati		Print & K Turke"		
School Database Bolar concepting Law COM mod R & Smool	Harpene Lab.	4355	Frot. R. B. Sech, 34 J. P. R.	last.	
Solar Landral (Value) (2010) (2010) Solar Landra Caser and Call (2010) (2010) Solar Antonia Caser and Call (2010) (2010) (2010) Solar Antonia Caser and Call (2010) (2010	CONSIGNATION AND DESCRIPTION		Concert agend a la conce		
Advant United and Advance and Adva	The local diversion of		1	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	
Horizon Lib.				and the second second	
Harden					

- (c) Second scanned telephone directory image
- (d) Corrected image of (c)

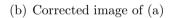
Figure 2.3: Deskewed document image of scanned telephone directory with scan line scheme



(a) Third scanned telephone directory image



(c) Fourth Scanned telephone directory image





(d) Corrected image of (c)

Figure 2.4: Deskewed images applying the Scan line scheme

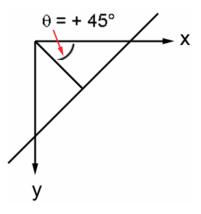


Figure 2.5: Normal Representation of a line

called parameter space) yield the equation of a single line for a fixed pair  $(x_i, y_i)$ . Furthermore, a second point  $(x_j, y_j)$  also has a line in parameter space associated with it, and unless they are parallel, this line intersects the line associated with  $(x_i, y_i)$  at some point (a', b'), where a' is the slope and b' the intercept of the line containing both  $(x_i, y_i)$  and  $(x_j, y_j)$  in the xy-plane. In fact, all the points on this line have lines in parameter space that intersect at (a', b').

In principle, the parameter-space lines corresponding to all points  $(x_k, y_k)$  in the xy-plane could be plotted, and the principal lines in that plane could be found by identifying points in parameter space where large numbers of parameter-space intersect. A practical difficulty with this approach, however, is that a (the slope of a line) approaches infinity as the line approaches the vertical direction. This problem can be solved by using normal representation of a line:

$$x\cos\theta + y\sin\theta = \rho$$

Fig. 2.5 represents the geometrical interpretation of the parameters  $\theta$  and  $\rho$ . A horizontal line has  $\theta = 0^{\circ}$ , with  $\rho$  being equal to positive *x*-intercept. Similarly, a vertical line has  $\theta = 90$ , with  $\rho$  being equal to positive *y*-intercept or  $\theta = -90$ , with  $\rho$  being equal to negative *y*-intercept. Each sinusoidal curve in figure 1.2 represents the family of lines that pass through a particular point  $(x_k, y_k)$  in the *xy*-plane. The intersection point  $(\rho', \theta')$  in fig. 2.5 corresponds to the line that passes through both  $(x_i, y_i)$  and  $(x_j, y_j)$  in fig. 2.5.

The computational attractiveness of the Hough Transform arises from subdividing the  $\rho\theta$  parameter space into so-called accumulator cells where  $(\rho_{min}, \rho_{max})$  and  $(\theta_{min}, \theta_{max})$  are the expected ranges of the parameter values:  $-90^{\circ} \leq \theta \leq 90^{\circ}$  and  $-D \leq \rho \leq D$ , where D is the maximum distance between opposite corners in an image. The cell at coordinates (i, j), with accumulator value A(i, j), corresponds to the square associated with parameter-space coordinates  $(\rho_i, \theta_j)$ . Initially, these cells are set to zero. Then, for every non-background point  $(x_k, y_k)$  in the xy-plane, we let  $\theta$  equal each of the allowed subdivision values on the  $\theta$  axis and solve for the corresponding  $\rho$  using the equation  $\rho = x_k \cos\theta + y_k \sin\theta$ . The resulting  $\rho$  values are then rounded off to the nearest allowed cell value along  $\rho$  axis. If a choice of an angle  $\theta_q$  results in solution  $\rho_p$ , then we let A(p,q) = A(p,q) + 1. The number of subdivisions in the  $\rho\theta$ -plane determines the accuracy of the collinearity of the points.

The steps for finding skew angle using Hough transform is as follows:

- 1. For each non-background pixel  $P(x_i', y_i')$ .
- 2. Calculate the corresponding values of  $\rho$ ,  $\rho_j$  for each  $-90 \le \theta_i \le 90$ . The value of  $\rho$  is rounded of to the nearest allowed cell value along  $\rho$  axis.
- 3. Increment the corresponding Hough matrix cell, H(j, i), by one.

The above process results in a Hough matrix, whose each cell, (i, j), gives the number of points that lie on the line with parameters  $\rho$  and  $\theta$ ,  $(\rho_i, \theta_j)$ . Each column of the Hough matrix gives all points that lie on a set of parallel lines, irrespective of the  $\rho$  values. Thus, finding the variance v of the values along each column gives us the variance in the number of background pixels that lie on a set of parallel scan lines. Again, the angle of skew is the angle at which this variance is maximum.

#### 2.3.1 Implementation of Hough Transform

An input image is taken and Hough transform was implemented. The value of  $\theta$  was incremented across the rows and the value of  $\rho$  was incremented across the columns. The variance was calculated for values in each column i.e. variance in the number of

$Angles \Rightarrow$	Calculated Skew $(\theta_s)$	Original Skew $(\theta)$
Scanned Document Images $\Downarrow$	Calculated Silew $(v_s)$	Oliginal bicw (0)
fig $2.6(a)$	1	2
fig 2.6(c)	1	2
fig $3.3(a)$	-8	-8
fig $3.3(c)$	6	6
fig $3.4(a)$	-7	-7
fig $3.4(c)$	15	15
fig $3.5(a)$	16	16
fig $3.5(c)$	-18	-18

Table 2.2: Comparative Results in Skew of different scanned document images as calculated and original angles (as measured) using Hough Transform Method.

points between the number of points that lie on a set of parallel lines. The  $\theta$  that gave the maximum variance is the skew angle.

The following are the results of the corrected figures of the scanned document images using hough based detection: fig. 2.7, fig. 2.8, fig. 2.9

### 2.4 Time Complexity

Let  $N_{nb}$  be the number of non-background pixels in the scanned document image. And let  $N_{\theta}$  be the number of angles used to calculate the Hough Transform. Then the time complexity is  $O(N_{nb}N_{\theta})$ 

#### THE SOCIAL EMERGENCY

66 THE EXCLUSION PARAMETERS (1) laundry-girls, for example, who, after a  $d_{M_1'}$ work of ten hours at the warhines, go at high; to their boarding-bouses where they was dishes to else out a living. — then this program would not be unlike the advice of a physica  $d_{M_1}$  by the program with tuberculusis that 1. would not be unlike the notices of a payming who tells a poor man with tuberculosis that is unset go to the country for a year and love a cream and agay. Even in the case of wage-earning girls who

eream taid 0.029. Twen in the case of wage-earning girls whe adopt loose ways to satisfy extravagant design, their tastes are established by women of the wealthy and mixile classes. The kissure of these women is due to their wage-earning is-ters, who in factories and mills make the ool-prepare food stuffs, and do all sorts of tash that formarly kept scenars of the upper class at home. Through the instinct of insisten-routhined with the Auserian feeling of demis into a many a working-sid. . Other factors are industrial arrangements be completed and many a working-sid. . Other factors are industrial arrangements be completed for a part of the year, and they without providing for their further employ ment, turn them loose into cities where our saloons welcome them and cash their classi

THE SOCIAL EMERGENCY

should be here made of this report, because its abould be here made of this report, because it, total effect is to minimize economic same of prostitution, placing the responsibility eig-where than on industrial conditions. It is a be noted, however, that it diese emphasize the indirect effects of poverty, and does apeak of the moral danger lutking in certain occupation, and of the bad effects of the lack of industral education.

education. More definite responsibility for vice is as-celled to how wages in the reports of vice con-missions. The Chicago Report says that of one group of 119 immoral women, 18 eams fram department stores, and 28 solid that they hal taken up the career for the need of mong-The Portland Report propents 22 women as "Cases in which Low Wage and Vice are closely associated."<sup>11</sup> The Report continues:-

In presenting the foregoing table and statuments from girls, this commission does not take the pet-tion that the low wags of self-supporting galaxies does contributing cause of this followports, real-ing that there are thousands of girls who would show the utnost hardships before yielding them-selves to those who are ready to solice them. The \* Report of Portland Vice Commission, p. 176.

#### (a) First scanned image from book

(b) Resulting image from first scanned image

THE SOCIAL EMERGENCY should be here made of this report, because as should be here made of this report, because is, total effect is to maintize economic unity, of prostitution, placing the responsibility ele-where than an industrial conditions. It is to he noted, however, that it does emphasize in indirect effects of paverty, and does used at the moral danger inriting in certain occupation, and of the had effects of the lack of industrial education.

education. More definite responsibility for vice is as crited to low wages in the reports of vice exa-missions. The Chicago Report says that of one missions. The Chicago Report says that it one group of 119 immoral worker. Is same from department shores, and 35 said that they had taken up the career for the need of mong. The Portland Report presents 23 women as "Cases in which Low Wage and Vice see closely associated."<sup>1</sup> The Report continees:-

In presenting the locupoing to be and statement from girk, this commission does not take the pair ton that the low ways of self-supporting grin is in side contributing cause of their delong mericy of the ing that there are threasand of grins who could once the unnext banddings before yielding ther-selves to those who are ready to solve them. To \* Report of Forthand Vice Commission, p. 179.

#### THE SOCIAL EMERGENCY

should be here made of this report, because its should be here made to this report, occame hy total effect is to minimize economic cause, of prostitution, placing the responsibility else, where than on industrial conditions. It is to be noted, however, that it does emphasize the indirect effects of poverty, and does speak of the moral danger lurking in certain occupation, and of the bad effects of the lack of industria education.

education. More definite responsibility for vice is as-reliade to low wages in the reports of vice con-missions. The Chicaro Report says that of me-group of 110 immonal women, 18 came from department stores, and 88 said that they had taken up the carver for the need of meany. The Portland Report presents 32 women as "Cases in which Low Wage and Vice are closely associated."<sup>1</sup> The Report continues.—

In presenting the foregoing table and statement from girls, this examination does not take the pai-tion that the low wages of self supporting girls in the also contributing cause of their delanguary, self-ing that there are thousands of girls sub-would en-dance the utness harddings before yielding their solves to those who are ready to seduce them. The \* Report of Portland Vice Commission, p. 178.

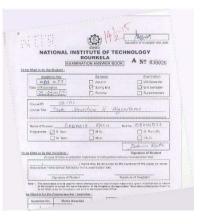
(c) Second scanned image from book (d) Resulting

scanned image

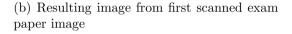
image from second

Figure 2.6: Deskewed document image of scanned book page with the Hough transform scheme



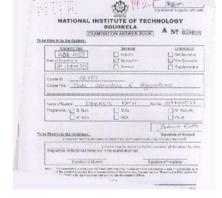


(a) First scanned exam paper image



	NATION ANSWER BOOK	A Nº 030026
To be filled in by the Student.: <u>Academy, Tak</u> 20 <u>QL</u> - 20 <u>QT</u> Date of Examinator	Sensitiat Auturni Storing Ked	Mit Semester
Course Title Thata	Structure & Hig	oruthims
De	BASIS RATH	Roll NO 108 MIND 3
Name of BuddentE Programme : B. Tech. M. Tech.	BASIS RATH	2 Ph. D. Zudanie Ro

(c) Second scanned exam paper image



(d) Resulting image from second scanned exam paper image

Figure 2.7: Deskewed document image of scanned exam paper with Hough transform scheme

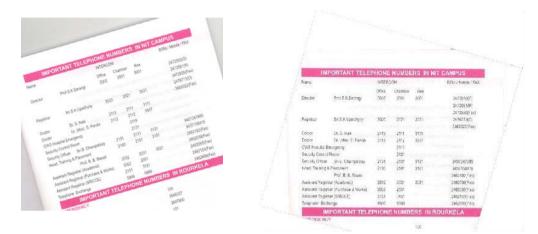
			Prior J. Bera, Sri G. Beh	era]	
		4208	[Prof. S. Bhattacharyya, S	in P. K. Mohanty]	
	Computer Science & Engineering	: (CS)			
	Office	2350	[Sri G. Nayak; Sri K. Naya	ak]	
Camputer & Engliseering: (Cs)	HOD [Prof. B. Majhi]	2351	3355	hod-cs	
	FAX	2351			
Prof. B. Marca	Faculty				
	Baliarsingh R.	2359	3359	rbsingh	B-21
Pacuity 2351 holica Batatringh R.	Chinara S.	2361	3361	suchismita	D - 64
Chitary 6. 2359 Hate	Jena S.	2353	3353	sjena	TF-13
Junu 5. 2361 Steel rooman	Jena S.K.	2352	3352	skjena	A-8
Arca S.K. 2353 JUS3 Richards D. 4	Khilar P.M.	2354	3354	pmkhilar	D -17
2002 2002 Elena TE	Maihi B.	2355	3355	bmajhi	D-6
2254 3354 Motore A-8	Mohapatra D.P.	2356	3356	durga	D - 29
Rath S.K. 21Ke house	Rath S.K.	2357	3357	skrath	A - 15
Sa Paring Kumar 2357 North Deta	Sa Pankaj Kumar	2362	3362	pankalksa	TE-14
Sahu B.D. 2902 steel Mean 0-29	Sahu B.D.	2358		bdsahu	
Tuta A K 2368 Rest	Turuk A.K.	2360	3360	akturuk	D-22
	Laboratories & Facilities			Contract of Contra	
Laboratory (M-K-C)	Software Laboratory (M. K. P. H wate	1 4351	(Prof. B. Maihi, Sri A. K. D	las : Sri M. K. Patr	naiki
Million Provinces and All States	Advanced Database Engg. Leb	4352	[Prof. S. K. Jena]	an brodil (1997)	
rage Processing & Cluster computing Lab 452 Prof. & Maph, Bi A. K. Das ; Bir M. K. Pathakij formation Security & Cluster computing Lab 433 Prof. & K. Anaj Hornes Lab, Prof. & K. Pathakij	Image Processing & Cluster computing Lab	4353	(Prof. B. Maihi, Sri M. K. F	Patnaiki	
theary Lab. (Prof. A. K. Tursul)	Information Security & Communication Lab	4354	(Prof. A. K. Turuki	CALL ST	
4365 (Prof. R. B. Singh, Sn.J. R. H. Nag)	Hardware Lab	4355	[Prof. R. B. Singh, Sri J. P	H Naol	
and an and a second and a secon			P	and a second sec	
			9		

(a) First scanned image from telephone direc- (b) Resulting image from first scanned image tory from telephone directory

	irma Anarysis Laboratory lackstrig	4205 4205	Frot. 3 B. rea		
	<ul> <li>Provi unborotary</li> </ul>	1307	Frol. J. Bers. 3		
4509 and 5 K The general would	Cito-Detamic Laboratory	4,786		natys, Sr P K. Motort	1
400 port 5.K. Jack Sold Laterall, 400 port 3 Rev. 50 d Laterall, 100 port 3 Rev. 50 d Laterall,	Computer Science & Engine		The start of the start of	contraction of another of	0. 1354
	Office				
428 pays a family for a family for a family for a family for a	ICC Prof IS Made	2161	1301	hoc or	
Operation (CA)         Data Marka Ser K         Data           250         255         505         6,17           Mark         255         8,17	FAX	Zm	100013	11111	
2050 2055 0.21	Family				
2351 (build) 0.64	Delignation R	2349	3369	rtzs-reat	8-21
and and a second second	Obstana S	2321	736*	an thirth an	D-54
*69 SEC A.**	Jorn S	2353	3353	sena	15-13
ADD D D D D D D D D D D D D D D D D D D	Jana G.K.	2352	3362	Anni	A-5
7253 1551 press	Shia PM	2054	1334	renshi or	0.1
1004	Mobile.	2356	35/95	Cenasilie	B-6
100 555	Mohapatha D.P.	3350	3030	turge.	E-29
koles stal	Forth 8 X	3357	3337	strath	A - 15
10 10 10 10 10 10 10 10 10 10 10 10 10 1	Sa Penke Northe	7387	3982	panénjesa.	15-14
554	sale B.D.	2362		txtachu	
258	Turus A.K.	2385	3350	aktinuk	6+27
7960 3364 K. Data (Shi M. N.	Lancratorica & Facilities			5.226	
7955 7950 7950 7950 9956 8 Mallo BAX K. Dan, Ser M. K. Portali 1957 8 Mallo BAX K. Dan, Ser M. K. Portali	Bothware Caboratory ( M)-Y 11404	n=34; ) 4351	Free B Megn:	SHA K, Dat ; SHM K P	anai)
the test of test o	furtement Database Engly, Lak	1352	Frot 3 K	e:	
Party And Albert and Break	image Property & Cluster comput	ng Lat: 4385	(Prof. 8 ML(H	Sti M, K, Pelnakl	
Cont UP Lin Alor opt A to black	Internation Secenty & Communicati	n Lab (1356	Pret A K T.s	6	
Clasher canopholistic List proc. R. B. Scholl	Harpwore Lab.	-155	Prot. R. B. Sin	ch, Sin J. F. H. Nogl	
Int C III International Construction (III) International Construction (I	and and an and a second se		-		. Se
	1 mar 1 m		6.0		

(c) Second scanned image from telephone di- (d) Resulting image from second scanned imrectory age from telephone directory

Figure 2.8: Deskewed document image of scanned telephone directory with Hough transform scheme



(a) Third scanned image from telephone di- (b) Resulting image from third scanned image rectory from telephone directory



(c) Fourth scanned image from telephone di- (d) Resulting image from fourth scanned imrectory age from telephone directory

Figure 2.9: Deskewed images applying the Hough transform scheme

# Chapter 3

# Skew Detection based on Base-Point

In this chapter we are going to discuss the methods of skew detection using a staight line fitting algorithm. The various steps involved and the additional criteria which increases the efficiency of these algorithms have been discussed. Finally, the implementation with suitable examples are given.

### **3.1** Selection of a sub-region and objects

#### 3.1.1 Sub-region selection

Text lines in a document are generally parallel to one another in the horizontal direction, and the space between two successive text lines is relatively constant. Since scanning every pixel in the whole document image is time-consuming, it is appropriate to select a suitable sub-region to calculate the text line direction that corresponds to the skew angle of the image.

Though pixels in one page image express all kinds of information, it is timeconsuming to analyze all the pixels in the image. Connected component, which is the aggregation of related pixels, can also express the information in many aspects such as page layout [5].

In a scanned document image, there might be some black edges that not only affect the algorithm accuracy but also increase the computing cost. In order to avoid the possible negative effects of black edges, the edges of a document image should not be included in the selected sub-region. Furthermore, the size of the sub-region should be carefully selected to achieve higher speed and better accuracy. The selected sub-region R should satisfy the following condition:

$$R = \{(x, y) | w_1 \le x \le w_2, h_1 \le y \le h_2, (w_2 - w_1) \ge W_c, (h_2 - h_1) \ge T_h\}$$
(3.1)

Here  $W_c$  is the average width of the alphanumeric characters, and  $T_h$  is the space threshold between the successive text lines. Given that the width of a document image is W and the height is H, the left boundary of the sub-region should be  $w_1 = W/3$ , the right boundary  $w_2 = W2/3$ , the top boundary  $h_1 = H/3$  and the bottom boundary  $h_2 = H2/3$ . Statistically, the number of connected components in one text line n should be over 10, and the number of text lines in the subregion k should be more than 3, which can ensure the precision of this algorithm.

At the same time the relationships between the adjacent connected components are analyzed with some algorithms such as projection, which can make sure the selected sub-region contains only one text column.

#### 3.1.2 Objects choosing

The bounding box of every connected component is generated firstly. And a single character or touched characters contained in the bounding box are regarded as an object.

Statistically, most of the alphanumeric objects bottoms are on a baseline, such as 'A',' s',' x', etc. Only a very few alphanumeric objects' bounding boxes pass through a baseline, such as 'p',' q', etc. The sizes of the punctuation mark objects are obviously smaller than the alphanumeric ones. In order to remove the negative effects of these punctuation marks, only the objects satisfying the following condition can be selected as candidates for the skew detection algorithm:

$$C = \{C_i | W(C_i) \ge D_w \lor H(C_i) \ge D_h, 1 \le x \le k\}$$
(3.2)

Here C is the set of candidates for the skew detection algorithm,  $W(C_i)$  and  $H(C_i)$ are the width and the height of the bounding box of the object  $C_i$ , respectively,  $D_w$ 

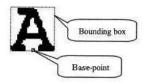


Figure 3.1: Representation of Bounding-box and Base-point for a scanned character is the threshold of the width of the objects' bounding box,  $D_h$  is the threshold of the height of the objects' bounding box, and k is the number of candidate objects.

### 3.2 Base-points clustering

#### 3.2.1 Definitions

**Definition 1.** Base-point of an object is the center at the bottom within the bounding box of an object fig. 3.1.

**Definition 2.** Base-group is a group containing all the base-points in the same text line.

#### 3.2.2 Base-points clustering

In a pure text sub-region where the text lines are parallel, the base-points in different text lines can be grouped into different base-groups according to the space threshold  $T_h$ . The following is the detailed procedure.

Step 1. Initialize every base-point so that it is not in any base-group, and set k = 0.

Step 2. In the selected sub-region R, if the lefttop base-point  $P_i(x_i, y_i)$  not in any basegroup is found, set k + i and put  $P_i$ , into a new base-group G(k).

Step 3. Within the rectangle area  $\{(x_i, y_i - T_h/2)\}, \{(w2, y_i + T_h/2)\}$ , if the leftmost base-point  $P_j(x_j, y_j)$  not in any base-group is found, put  $P_j$  into G(k) and set  $P_i = P_j(i.e.x_i = x_j, y_i = y_j)$ . Repeat this step until all the base-points are in a certain base-group within this rectangle area.

$Angles \Rightarrow$	Calculated Skew $(\theta_s)$	Original Skew $(\theta)$
Scanned Document Images $\Downarrow$	Calculated Skew $(0_s)$	
fig $2.6(a)$	1	2
fig $2.6(c)$	1	2
fig $3.3(a)$	-4	-8
fig $3.3(c)$	-2	6
fig $3.4(a)$	-1	-7
fig $3.4(c)$	-1	15
fig $3.5(a)$	-4	16
fig $3.5(c)$	-3	-18

Table 3.1: Comparative Results in Skew of different scanned document images as calculated and original angles (as measured) using Base-Point Method.

Step 4. Go to Step 2 until all the base-points in the sub-region R have been put into different base-groups.

Apply straight line fitting, using least squares method, for each of the cluster obtained at the end of Step 4 to get the slope of the line that is a best fit for each cluster. Take the average of all slope values obtained in the previous step. This is our skew angle.

### 3.3 Time Complexity

Let the number of pixels in the subregion R be  $N_R$ . Then the base-point algorithm has a time complexity of  $O(N_R)$ .

#### THE SOCIAL EMERGENCY

Set THE SOCIAL RADECENCY laundrysgick, for estample, who, after a  $d_{m_1'}$ work of ten hours at the unachines, go at tight to their boarding-bouses where they was dishes to elso aut a Kring. – then this program, would not be unlike the address of a physica, who tells a poor man with tuberculosis that is must go to the country for a year and live an errorm and enze.

who tells a poor man want two-remains that he must go to the country for a year and live  $a_0$ cream and upper starting grids who adopt loose ways to asticity extravagant desing their tasks are established by women of he would be and mildle classes. The leaves of these women is due to their wage-saming si-tees, who in factories and mills make the obj-propare food stuffs, and do all sorts of task-than formaries (sept women of the upper class at home. Through the instinct of initiation outdoed with the American feeling ideans may, the habits of the well to do Octermine the ambition of many a working sint. Where factors are industrial arrangement which segregate men in construction and im-which are providing for their further employ-ment, turn them loose into cities where any analogous welcome them and each their classe

saloons welcome them and cash their circla

THE SOCIAL EMERGENCY

should be here made of this report, because 2,3

total effect is to minimize economic usage of prostitution, placing the responsibility else, where than on industrial conditions. It is to be noted, however, that it does emphasize the

indirect effects of poverty, and doss speak of the moral danger inrking in certain occupation, and of the had effects of the lack of industrial

and of the had effects of the lack of industria education. More definite responsibility for vice is as effect to low wages in the reports of vice exp-missions. The Chicago Repertasys that of one group of 119 immeral worker, 18 came from department stores, and 35 said that they had taken up the career for the need of means "Cases in which Low Wage and Vice se closely associated."<sup>1</sup> The Report continues:-

In presenting the forspoing to low and statement from girk, this commission does not take the pair it on that the low ways of self-supporting girks is also contributing cause of their d-long means pair ing that there are through the d-long means pair ones: the unnext bandwight above widdless there we tree to those who are ready to schere them. To

\* Report of Forthand Vice Commission, p. 175.

THE SOCIAL EMERGENCY

should be here made of this report, because  $h_{t}$ total effect is to minimize concentic causes of prostitution, ploning the responsibility else, where than on industrial conditions. It is to be noted, however, that it does emphasize the indust effects of poverly, and does apeak of the noral danger lutking in certain occupation, and of the bad effects of the lack of industral education. More definite responsibility for vice in a schedel to low rages in the reports of vice con-missions. The Chicago *Input* says that of ag-group of 119 immed rooms. It is sume from depactment stores, and 38 sold that they had taken up the causer for the need of mong a "Dee Yorldand *Report* properts 22 women as "Deey insolid *Report* properts 22 women as "Deey massected." The *Report* continues-tion of the theory appear in the store of the portion the forewing take and Vice are closely associated. "In the *Report* continuesshould be here made of this report, because its

In presenting the foregoing table and statements from girls, this commission draw not take the pic-tion that the low wages of self-supporting girls if a sole contributing cause of their delinquence, real-ing that there are thousands of girls who would make the uncuts havefolgo hofter yielding them-selves to those who are ready to solice them. The \* Report of Portland Vice Commission, p. 176.

#### (a) First scanned image from book

60

(b) Resulting image from first scanned image

#### THE SOCIAL EMERGENCY

a THE BOELL EMERGENCY should be here made of this report, because in total effect is to minimize economic cause, of prositution, placing the responsibility else, where than on industrial conditions. It is to be noted, however, that it does emphasize the indirect effects of poverty, and does speak of the moral danger lurking in certain occupation, and of the had effects of the lack of industrial character. education.

education. More definite responsibility for vice is a minimum variant in the reports of vice equ-minimum. The Chicago Report mays that of me group of 110 immoral women, 16 came from department stores, and 38 shall that they ini-taken up the career for the need of monsy. The Portland Report presents  $\pm 20$  women a "Cases in which Low Wage and Vice are closely associated."<sup>11</sup> The Report continues.—

In presenting the foregoing table and statement from girls, this examination does not take the pai-tion that the low wages of self supporting girls in the also contributing cause of their delanguary, self-ing that there are thousands of girls sub-would en-dance the utness harddings before yielding their solves to those who are ready to seduce them. The \* Report of Portland Vice Commission, p. 178

(c) Second scanned image from book (d) Resulting from second image scanned image

Figure 3.2: Deskewed document image of scanned book page with Base-point scheme



(a) First scanned exam paper image



(c) Second scanned exam paper image

(四)15 NATIONAL

(b) Resulting image from first scanned exam paper image



(d) Resulting image from second scanned exam paper image

Figure 3.3: Deskewed document image of scanned exam paper with Base-point scheme

Camputer Spience & Eng Office	littere	4208 (Prot. 9, 8	ira, Sili G. Deteraj Vatlachuryva, Sol P. K. M	in a Setul	HOD Prot. B. M.	230 (h) 235		SKK Nayokj	
Cinde			Solp K M	Planty	Faculty	239	1	P00-da	
HOD PHY. R. Majo		350 (Sei G. Nitur	R. SH K. Nayanj		Batenings R				
11.80	0 <del></del>				Chinary 9.	2359	3359		
Faculty	23	53	hod-ca		Jana S	2501	3367	rbsingh	0-21
Battersingh R.					Jona 8 sc	2553	3953	<b>NUMBER</b>	0-64
Chinara S.	235				RDdar #M	2352	2352	njona	17-13
Anta E	236	3301	rbsingh	8-21	Maps w.	2154	1354	siljena	A-8
Hena S.K.	2363	3353	Futhenuta	D+64	Mohapara D.P	2355	3258	prokhilge.	D-17
Khlar P.M.	2362	3352	6jena	TF-13	Rath E.K.	7356	3756	Droughy.	D-8
Magra B.	2354	3354	sijena		Ta Parkaj Kumar	2357	3387	durps	0.29
Mohapatha C.P.	2355	3355	DTAbilar	A-8	Sahu B.D.	2302	3347	Photo	A-15
Rath S.K.	2356	3356	breahi	D -17	TUNKAK	2358		DANKARA	TF-14
Sa Parikaj Kumar	2357	2187	durge	D-6	Laboraturies & Facilities	2363	3360	bdaahu	Store.
Sahu B.D.	2362	1362	Nami	0-29	Software / Acceletion		0000	akturak	0.22
TURNAK	2358	4.852	pankajkoa	A-15	Software Laboratory (Mr. P., et al. Advanced Database Ergp 1.ap	w.78.1 4351			ALCON.
Laboratories & Facilities	2360	3360	bdtahu	77-14	Image Personal Engl Lab	4352	PHD. B. MAPI, SHA	K. Das ; Sn M. K. Pate	initi i
		1080	Office 4		Image Processing & Cluster computing Internation Security & C	Lab #351	Prol. 9. K. Jenaj		any .
Software Laboratory ( M3-K - Pirulet Hund Idvanced Detabase Engl. Lab	2 4351			0-22	Harmation Security & Communication Hardware Lab		Prof. B. Math. Sri M. J	K Patrak(	
NOR Processing 6	4380	PTOL B. MAPI, SHA.	K. Dat ; Sti M. K. Patral		CID.				
tormation formed a cluster computing Lat	4313	Prot. S. K. Jerul	, an M. K. Palnal	k)		1000	Pot R & Shigh, Srij	P.H. Nati	
formation Security & Construction Lab retwary Lab.		Prof. B. Math, SH.M.	C Patricia			0			
and the		Prot. R. B. Singh, Sri J							

(a) First scanned image from telephone direc- (b) Resulting image from first scanned image tory from telephone directory

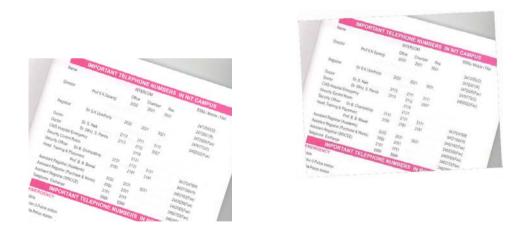


(c) Second scanned image from telephone di- (d) Resulting image from second scanned imrectory age from telephone directory

Figure 3.4: Deskewed document image of scanned telephone directory with Base-point scheme



(a) Third scanned image from telephone di- (b) Resulting image from third scanned image rectory from telephone directory



(c) Fourth scanned image from telephone di- (d) Resulting image from fourth scanned imrectory age from telephone directory

Figure 3.5: Deskewed images applying the Base-point scheme

# Chapter 4 Conclusions

The work in this thesis, primarily focuses on skew detection in scanned document images. The work reported in this thesis is summarized in this chapter. Sec. 4.1 lists the pros and cons of the work. 4.2 provides some scope for further development.

### 4.1 Achievements and Limitations of the work

The scan line algorithm, based on calculating the coordinates of lines in the image plane and finding the number of non-background image pixels on a line and then finding the skew angle on the basis of variance is a successful method, in terms of skew detection and was found to reliably detect a range of skew angles over a range of text document images. However, it has a drawback of being painstakingly slow as compared to the other methods.

The hough transform method was found to be as good as the scan line algorithm in terms of skew detection and has a much lesser time complexity.

The above two methods were found to provide the correct skew angle even when the document image consists of many lines that are parallel to the text, since these lines increase the number of votes because the number of votes along the direction of these lines is much more as compared to any other direction. These methods are particularly useful for detecting skews in documents that contain for scripts which have lines as a major part of the script. For example, many Indian scripts such as Devanagri, Bangla, Gurmukhi, etc. have lines, called headlines, connecting letters of a word. The base point method has a much lesser time complexity than the other two, since it operates on eigen points instead of working on all non-background image pixels. Since this method relies on connected components to find the eigen points, if the connected components so found are not distinct characters, the base point method can result in false results. This was evident in the test images.

### 4.2 Further Development

The algorithms we have discussed works well with tilted skews and not with shear skews in scanned document images. We can try to implement and/or modify the existing algorithm so that it can detect also detect shear skews in scanned document images.

## Bibliography

- R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison Wesley, 2nd edition, 1992.
- [2] Anil K. Jain. Fundamentals of Digital Image Processing. Prentice Hall of India, 2008.
- Bidyut B. Chaudhuri. Reading Systems: An Introduction to Digital Document Processing. Springer London, 2007.
- [4] Katherine Marsden. Skew detection and correction overview. http://www.eecs.berkeley.edu/ fateman/kathey/skew.html.
- [5] Chandan Singha, Nitin Bhatiab, and Amandeep Kaur. Hough transform based fast skew detection and accurate skew correction methods. *Pattern Recognition*, 41:3528 – 3546, 2008.
- [6] Yang Cao, Shuhua Wang, and Heng Li. Skew detection and correction in document images based on straight-line fitting. *Pattern Recognition Letters*, 24:1871 – 1879, 2003.
- [7] Pankaj Kumar Sa. On the Development of Impulsive Noise Removal Schemes. Master's thesis, National Institute of Technology Rourkela, 2006.