

Optical Character Recognition and analysis of historic data

Mahamat Annour Abdallah (mahamat@aims.ac.za)

محمد النور عبدالله

African Institute for Mathematical Sciences (AIMS)

Supervised by: Prof B.M. Herbst
Stellenbosch University , South Africa

18 May 2017

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa



Abstract

Optical Character Recognition has been used for many years to convert text documents which have been written by hand or typed into digital text. In this essay we present some of methodology of how OCR systems works by taking a look at the mathematical concepts associated with it, including Radon Transform and its properties.

For this project, some famous open source OCR systems will be investigated and evaluated for accuracy then used for analysis of some samples of historic documents dating back to the 19th century.

الذّياجة

لقد استخدامة تقنية التعرف الضوئي على الرموز لسنوات عديدة لتحويل الوثائق النصية التي تمت كتابتها بخط اليد أو طبعة عبر الآلة إلى نص رقمي.

في هذا المقال سوف نقدم منهجية عامة لكيفية عمل هذا النظام وذلك عبر تسليط الضوء على بعض المفاهيم الرياضية المرتبطة به، مع الأخذ في الاعتبار تحويل رادون وخصائصه.

في هذا المشروع، سيتم استخدام بعض أشهر أنظمة التعرف الضوئي على الرموز مفتوحة المصدر وتقييم دقة أدائها عبر اختبار معالجة بعض العينات من الوثائق التاريخية التي تعود إلى القرن التاسع عشر الميلادي.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.



Mahamat Annour Abdallah, 18 May 2017

Contents

Abstract	i
1 Introduction	1
1.1 The OCR Technology	1
1.2 Overview of history of OCR	1
1.3 Statement of the problem	1
1.4 Overview and material presented in this project	2
2 The Radon Transform in OCR system	3
2.1 Radon transform	3
2.2 Properties of Radon Transform	5
3 The design space of the OCR system	9
3.1 Pre-processing	9
3.2 Segmentation	12
3.3 Recognition	14
3.4 Post processing	15
4 Data analysis and Results	16
4.1 OCR tools open source	16
4.2 Analysing data	16
4.3 Results	18
5 Conclusion and Future work	20
5.1 Conclusion	20
5.2 Future work	20
References	23

1. Introduction

1.1 The OCR Technology

Optical character recognition (OCR) is a program that used to convert a scanned or printed image document into a text document, then can be stored in any internationally known format that is used to represent text in electronic devices.

Suppose we have some documents in the form of non-digital files, whether written by hand or typed and want to make them in digital form, in order to be able electronically analyse and handle them, one could spend hours rewriting and debugging all those documents, or simply convert all those documents to digital format in minutes using a scanner and OCR software. Therefore, the optical character recognition (OCR) automatically translates images (which are usually captured by the scanner), to an editable text or Character to an international standard encoding scheme (e.g. ASCII or Unicode). In this essay we investigate the potential look forward this wonderful technique to re-digitize old manuscripts so that they can be editable and researchable. (Eikvil, 1993)

1.2 Overview of history of OCR

The first OCR ideas were designed between 1870 – 1931 when **Fournier d'Albe** developed the **Optophone** and **Tauschek's reading machines** as devices to aid blind people for reading. Between 1931 – 1954 the first OCR instruments were invented and applied in industry, and being able to interpret **Morse code** and read text, **The Smart Machinery Research Company** was the first company established to specialize in producing this kind of technology at that time. between 1954 – 1974 **Optacon** was developed and was the first portable OCR device used to digitize **Reader's Digest** coupons and postal addresses. This machine has designed to facilitate scanning. In the ear between 1974 – 2000 the scanners were widely used to read price tags and passports, and companies such as **Corporation**, **ABBYY** and **Kurzweil Computer Products** were established. The latter developed the first-line OCR software, able to read any text document with high accuracy. Currently OCR software is available on the Internet for free, on a very wide scale, through products such as Adobe Acrobat, WebOCR and Google Drive, But In order to get a more high accuracy with additional services using professional software, we still have to pay for it. There are also some high-performance open sources software such as **Tesseract** and **CuneiForm** which we will address in more detail during the course of this project. (Wikipedia)

1.3 Statement of the problem

Some South African economic historians are very interested in records that date back to the 19th century. These records are only available in printed form and before any analysis can start, the data has to be converted to text. For this project, some open source OCR systems will be investigated and evaluated for accuracy. In order to analyse the data the open source Tesseract and Cuneiform will take into consideration as the most accurate systems for optical character recognition open source.

1.4 Overview and material presented in this project

- Chapter two will focus on discussing the mathematical concept behind OCR by considering Radon Transform and its properties.
- Chapter three presents the design space of the OCR system, where we represent the structure and the different necessary steps that make this system successful and gives more accurate results.
- In chapter four will take some samples from our data and will be analyzed them, by using open source software and compare the results to figure out which is most efficient and have more accurate.
- Chapter five summarize the conclusion of our work and finally we will describe the future works.

2. The Radon Transform in OCR system

2.1 Radon transform

Mathematically we can compute the projection histograms of an image using the Radon transform (the projection histograms is counting the number of pixels in each column and row of a character image), this transform supposed by Austrian mathematician Johann Radon (1887 – 1956). states that image reconstruction from projection is possible.

The main motivation to consider Radon transform in this project, that it's suited for line parameter extraction even in the case of presence of noise. The Radon transform more efficient to converts a difficult global detection problem in the domain of the photos into a more easily local peak detection problem in corresponding line parameters in the domain (Toft and Sørensen, 1996) and the Random transform is able to transform two dimensional images with lines into a domain of possible line parameters, this have lead to many line detection applications within image processing, computer vision, and seismic (Miciak, 2008).

As mentioned above, the Radon Transform allows the reconstruction of the two-dimensional images from the line parameter projections. The Radon Transform computes these projections using the technique of rotational scanning. A set of line integrals of the image's pixel values (using, for example, grey-scale weights) are computed along parallel paths, spaced one pixel apart, for each perspective (angle) chosen for scanning. The image is represented by an number of these sets of line integrals (projections), each from a different angle ¹ Figure 2.1 illustrates the process from one such angle. The Radon Transform offers a general method for computing the projection along any angle θ .

2.1.1 Definition. For a given function f , whose domain is the plane, then for each pair (x', θ) the Radon Transform of f is defined, by

$$\mathfrak{R}(x', \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \delta(x \cos \theta + y \sin \theta - x') dx dy, \quad (2.1.1)$$

where $\mathfrak{R}(x', \theta)$ is a projection of $f(x, y)$ on the axis x' of θ direction and $\delta(\cdot)$ is the delta function which is was introduced by Dirac (Feeman, 2010). So that value of a 2-D function at an arbitrary function at an arbitrary point is uniquely obtained by the integrals along the lines of all direction passing the point (Asl and Sadremomtaz, 2013)

The delta function is special in the sense that is defined as follows:

$$\delta(\tau) = 0 \text{ for } x \neq 0, \quad \int_{-\infty}^{\infty} \delta(\tau) dx = 1. \quad (2.1.2)$$

And x' is defined as follows:

$$x' = x \cos \theta + y \sin \theta, \quad (2.1.3)$$

where x' is the perpendicular distance of the beam form the origin and θ is the angle of incidence of the beams (Toft and Sørensen, 1996).

¹The perspectives are rotated around the centre of the image.

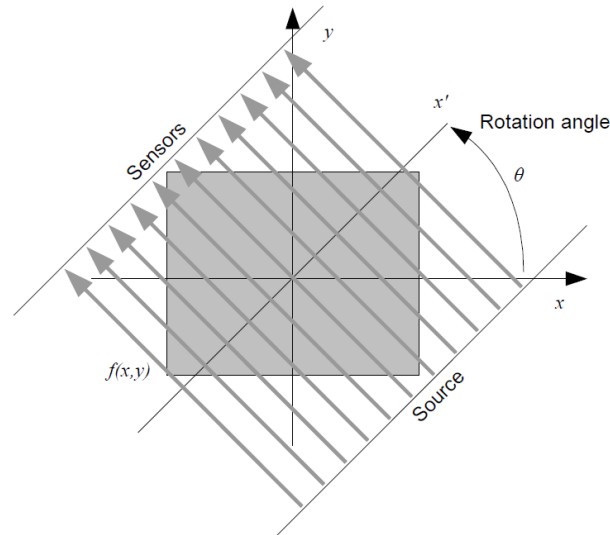


Figure 2.1: Single projection at a specified rotation angle

The Radon Transform allows us to compute the projection of image intensity along the axis x' lying at angle θ (counterclockwise) from the x -axis. The origin is located at the center of the image, which all axes pass through. The projection of the image into the x -axis, for example, given by the Radon Transform with $\theta = 0$ (and similarly for the y axis when $\theta = \frac{\pi}{2}$). The projection at some θ of a simple shape is illustrated in figure 2.2a.

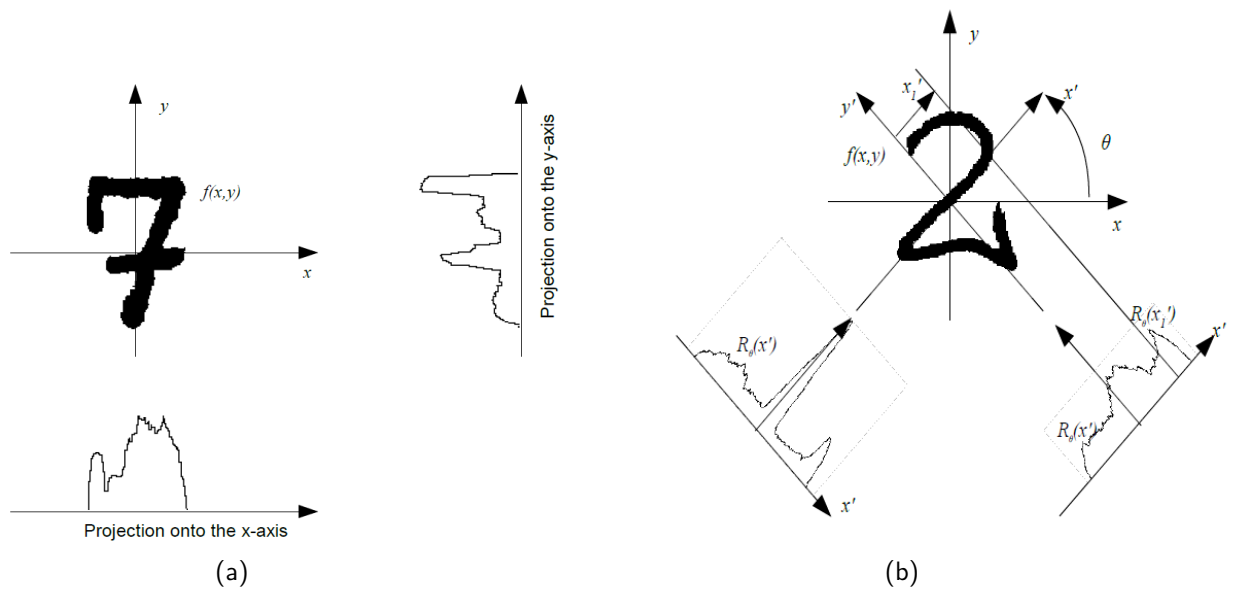


Figure 2.2: (a) show the horizontal and vertical projection of a simple function by using Radon Transform, (b) The geometry of Radon Transform.

The figure 2.2b show us the geometry of the Radon transform. Although the Radon transformation expresses the projection by the 2D integral on the (x, y) -coordinate, the projection is more naturally expresses by an integral of one variable since it is a line integral (Miciak, 2008). Since the (x', y') -coordinate along the direction f projection is obtained by rotating the (x, y) -coordinate by θ , we can

represent the relationship between two direction is expressed as following

$$x' = x \cos \theta + y \sin \theta, \quad y' = -x \sin \theta + y \cos \theta \quad (2.1.4)$$

From the equation 2.1.1 we can define the discrete equation of the radon transform as follows

$$\Re(x', \theta) = \sum_{x=1}^M \sum_{y=1}^N f(x, y) \delta(x \cos \theta + y \sin \theta - x') dx dy \quad (2.1.5)$$

where M, N are the dimensions of rows and columns for the given image matrix, $\Re(x', \theta)$ as we mention above the integral of $f(x, y)$ along the line $x' = x \cos \theta + y \sin \theta$.

In order to make the idea more clearly let suppose x'_1, x'_2, \dots, x'_n be the perpendicular distance of a line from the origin, and the projection angle $\theta_1, \theta_2, \dots, \theta_m$ are the angles formed corresponds to x'_1, x'_2, \dots, x'_n by the distance vector.

The magnitude $\Re(x'_i, \theta_k)$ represent to us the amount of the Radon Transform of point (x'_i, θ_k) , so that Radon Transform matrix (Gan and He, 2011) is define as follows:

$$\Re(x'_i, \theta_k) = \begin{bmatrix} \Re(x'_1, \theta_1) & \Re(x'_1, \theta_2) & \dots & \Re(x'_1, \theta_n) \\ \Re(x'_2, \theta_1) & \Re(x'_2, \theta_2) & \dots & \Re(x'_2, \theta_n) \\ \vdots & \vdots & \ddots & \vdots \\ \Re(x'_m, \theta_1) & \Re(x'_m, \theta_2) & \dots & \Re(x'_m, \theta_n) \end{bmatrix} \quad (2.1.6)$$

2.2 Properties of Radon Transform

The strongly property of this transform is its ability to extract lines (curves in general) from the high image noise, thus following some of interesting properties relating to the application of affine transformation, and we can use Radon transform to compute any translated, rotated or scaled image and the properties of Radon transform as follows:

Note we are going to consider, The parameter set of $t \in \{0, \infty\}$ and $\theta \in \{0, 2\pi\}$ describes every element of the Radon Transform, since $\Re f(t, \theta) = \Re f(-t, \theta + \pi)$ (Toft and Sørensen, 1996), the following are the important properties of this transformation.

- **The linearity**

The Radon Transform is a linear transform, suppose if α_i is an array of constants, and f_i is functions then

$$\Re \left(\sum_i \alpha_i f_i \right) = \sum_i \alpha_i \Re f_i \quad (2.2.1)$$

In other order if we have for example two given function f_1 and f_2 are both define in the plan and two any constants α and β then

$$\Re(\alpha f_1 + \beta f_2) = \alpha \Re f_1 + \beta \Re f_2, \quad (2.2.2)$$

From Fig.2.1 each projection from the point source (t_i, θ_k) represent Radon Transform along the plan (x, y) , so that the whole image will be represented by summation of all point source. therefore the linearity is vary important in Radon Transform.

- **The shifting**

This property of Radon allows us easily correct such images by shifts elements of an array along any dimension by any number of elements to another dimension, let assume that a function $g(x, y)$ is shifted so that

$$h(x, y) = g(x - x_0, y - y_0) \Rightarrow \quad (2.2.3)$$

$$\Re h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x - x_0, y - y_0) \delta(t - x \cos \theta - y \sin \theta) dx dy \quad (2.2.4)$$

Let us introduce new variables: $x' = x - x_0$, $y' = y - y_0$ Then

$$\Re h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \delta(t - x_0 \cos \theta - y_0 \sin \theta - x' \cos \theta - y' \sin \theta) dx' dy' \quad (2.2.5)$$

$$= \Re g(t - x_0 \cos \theta - y_0 \sin \theta - \tau) \quad (2.2.6)$$

Thus the shift won't affect the variable $\tau = x' \cos \theta + y' \sin \theta$, it only affects is on t -coordinate by the transform of θ .

- **The rotation**

The advantage of rotation property of Radon Transform is that the Radon Transform of the rotation of $f(x, y)$ by angle ϕ leads to a circular shift of the Radon Transform of original $f(x, y)$ in the variable θ , that mean $\Re f(t, \theta) \Rightarrow \Re g(t, \phi_0 - \theta)$, so that we will consider the polar coordinate (r, ϕ) , then Radon transform in 2-D reads:

$$\begin{aligned} \Re f(t, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(r, \phi) \delta(t - r \cos \phi \cos \theta - r \sin \phi \sin \theta) |r| dr d\phi \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(r, \phi) \delta(t - r \cos(\phi - \theta)) |r| dr d\phi \end{aligned}$$

Let assume that the angle of rotation is ϕ_0 , therefore the function change as $h(r, \phi) = g(r, \phi - \phi_0)$ Thus

$$\Re h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(r, \phi - \phi_0) \delta(t - r \cos(\phi - \theta)) |r| dr d\phi$$

Now let introduce a new notations: $\phi' = \phi - \phi_0$ Thus

$$\Re h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(r, \phi') \delta(t - r \cos(\phi' + \phi_0 - \theta)) |r| dr d\phi' = \Re g(t, \phi_0 - \theta) \quad (2.2.7)$$

we can observe that, if the coordinate system (x, y) is turned by amount ϕ_0 then also the Radon Transform will returned by ϕ degree. Thus the rotation property plays an important role in rotating the image matrix (Hu, 1962).

- **Scaling**

The scaling along both axes (x, y) in the spatial domain results in the scaling along the t axis in the Radon domain and scaling of the value of the transform (Arodź, 2004), so that let assume $0 < a, 0 < b$ constants, and scaling coordinate of variables as follows $h(x, y) = g(\frac{x}{a}, \frac{y}{b})$, then the Radon Transform of $h(x, y)$ is

$$\Re h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g\left(\frac{x}{a}, \frac{y}{b}\right) \delta(t - x \cos \theta - y \sin \theta) dx dy$$

Again let consider new notations $x' = \frac{x}{a}, y = \frac{y}{b}$ and γ variable, then we have

$$\begin{aligned}\mathfrak{R}h(t, \theta) &= ab \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \delta(t - ax' \cos \theta - by' \sin \theta) dx' dy' \\ &= \frac{ab}{|\gamma|} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \delta\left(\frac{t}{\gamma} - x' \frac{a \cos \theta}{\gamma} - y' \frac{b \sin \theta}{\gamma}\right) dx' dy'\end{aligned}$$

let assume that γ can be chosen and θ will be replaced by variable θ' as follows

$$\cos \theta' = \frac{a \cos \theta}{\gamma}, \quad \sin \theta' = \frac{b \sin \theta}{\gamma} \quad \text{and} \quad t' = \frac{t}{\gamma} \quad (2.2.8)$$

Thus

$$\mathfrak{R}h(t, \theta) = \frac{ab}{|\gamma|} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \delta(t' - x' \cos \theta' - y' \sin \theta') dx' dy' = \frac{ab}{|\gamma|} \mathfrak{R}g(t', \theta') \quad (2.2.9)$$

So that t', θ' must be founded as function of t, θ , and we can express γ from 2.2.8 so we get

$$\frac{a \cos \theta}{\cos \theta'} = \gamma = \frac{b \sin \theta}{\sin \theta'} \Rightarrow a \tan \theta' = b \tan \theta' \quad \text{Thus} \quad \theta' = \arctan\left(\frac{b}{a} \tan \theta\right)$$

then we get

$$\cos^2 \theta' + \sin^2 \theta' = 1 \Rightarrow \gamma^2 = a^2 \cos^2 \theta + b^2 \sin^2 \theta \Rightarrow \gamma = \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta}$$

Therefore

$$\mathfrak{R}h(t, \theta) = \frac{ab}{\sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x', y') \delta(t' - x' \cos \theta' - y' \sin \theta') dx' dy' = \mathfrak{R}g(t', \theta')$$

Thus

$$t' = \frac{t}{\sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta}} \quad \text{and} \quad \theta' = \arctan\left(\frac{b}{a} \tan \theta\right) \quad (2.2.10)$$

To summarize the equation 2.2.9 we have

$$\mathfrak{R}h(t, \theta) = \frac{ab}{|\gamma|} \mathfrak{R}g(t', \theta') \quad (2.2.11)$$

$$= \frac{ab}{|\gamma|} \mathfrak{R}g\left(\frac{t}{\gamma}, \arctan\left(\frac{b}{a} \tan \theta\right)\right), \quad \gamma = \sqrt{a^2 \cos^2 \theta + b^2 \sin^2 \theta} \quad (2.2.12)$$

By we assume that $0 \leq \theta < \pi$, then the arctan function is

$$\arctan\left(\frac{b}{a} \tan \theta\right) = \begin{cases} \arctan\left(\frac{b}{a} \tan \theta\right) & \text{if } 0 \leq \theta < \frac{\pi}{2} \\ \arctan\left(\frac{b}{a} \tan \theta\right) + \pi & \text{if } \frac{\pi}{2} \leq \theta < \pi \end{cases} \quad (2.2.13)$$

• The convolution

The property of convolution for Radon in the two dimensions (x, y) has an important consequence on the reconstruction of an image. Let us assume that the function $h(x, y)$ 2D convolution of the functions $f(x, y)$ and $g(x, y)$ therefore.

$$h(x, y) = f * g(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) g(x - x_1, y - y_1) dx_1 dy_1 \quad (2.2.14)$$

The Radon transform of convolution $h(x, y)$ is given by

$$\mathfrak{R}h(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) g(x - x_1, y - y_1) \delta(t - x \cos \theta - y \sin \theta) dx_1 dy_1 dx dy$$

The integral in around $dx dy$ is the Radon transform of the $f(x, y)$ function shifted to the points (x_1, y_1) , by using the shift theorem, so that

$$\begin{aligned} \mathfrak{R}h(t, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x - x_1, y - y_1) \delta(t - x \cos \theta - y \sin \theta) dx_1 dy_1 dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) \mathfrak{R}g(t - x_1 \cos \theta - y_1 \sin \theta, \theta) dx_1 dy_1 \end{aligned}$$

Now let us insert a new integration of t_1 with a Dirac delta function, and carry out the integration according to x_1, y_1 then we get

$$\begin{aligned} \mathfrak{R}h(t, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) \int_{-\infty}^{\infty} \mathfrak{R}g(t - t_1, \theta) \delta(t_1 - x_1 \cos \theta - y_1 \sin \theta) dx_1 dy_1 dt_1 \\ &= \int_{-\infty}^{\infty} \mathfrak{R}g(t - t_1, \theta) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, y_1) \delta(t_1 - x_1 \cos \theta - y_1 \sin \theta) dx_1 dy_1 dt_1 \\ &= \int_{-\infty}^{\infty} \mathfrak{R}f(t_1, \theta) \mathfrak{R}g(t - t_1, \theta) dt_1 \\ &= \mathfrak{R}f(t, \theta) * \mathfrak{R}g(t, \theta) \end{aligned}$$

Thus the Radon Transform of $2D$ is a $1D$ convolution of the Radon Transformed function with respect t , ([Toft and Sørensen, 1996](#)).

3. The design space of the OCR system

The optical character recognition system requires many steps to completely recognize characters and produce machine encoded text. In this section, we focus on the technique of OCR systems. After the image has been obtained, usually by using scanner, The literature review in the field of OCR requires various method of processing or phases which applied to the image to perform the many different vision tasks required, those steps represented in the following stages:

Pre-processing, Segmentation, Radon transform, and Recognition. The block diagram showing in Figure 3.1 represent to us the optical character recognition system (Chandarana and Kapadia, 2013a) (Miciak, 2008)

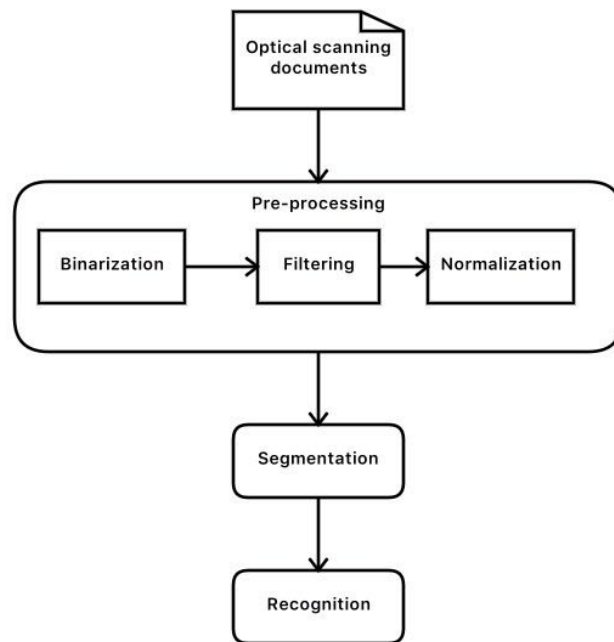


Figure 3.1: Diagram show the Optical Character Recognition system.

3.1 Pre-processing

Pre-processing plays an important role in an OCR system. The aims of this stage to produce data that are easy for the Character Recognition systems to operate accurately. whereas pre-processing perform reducing the noise and normalize the data, in order to achieve those objectives, the following techniques are used in the pre-processing stage.(Chandarana and Kapadia, 2013b) (?)

3.1.1 Binarization. The binarization (thresholding) of image is the starting step of most image processing system, refers to the conversion of a gray-scale image into a binary image. The colourful image from the acquisition unit (scanner) represented by 3 coefficients Red, Green and Blue. (Miciak, 2008) (Vamvakas et al., 2008) the image must be converted to the image with 256 levels of grey scale (0 to 255 pixel values) In order be represented as a binary image (0 to 1 pixel values) by using a threshold value. The pixels lighter than the threshold are turned to white and the remainder to black pixels.(Singla

and Yadav, 2014) as showing in Figure 3.2, the threshold image function $f(x, y)$ is defined as follows:

$$f(x, y) = \begin{cases} 1 & \text{if } g(x, y) > T \\ 0 & \text{if } g(x, y) \leq T, \end{cases} \quad (3.1.1)$$

where $g(x, y)$ is the gray level of point (x, y) and T is a threshold that separate the represented model (Jain et al., 1995). The Figure 3.2a represent to us sample from the historical data that we working on

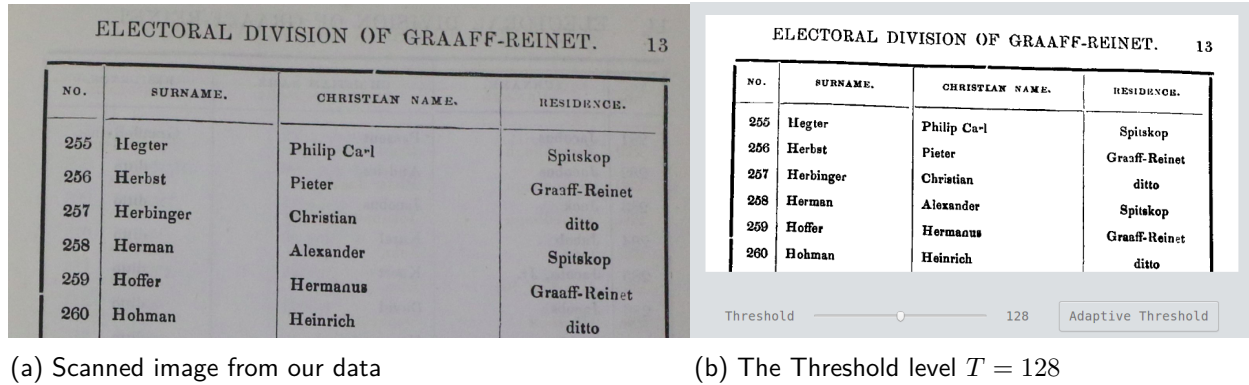


Figure 3.2: Diagram show the process of Threshold

as scanned image, and Figure 3.2b its converted result to binary image by considering Threshold level $T = 128$, thus all the pixels lighter than this amount are turned to whit (background of image) and remind to black (the other information).

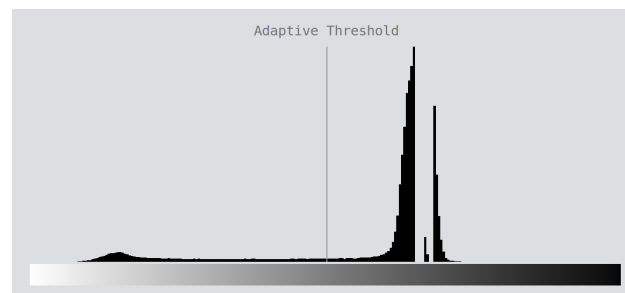


Figure 3.3: Diagram show level of grey scale and Threshold

The Figure 3.3 show us the level of grey scale from white 0 to Black 1, the line represent to us the adaptive Threshold between white and black pixel. The figure 3.4 represent example of a binarization method for a letter A (Dhaka et al.).

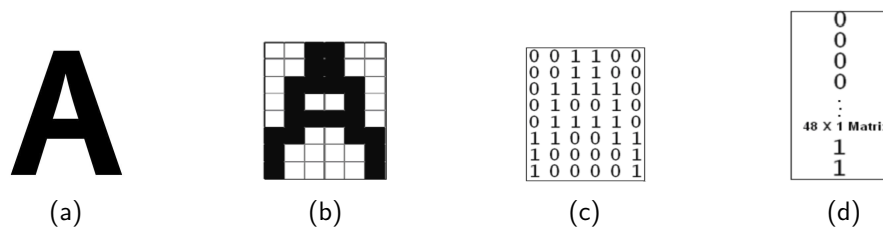


Figure 3.4: Digitization Process

3.1.2 Filtering. The filtering is used for improving the quality of the image, by remove noise and diminish spurious points emphasizing details and make the processing of the image more easily for the next steps, filtration of digital images is obtained by convolution operation. That by making the new value of point of image that represented by grey scale counted on the basis of neighbouring points value (Miciak, 2008) .

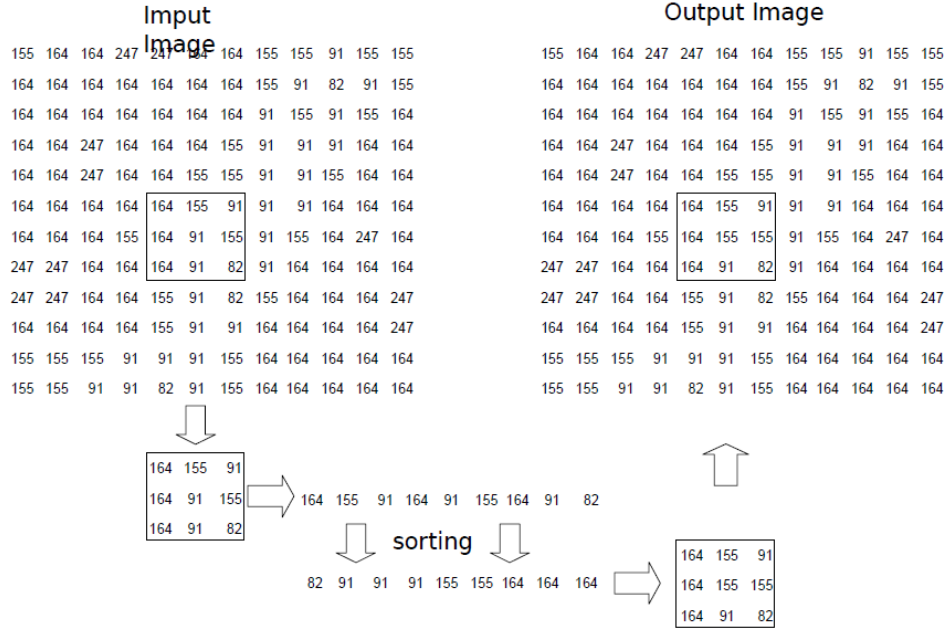


Figure 3.5: Diagram show th Median filtering.

The basic idea of that is to convolute a predefined mask with the image to assign a value to a pixel as a function of the gray values of its neighbouring pixels. Every value is classified and it has influence on new value of point of the image after filtration (Vamvakas et al., 2008) The applied filter is median filter, with mask 3×3 .

3.1.3 Normalization. The normalization is a process that changes the range of pixel intensity values(Chandarana and Kapadia, 2013b). The character normalization is applied for standardization size of the character. Since the scanned image might have different distortion such as: translation, rotation and scaling. Images there are translated, rotated and expanded or decreased. The typical solutions take into consideration the normalization coefficients and calculate the new coordinates given by:

$$\underbrace{[x, y, 1]}_{\text{Normalize character}} = [i, j, 1] \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -I & -J & 1 \end{bmatrix}}_{\text{Translation}} \times \underbrace{\begin{bmatrix} m_i & 0 & 0 \\ 0 & m_j & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Scaling}} \times \underbrace{\begin{bmatrix} \cos \beta & \sin \beta & 0 \\ -\sin \beta & \cos \beta & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{Rotation}} \quad (3.1.2)$$

Where I, J are the center of mas of the image and given by:

$$I = \frac{\sum_i \sum_j i f(i, j)}{\sum_i \sum_j f(i, j)} \quad \text{and} \quad J = \frac{\sum_i \sum_j j f(i, j)}{\sum_i \sum_j f(i, j)} \quad \text{where} \quad f(i, j) = \begin{cases} 0 & \text{white} \\ 1 & \text{black} \end{cases} \quad (3.1.3)$$

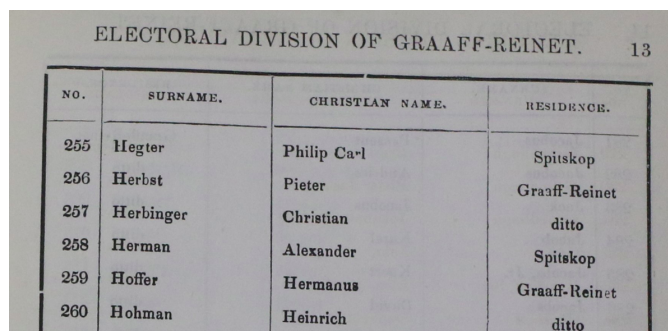
By normalization, the character is made to fit into a standard size array. Any size of characters and shape can be processed and matched with the normalization technique. so we consider the new coordinate

system where center is equals to center of mass of the character. The value of angle rotation is according to main axes of the image. The value of scale coefficient is calculated by mean value of variation of the character. So the center of mass of the character is good candidate point of the center of image as a product of normalization stage (Miciak, 2008).

3.2 Segmentation

The pre-processing stage yields a clean document in the sense that a sufficient amount of shape information, high compression and low noise on a normalized image is obtained. In this stage, the document will be segmented into its sub-components (Arica and Yarman-Vural, 2001). The top-down and left-right segmentation approach will be applied. where the text will be segmented into lines and words and finally each word can be segmented into individual characters, using the vertical and horizontal projection of Radon transform and then be classified or identified (Singla and Yadav, 2014). The aim of the segmentation to found out the position of each individual characters and composite characters in the image in order to be recognized in next stage (Vamvakas et al., 2008).

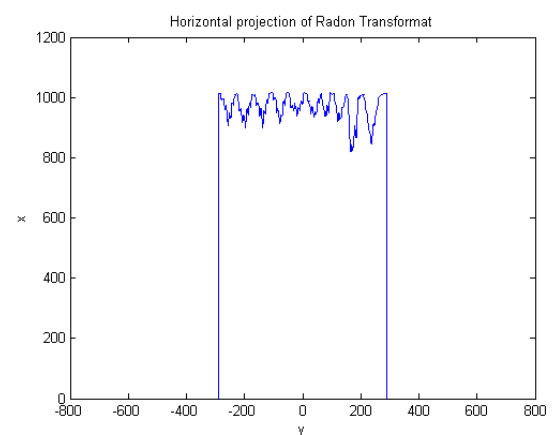
3.2.1 Line segmentation. Line segmentation is the first step of the segmentation process. This technique considers a pixel-based horizontal projection by using Radon Transform which is a connected component-based projection to separate text lines, the white space between the text lines is used to segment the text lines. when the gap between lines should show zero value of summation of pixels. then the system considers this zero value as the gap between the lines. so that the lines have been segmented and stored to be used for words segmentation (Khidhir, 2011) (Miciak, 2008). Figure 3.6 show the line segmentation using Radon Transform.



The input image shows a document titled "ELECTORAL DIVISION OF GRAAFF-REINET. 13". It contains a table with four columns: NO., SURNAME, CHRISTIAN NAME, and RESIDENCE. The table lists six entries with their respective details.

NO.	SURNAME.	CHRISTIAN NAME.	RESIDENCE.
255	Hegter	Philip Carl	Spitskop
256	Herbst	Pieter	Graaff-Reinet
257	Herbinger	Christian	ditto
258	Herman	Alexander	Spitskop
259	Hoffer	Hermanus	Graaff-Reinet
260	Hohman	Heinrich	ditto

(a) Input image



(b) Radon Transform of Image in Fig 3.6a

Figure 3.6: Diagram show line segmentation using Radon Transform

3.2.2 Word Segmentation. The segmentation of words be determined by making a vertical projection in an image, and as we know in generally in English script, the spacing between the words is greater than the spacing between the characters in a word as showing in a figure 3.7. So that the width of the zero-valued valleys are more between the words in the line as compared to the width of zero-valued valleys that exists between characters in a word (Shinde and Chougule, 2012) therefore the technique is to find the spacing between the words by taking the vertical projection of Radon Transform on the

image by follows the same concept that mentioned in the previous step(Chandarana and Kapadia, 2014). Figure 3.7 show us words segmentation by using Radon Transform.

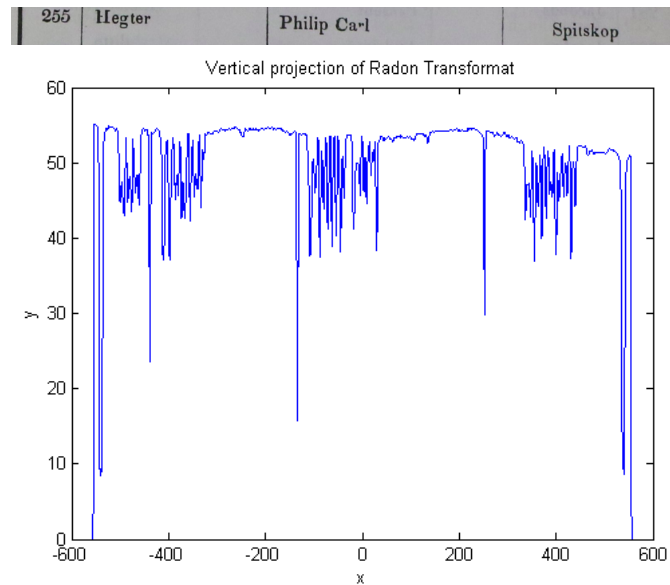


Figure 3.7: Diagram show words segmentation using Radon Transform

3.2.3 Character Segmentation. The character segmentation is the last step of segmentation stage, the method is the same as word segmentation, but the only difference is the summation of zero pixels between the different characters are less than summation of zero pixels between the words. Here finally the character has been isolated to be recognized in next steps(Singla and Yadav, 2014)(Shinde and Chougule, 2012). The figure 3.8 show the Word and Character Segmentation.

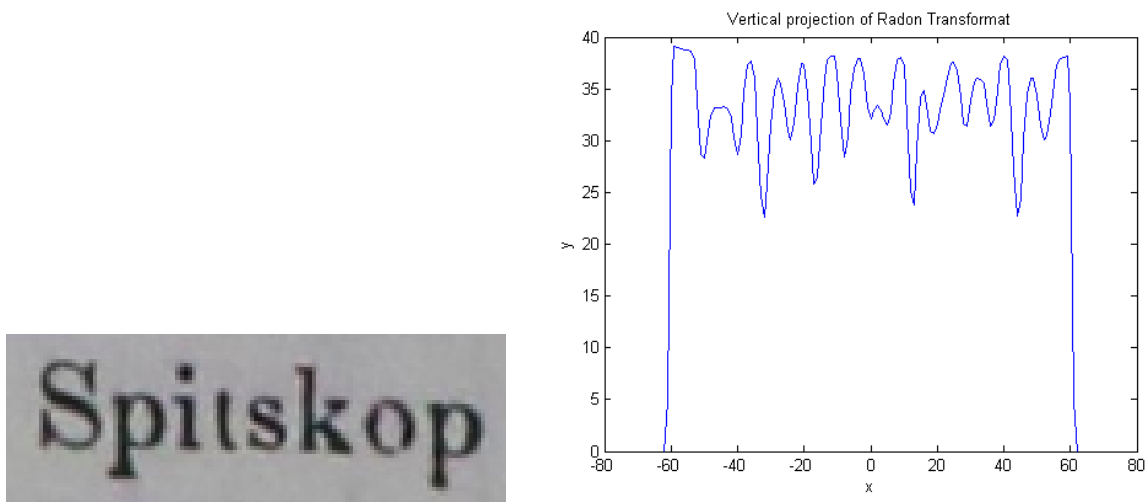


Figure 3.8: Diagram show character segmentation using Radon Transform

3.3 Recognition

This stage of the recognition system is the moment for making the decision of classification the characters by using a neural network. The neural network is a mathematical process that simulates in its way of operation the human brain in the recognition of sounds and images (Singla and Yadav, 2014). After each character has been isolated from the segmentation stage will be represented as a matrix of pixels, where each pixel will be represented by a binary format 0 and 1, which feed into a neural network that has been trained to be able to make a good Association between the previously stored knowledge and the input image of character that is represented by a concatenation of 0,1 (Sharma and Chaudhary, 2013).

In OCR system we benefice from the neurological intelligence of this system to analysis a large data of binary values, by use a simple processing units called neurons to stores the experimental information and make them available to the user by adjusting the weights, which has been trained in each connection process, by made adjustments to the weights to obtain the highest probability of matching with input data (Dhaka et al.). However the characters will be labelled using softmax classifier function (Hentschel, 1998) which take the information from input section as binary character and give the matching probability of this character with the stored characters, where the choosing depends on the high probability of matching which based On previously stored knowledge, Figure 3.9 show us the recognition stag of OCR.

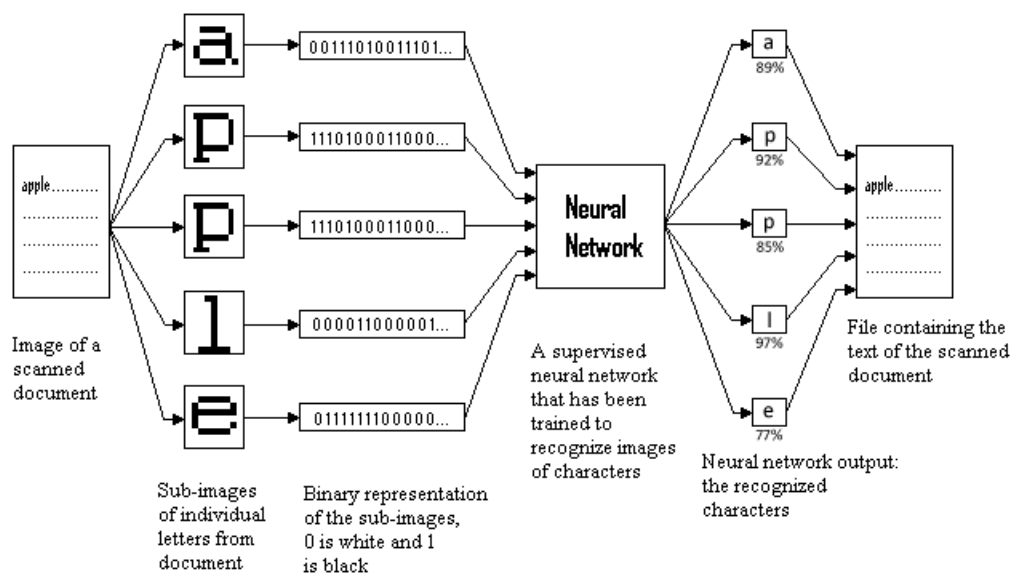


Figure 3.9: Recognition stage of OCR system

The neural networks system consisting of 3 separate sections: Input, Hidden and Output layer. The Hidden layer is a most important section in this system, it contains all complex mathematical operations for classification, the most neural network composed by multi hidden layers of interconnected elements (Eikvil, 1993), each layer gives output of computing the sum of weights, This operation is a nonlinear function, although increasing the number of hidden layers increases the likelihood of output accuracy (Hentschel, 1998).

3.4 Post processing

The process of recognizing the isolated characters has already been done from the previous step, but nevertheless, these character in themselves do not contain the information desired. The post processing stage might be considered as a complementary phase for OCR system to improve the accuracy of the result (Kolak and Resnik, 2005).

In OCR system whatever the system is accurate, it stays very difficult to obtain a result 100% accurate (Eikvil, 1993). Therefore in this stage characters are grouped together and then the errors are detected and corrected, very often the dictionary is used for detection and correction of the words it accomplished. The grouped words are considered as input to the dictionary then by using machine learning the probability of matching will be evaluated and investigated, so that the most correction was done.

4. Data analysis and Results

The data we seek to analyse here is the names of some of South African's voters those belong the period between 1873 and 1900, which is about 60 volumes with pages of one thousand and a few hundred for each of them, the aim of the analysing this data is to check how far the accuracy of one of the best open source OCR, Tesseract to converting those data to texts. Therefore we have selected some samples and tested them, we will address the result in the following chapters.

4.1 OCR tools open source

Tesseract and CuneiForm are the most accurate OCR open source, under Linux(Canonical Ltd) therefore we consider both of them in this project looking for which is most accurate.

4.1.1 Tesseract. Is one of the most open source OCR engine, It was developed by HP in between 1984 to 1994, and It was modified and improved in 1995 with greater accuracy when it was one of the top 3 engines in that time. Tesseract does not come with a GUI and is instead run from the command line interface, it is available for Linux, Windows and Mac OS X(Vithlani and Kumbharana, 2015). In late 2005, HP released Tesseract for open source, released under the Apache License, Version 2.0. Now it is available at <http://code.google.com/p/tesseract-ocr/>(Patel et al., 2012) In 2006 Tesseract was considered one of the most accurate open-source OCR engines then available(Canonical Ltd).

4.1.2 CuneiForm. Is one of the most accurate open source OCR system developed by Russian software company Cognitive Technologies as a commercial product in 1993. In 2008, Cognitive Technologies released CuneiForm for open source, under the BSD variant license(Canonical Ltd).

4.2 Analysing data

- **Tesseract**

The version that use it here is 3.03 with operating system Linux. It's easy we can install it by using following commend:

```
admin@user:~$ sudo apt-get install tesseract-ocr
```

After successful installation, we use the following command to convert imge to text:

```
admin@user:~$ tesseract <path to image> <basename of output file> -l <language>.
```

The language corresponding international codes (loosely based on *ISO 963 – 2*⁽¹⁾). e.g. -eng, -deu, -fra, -ita, -ndl, -por, -spa, ...

The Tesseract will automatically give the output file a .txt extension. The Figure 4.1 show us one sample from the data and the result.

- **CuneiForm**

The version that use it here is 1.1.0 with operating system Linux. It's easy we can install it by

(1) *ISO 639* is a set of international standards that lists short codes for language names.

using the following commend:

```
admin@user:~$ sudo apt-get install cuneiform
```

After successful installation, we use the following command to convert img to text:

```
admin@user:~$ cuneiform <path to image> -l <language> -o <basename of output file>
```

The cuneiform also will automatically give the output file a .txt extension. The Figure 4.2 show us one sample from the data and the result by using CuneiForm software.

NO.	SURNAME.	CHRISTIAN NAME.	RESIDENCE.
255	Hegter	Philip Carl	Spitskop
256	Herbst	Pieter	Graaff-Reinet
257	Herbinger	Christian	ditto
258	Herman	Alexander	Spitskop
259	Hoffer	Hermanus	Graaff-Reinet
260	Hohman	Heinrich	ditto
261	Hoogen, van den	Hermanus Johannes	ditto
262	Hop	Jan	ditto
263	Hornaby	William	ditto
264	Hosking	John Woodcock	ditto
265	Hoy	Charles	ditto
266	Howles	Jan	ditto
267	Hudson	Hougham	ditto
268	Hudson	George	ditto
269	Hugo	Thomas Arnoldus	ditto
270	Hugo	Schalk Johs. Burger	ditto
271	Hurford	George Frederic	ditto
272	Hurford	John	ditto
273	Humphrey	Charles Kennedy	ditto
274	Humphrey	Arthur H.	ditto
275	Ingle	Thomas J.	ditto
276	Isaac	Willem	ditto
277	Isaac	James	ditto
278	Isaac	Thomas Lamb	ditto
279	Isaac	Africa	ditto
280	Jacobus	Robert	ditto

(a) Original

ELECTORAL DIVISION OF GRAAFF-REINET.

```
13
N0. SURNAME. CHRISTIAN NAME. : RESIDENCE.
255 Hegter Philip Carl I Spitskop
256 Herbst Pieter Graaff- Reinet
257 H erbinger Christian ditto
258 Herman Alexander Spitskop
259 Hofl'er Hermanul G raaffl Rein at
260 H ohman Heinrich ditto
261 Hoogen, van den Hermanus Johannes ditto
262 Hop J an ditto
263 Hormby William ditto
264 Hosking John Woodcock ditto
265 Hoy Charles ditto
266 H owles J an ditto
267 Hudson Hougham ditto
268 Hudson George ditto
269 Hugo Thomas Arnoldus ditto
270 Hugo Schalk Johs. Burger ditto
271 Hurford George Frederic ditto
272 H urford J oh n ditto
273 Humphrey Charles Kennedy ditto
274 Humphrey Arthur H . ditto
275 Ingle Thomas J. ditto
276 Isaac Willem ditto
277 Isaac ' James ditto
278 Isaac Th omns In mb ditto
279 Isaac Africa ditto

280 J acobus Robert ditto
```

(b) Tesseract

Figure 4.1: Single projection at a specified rotation angle

The Figure 4.1a show image take of the list of some persons residing in the electoral division of Graaff-Reinet in 1880. The Figure 4.1b show the best result as txt format that get by using Tesseract software in this experiment. when the rat of accurate is 95% that represent by Table 4.1, where from 104 words we have only got 4 errors.

It's not that rate of accuracy we seek to, but is the best result we got from this experiment.

NO.	SURNAME.	CHRISTIAN NAME.	RESIDENCE.
255	Hegter	Philip Carl	Spitskop
256	Herbst	Pieter	Graaff-Reinet
257	Herbinger	Christian	ditto
258	Herman	Alexander	Spitskop
259	Hoffer	Hermanus	Graaff-Reinet
260	Hohman	Heinrich	ditto
261	Hoogen, van den	Hermanus Johannes	ditto
262	Hop	Jan	ditto
263	Hornsby	William	ditto
264	Hosking	John Woodcock	ditto
265	Hoy	Charles	ditto
266	Howles	Jan	ditto
267	Hudson	Hougham	ditto
268	Hudson	George	ditto
269	Hugo	Thomas Arnoldus	ditto
270	Hugo	Schalk Johs. Burger	ditto
271	Hurford	George Frederic	ditto
272	Harford	John	ditto
273	Humphrey	Charles Kennedy	ditto
274	Humphrey	Arthur H.	ditto
275	Ingle	Thomas J.	ditto
276	Isaac	Willem	ditto
277	Isaac	James	ditto
278	Isaac	Thomas Limb	ditto
279	Isaac	Africa	ditto
280	Jacobus	Robert	ditto

(a) Original

E1 ECYORAI DIVISION ()F GPAAFF-B,EINEI'. 13
 NO. SURNAME. CHRtSTKAN NAME,; RESlDEXCE.
 955 l1eg ter Philip Ca.l 8pitskop
 256 Herbs t pieter 0 raa8'-Reinet
 957 Herbinger Christian ditto
 MS Herman Alexander Spitskop
 %9 8o8er Hermangs 0 raafF- Rein st
 960 8 ohman Heinrich ditto
 ,.1
 261 Hoogen, van den Hermanns Johannes ditto
 262 Hop Jan ditto
 968 Hornsby William ditto
 Q64 Hos king John % oodcock ditto
 965 Hoy Charles| ditto
 Q66 Howles Jan ditto
 Q67 Hudson Hoggham ditto
 Q68 Hudson George ditto
 Q69 Hugo Thomas Arnoldus ditto
 970 Hugo Schalk Johs. Burger ditto
 Hgrford George Frederic ditto
 972 Hgrford John (lit to
 Q73 Humphrey Charles Kennedy ditto
 Q74 Hgmphrey A.rthgr H. ditto
 Q75 Ingle Thomas J'. ditto
 Q76 Isaac %illem ditto
 Q77 Isaac . James ditto
 Q78 Isaac Thomas Limb ditto
 Q79 Isaac Africa ditto
 Q80 Jacobgs B,obert ditto

(b) Cuneiform

Figure 4.2: Single projection at a specified rotation angle

The Figure 4.2a show us the same image for 4.1a. The Figure 4.2b show the better result as txt format that get by using Cuneiform software in this experiment. when the rat of accurate is 64% that represent by Table 4.1, whereas from 104 words we get 37 errors, this is high value of errors comparing by Tesseract in this experiment.

4.3 Results

From the data we have select 10 pages from different volumes those blown the age between 1872 to 1880 the table 4.1 show us text results by using OCR open Source Tesseract and CuneiForm.

Table 4.1: Table Title

Volumes reference	Th year	Number of words in page	Tesseract accuracy	Cuneiform accuracy
<i>CCP 11/1/1</i>	1873	94	93%	48%
<i>CCP 11/1/2</i>	1872	109	88%	49%
<i>CCP 11/1/3</i>	1875	90	92%	54%
<i>CCP 11/1/4</i>	1874	128	82%	1%
<i>CCP 11/1/5</i>	1877	101	89%	4%
<i>CCP 11/1/6</i>	1877	121	77%	2%
<i>CCP 11/1/7</i>	1877	188	63%	<i>Failure</i>
<i>CCP 11/1/8</i>	1878	114	67%	55%
<i>CCP 11/1/9</i>	1879	226	92%	41%
<i>CCP 11/1/10</i>	1880	104	95%	64%

5. Conclusion and Future work

5.1 Conclusion

5.2 Future work

Acknowledgements

References

- N. Arica and F. T. Yarman-Vural. An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 31(2): 216–233, 2001.
- T. Arodź. On new radon-based translation, rotation, and scaling invariant transform for face recognition. *Computational Science-ICCS 2004*, pages 9–17, 2004.
- M. N. Asl and A. Sadremomtaz. Analytical image reconstruction methods in emission tomography. *Journal of Biomedical Science and Engineering*, 6(01):100, 2013.
- Canonical Ltd. Ubuntu official documentation. , <https://help.ubuntu.com/community/OCR>, Accessed April 2017.
- J. Chandarana and M. Kapadia. A review of optical character recognition. (*IJERT*), 2:1991–1994, 2013a.
- J. Chandarana and M. Kapadia. A review of optical character recognition. In *International Journal of Engineering Research and Technology*, volume 2. ESRSA Publications, 2013b.
- J. Chandarana and M. Kapadia. Optical character recognition. *IJETAE Tranjact*, 4(5), 2014.
- V. Dhaka, M. Kumar, and H. Sharma. Character recognition of offline handwritten english scripts: A review.
- L. Eikvil. Optical character recognition. *citeseer. ist. psu. edu/142042. html*, 1993.
- T. G. Feeman. The mathematics of medical imaging. *Springer*,, 2010.
- J.-Y. Gan and S.-B. He. Nonlinear radon transform and its application to face recognition. *Pattern Recognit. Artif. Intell*, 4:405–409, 2011.
- D. Hentschel. Creation of a neural network to assist in deciphering degraded ancient hebrew texts. 1998.
- M.-K. Hu. Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8 (2):179–187, 1962.
- R. Jain, R. Kasturi, and B. G. Schunck. *Machine vision*, volume 5. McGraw-Hill New York, 1995.
- A. Khidhir. Use of radon transform in orientation estimation of printed text. In *ICIT 2011 The 5th International Conference on Information Technology*, 2011.
- O. Kolak and P. Resnik. Ocr post-processing for low density languages. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 867–874. Association for Computational Linguistics, 2005.
- M. Miciak. Character recognition using radon transformation and principal component analysis in postal applications. In *Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on*, pages 495–500. IEEE, 2008.
- C. Patel, A. Patel, and D. Patel. Optical character recognition by open source ocr tool tesseract: A case study. *International Journal of Computer Applications*, 55(10), 2012.

- J. Patel, A. Shah, and H. Patel. Skew angle detection and correction using radon transform. *International Journal of Electronics, Electrical and Computational, System, ISSN*, 2015.
- A. Sharma and D. R. Chaudhary. Character recognition using neural network. *International Journal of Engineering Trends and Technology (IJETT)*, 4(4):662–667, 2013.
- A. A. Shinde and D. Chougule. Text pre-processing and text segmentation for ocr. *International Journal of Computer Science Engineering and Technology*, pages 810–812, 2012.
- S. Singla and R. Yadav. Optical character recognition based speech synthesis system using labview. *Journal of applied research and technology*, 12(5):919–926, 2014.
- P. A. Toft and J. A. Sørensen. *The Radon transform-theory and implementation*. PhD thesis, Technical University of Denmark Danmarks Tekniske Universitet, Department of Informatics and Mathematical Modeling Institut for Informatik og Matematisk Modellering, 1996.
- G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis. A complete optical character recognition methodology for historical documents. In *Document Analysis Systems, 2008. DAS'08. The Eighth IAPR International Workshop on*, pages 525–532. IEEE, 2008.
- P. Vithlani and C. Kumbharana. Comparative study of character recognition tools. *International Journal of Computer Applications*, 118(9), 2015.
- Wikipedia. Optical character recognition. Wikipedia, the Free Encyclopedia, https://en.wikipedia.org/wiki/Optical_character_recognition#Character_recognition, Accessed MAY 2017.