Freie Universität Berlin

The Transition Method for Binarization

Doctoral Dissertation
by
Marte Alejandro Ramírez Ortegón

Supervisor:
Prof. Dr. Raúl Rojas González

Co-supervisor:
Dr. Ernesto Tapia Rodríguez

Berlin, Germany
February 2, 2011

The Transition Method for Binarization

Dissertation

presented by

Marte Alejandro Ramírez Ortegón

submitted to the

Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Under the supervision of

Prof. Dr. Raúl Rojas González

and co-supervision of

Dr. Ernesto Tapia Rodríguez

Berlin, February 2011

First Referee: Prof. Dr. Raúl Rojas González

Second Referee: Dr. Mariano José Juan Rivera Meraz

Day of Disputation: February 17, 2011

Abstract

This thesis introduces a novel binarization method based on the concept of t-transition pixel. It includes five main contributions. The first contribution is a generalization of edge pixels, namely t-transition pixel. Such pixels are characterized with high transition values computed with discriminant functions called transition functions. In particular, maxmin function is proposed and widely analyzed. The second contribution is the formalization of the transition method for binarization, and to a minor degree, for edge detection, and for detection of regions of interest. In this method, binarization is performed by extracting information only from transition pixels. Comparison studies show that it greatly outperforms other top-ranked binarization methods. Furthermore, potential applications in edge detection and detection of regions of interest are observed. Two minor contributions are derived from the transition method: unimodal thresholds for transition values, and morphological transition operators to extract and restore transition sets. The third contribution is a mathematical analysis of unsupervised measures for segmentation quality, in which the strengths of the weighted variance measure are proved. From this analysis, the uniform variance measure and measures based on logarithms of gray intensities are proposed. The fourth contribution is a mechanism for systematic comparison of the efficacy of unsupervised evaluation methods for parameter selection of binarization algorithms in optical character recognition (OCR). Moreover, a statistical test is proposed to compare measures based on an intuitive triad of possible results: better, worse or comparable performance. The fifth contribution is addressed in a new chapter, which introduces a novel unbiased and efficient slope estimator for linear regression model. The computational cost of this estimator is considerably lower than the current state of the art.

Contents

A mi madre

To my lovely Umporn

Acknowledgments

First and foremost, to my friend and co-supervisor, Ernesto Tapia, for working so hard on this thesis and truly understanding what I meant in my very confused explanations. And to my incomparable Professor Raúl Rojas for his full support during my doctoral studies.

I cannot fully express my gratitude to my Mexican friends in Germany, for their generosity, guidance and superb friendship. Thank you especially to Erik Cuevas, who encouraged me to take three wonderful adventures in my life: Guatemala, India, and Tokyo. My thanks also to Aldo Mirabal for his friendship from the very beginning, as well as to Freddy Villafuerte, Daniel Zaldivar, and Dan-El Vila who cheered me all the time.

Me, Ernesto, Daniel, Erik, Aldo, and Freddy.

Lilia Leticia Ramírez Ramírez and Edgar Alfredo Duéñez Guzmán, two dear friends, helped me uncountable times to write this thesis. Their careful reading

and always-practical suggestions have made a better thesis. To both of them, my deepest thankfulness.

My gratitute to The National Council on Science and Technology (CONACYT) of Mexico for its economic support during my whole doctoral studies (grant number: 218253), and the Freie Universität Berlin for being the host of my doctoral studies.

I am really happy to have met my friend Chang-Chien Fang, who taught me to value any trip in my life, and my dear Megumi Isoyama, who was an oasis of peace.

And finally, I would like to mention two extraordinary women who have touched my life. First, my mother, Maria Elena Ortegón Rivero, for her immeasurable love throughout my whole life. And my fiancée, Umporn Athikomrattaankul, chemist, wonderful chef, great player of detective games, and without a doubt the most astonishingly talented woman I have ever known.

Notation

Variables and Symbols

$\alpha, \beta, \ldots$	In algorithms, a lower case Greek letter denotes a tuning parameter.
$a, b, \ldots$	Lower case letter denotes a scalar.
$\boldsymbol{p}, \boldsymbol{q}, \ldots$ or $\boldsymbol{p}_{i,j}, \boldsymbol{q}_{i,j}, \ldots$	Lower bold letter denotes a pixel. I use sub-indexes in a pixel if the pixel position is referred.
$x, y, \ldots$	Sans Serif Font and italic shape denotes a random variable.
$\hat{\cdot}$	Estimator of $\cdot$.

Sets

$\mathcal{A}, \mathcal{B}, \ldots$	Upper case "calligraphic" letters denote sets.
$\mathcal{A} \cup \mathcal{B}$	Union of two sets. That is, the set containing all elements in either $\mathcal{A}$ or $\mathcal{B}$.
$\mathcal{A} \cap \mathcal{B}$	Intersection of two sets. That is, the set containing all elements that are in both $\mathcal{A}$ or $\mathcal{B}$.
$\lvert\mathcal{A}\rvert$	The cardinality of set $\mathcal{A}$.
$\hat{\mathcal{A}}$	Approximation set of $\mathcal{A}$. That is, $\lvert\mathcal{A}\rvert \approx \lvert\hat{\mathcal{A}}\rvert \approx \lvert\mathcal{A} \cap \hat{\mathcal{A}}\rvert$.
$p \in \mathcal{A}$	$\boldsymbol{p}$ is an element of $\mathcal{A}$.
$p \notin \mathcal{A}$	$\boldsymbol{p}$ is not an element of $\mathcal{A}$.

Reserved Set Names

$\mathbb{N}$	Set of natural numbers. That is, $\mathbb{N} = \{0, 1, 2, \ldots\}$

$\mathbb{Z}$	Set of integer numbers. That is, $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathcal{B}$	Background set.
$\mathcal{F}$	Foreground set.
$\mathcal{P}$	Union of foreground and background.
$\mathcal{P}_r(\boldsymbol{p})$	$\mathcal{P}_r(\boldsymbol{p}) \subset \mathcal{P}$ is a squared neighborhood centered at the pixel $\boldsymbol{p}$ of sides with length $2r + 1$.
${}_t\mathcal{P}$	t-transition set. That is, ${}_t\mathcal{P} = \{\boldsymbol{p} \mid \mathcal{P}_t(\boldsymbol{p}) \cap \mathcal{B} \neq \emptyset, \mathcal{P}_t(\boldsymbol{p}) \cap \mathcal{F} \neq \emptyset\}$
${}_t\mathcal{B}$	Negative transition set. That is, ${}_t\mathcal{B} = {}_t\mathcal{P} \cap \mathcal{B}$.
${}_t\mathcal{F}$	Positive transition set. That is, ${}_t\mathcal{F} = {}_t\mathcal{P} \cap \mathcal{F}$.
${}_t\mathcal{P}_r(\boldsymbol{p})$	Transition set in the pixel neighborhood. That is, ${}_t\mathcal{P}_r(\boldsymbol{p}) = {}_t\mathcal{P} \cap \mathcal{P}_r(\boldsymbol{p})$
${}_t\mathcal{B}_r(\boldsymbol{p})$	Negative transition set in the pixel neighborhood. That is, ${}_t\mathcal{B}_r(\boldsymbol{p}) = {}_t\mathcal{B} \cap \mathcal{P}_r(\boldsymbol{p})$
${}_t\mathcal{F}_r(\boldsymbol{p})$	Positive transition set the pixel neighborhood. That is, ${}_t\mathcal{F}_r(\boldsymbol{p}) = {}_t\mathcal{F} \cap \mathcal{P}_r(\boldsymbol{p})$

Functions and Images

$A, B, \ldots$	Capital letter denotes functions or images.
$\tilde{A}_S, \tilde{B}_S, \ldots$	Integral image.
$F(\boldsymbol{p})$	Value of the function or image F in the point $\boldsymbol{p}$.
$\arg\max_{x \in \mathcal{A}} \{F(x)\}$	The value of x that leads to the maximum value of $F(x)$ in set $\mathcal{A}$.
$\arg\min_{x \in \mathcal{A}} \{F(x)\}$	The value of x that leads to the minimum value of $F(x)$ in set $\mathcal{A}$.
$\max_{x \in \mathcal{A}} \{F(x)\}$	The maximum $F(x)$ value in set $\mathcal{A}$.
$\min_{x \in \mathcal{A}} \{F(x)\}$	The minimum $F(x)$ value in set $\mathcal{A}$.
$H_{F,\mathcal{A}}$	Histogram of F in set $\mathcal{A}$.
$H_{F,\mathcal{A}}(y)$	Frequency of the value $y = F(x)$ in set $\mathcal{A}$. This is, $H_{F,\mathcal{A}}(y) = \lvert\{x \in \mathcal{A} \mid F(x) = y\}\rvert$

Reserved Image Names

B	Binary image function. $B : \mathcal{P} \rightarrow \{0, 1\}$.
I	Gray-level image function. $I : \mathcal{P} \rightarrow \{0, 1, \ldots, g-1, g\}$.
V	Transition value image. $V : \{\mathcal{P}_s(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \rightarrow \{-g, -g+1, \ldots, g-1, g\}$.
T	Threshold image. $T : \{\mathcal{P}_r(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \rightarrow \{0, 1, \ldots, g\}$.

Probability and Distributions

$\mu_{F,A}$	Mean of F within a set $\mathcal{A}$.
$\sigma^2_{F,A}$	variance of F within a set $\mathcal{A}$.
$\hat{\mu}_{F,A}$	Estimator of the mean of F within a set $\mathcal{A}$.
$\hat{\sigma}^2_{F,A}$	Estimator of the variance of F within a set $\mathcal{A}$.
$E(x)$	Expected value of a random variable x.
$\Pr(\cdot)$	Probability.
$\Pr(p \in \mathcal{A})$	Probability that a pixel $\boldsymbol{p}$ belongs to $\mathcal{A}$.
$Var(x)$	Variance of random variable x.
$\phi(\mu, \sigma^2)$	Probability density function of normal distribution mean μ and variance σ^2.

$$\phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$\lambda(\tilde{\mu}, \tilde{\sigma}^2)$	Probability density function of lognormal distribution with parameters μ and σ^2.

$$\lambda(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right).$$

Chapter 1

Introduction

What is essential is invisible to the eye.

The Fox in The Little Prince
by Antoine de Saint-Exupéry
French writer (1900-1944)

Libraries such as **the General Archive of the Nation** (México) [52], **the Library of Congress** (United States of America) [63], and **the National Archives of Egypt** [64] have been digitalizing historical printed documents like ancient codices, maps, newspapers and books to preserve and spread their cultural heritage.

While digitization in itself is enough to preserve the contents of documents, a primordial benefit of digitization is the extraction of information from the digitalized images, and the access to this information through **digital libraries**.

A digital library is a portal site wherein the public can remotely search, visualize and download digitalized images of documents. Moreover, user may have access to historical and ancient documents whose physical consultation is unavailable due to security or preservation reasons. *digital library*

The main problem in the construction of digital libraries lies in the extraction of information from hundreds of thousands of ancient documents. For example, since the establishment of the National Archives of Egypt (NAE), the number of documents without indexing or classification accumulated in its stores has exceeded one hundred million. Because of this, the access to a certain document in

a particular subject, even if bibliographic records are available, is difficult. The digitalization of bibliographic records is the only feasible solution to that problem. The NAE plans to create a database with around 25 million records to index a hundred million documents.[1] This enormous labor requires the automation of the greatest possible number of processes.

The automatic extraction of information from scanned images of historical documents presents several difficulties. Documents use non-standard fonts and have different types and degrees of degradation, such as:

1. **Artifacts due to printing**: weak strokes, ink stains, smudged characters, **bleed-through**.

When the paper is too thin or the ink applied too heavily, the color can bleed or seep through to the other side. This is known as **bleed-through**.

2. **Artifacts due to aging**: dark spots (humidity or burns) and outlines of paper folds.
3. **Slanted characters**.
4. **Rotated characters**.
5. **Varying kerning** (space between characters),
6. **Varying leading** (space between lines).
7. **Line-break hyphenation**.

In this thesis, I tackle the problem in documents caused by artifacts due to aging and due to printing. My novel approach, which I have named the **transition method**, is based on a generalization of edge pixels.

My aim in this thesis is to systematically describe the transition method for identifying characters and relevant strokes in documents. Nevertheless, I also superficially describe how the transition method can be used for edge detection, and the detection of regions of interest. I also dedicate an entire chapter to exploring unsupervised measures to assess the binarization accuracy.

I assume that the objects of interest in a document can be distinguished by extracting diverse features based on the gray intensity and the spatial position of pixels. The transition method particularly exploits the distribution and spatial relationship of the difference of gray intensity between objects and background,

[1] Information about this project is available in:
http://www.nationalarchives.gov.eg/nae/Content?id=_37

modeling the distribution of gray intensities of both objects and background in small neighborhoods.

Even though the transition method has the potential to deal with uneven illumination, this thesis will focus only on images without sudden illumination changes in small neighborhoods.

1.1 Binarization

Conceptually, images often have a natural partition between foreground and background. Intuitively, **binarization** consists of estimating such a partition, where we consider as foreground the set of pixels in an image containing the objects of interest and the background representing the rest of the image. *binarization*

What constitutes foreground depends on the objects to be recognized. While OCR applications are interested in the location and extraction of ink with high contrast [50, 53], understanding the information to be extracted in documents can depend on the objects and their relationships. Figure 1.1, for instance, shows (left) a triangle and (right) grid lines with similar gray intensities. Both images contain dark pixels which certainly belong to the foreground. However, in (left) we may keep the triangle in the binary image, in (right) we could possibly remove the grid.

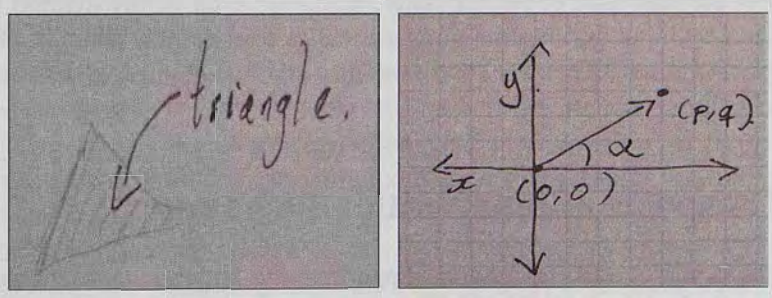


Figure 1.1 – *The context changes the definition of foreground.*

The previous example shows that binarization is a complex problem if only gray intensities are considered in the binarization process. Contextual information is needed to solve problems similar to those in Fig. 1.1. However, even though color images provide more information than those in gray intensities, few binarization researchers, like Kohmura and Wakahara [41], work directly with color images because of the computational load and analysis complexity. The rest of

the researchers transform an image from color to gray intensities before applying binarization algorithms. For example:

1. Chou et al. [13] developed a binarization system for images produced by cameras which deals with uneven illuminated images. They divide a gray-intensity image into several regions and decide how to binarize each region further.

2. Caron et al. [7] detect regions of interest characterizing each pixel with a template of gray intensities of 3×3, the frequency of which appears to obey a power law distribution.

3. Milewski and Govindaraju [53] presented a methodology for separating handwritten letters from background in carbon–copied medical forms. They compare the mean of gray intensities of small neighborhoods around the pixel of interest.

4. Both Chen et al. [11] and Mello et al. [50] binarize documents using gray-intensity images as input. Whereas Chen et al. generate the binary image from the edge image of the gray-intensity image, Mello et al. compute a threshold based on a weighted entropy equation.

I follow the approach of the previous examples. That is, my method takes a gray image I as input and returns a binary image B as output, wherein pixels in white represent the background approximation and pixels in black represent the foreground approximation.

1.2 Overview of binarization techniques

Several authors [79], [80] [87] have categorized the binarization algorithms according to where the information to compute the pixel threshold came from. In this manner, **global algorithms** label each pixel using information from the whole image while **local algorithms** rely on information from the pixel neighborhood. **Hybrid algorithms** combine information from the whole image and pixel neighborhood. Notice, however, that both global and hybrid algorithms can be transformed into local versions by restricting the analysis to the pixel neighborhood.

With the aim of overcoming composite foreground and background areas, all algorithms considered in this thesis were implemented as local algorithms even though some of them were originally global or hybrid algorithms.

:sholding

Thresholding algorithms are a particular type of binarization where a pixel is classified as foreground if the gray intensity of the pixel is darker than a threshold.

I categorized some thresholding algorithms related to my approach, based on which features of gray intensities the algorithm manipulates. Hence, a binarization algorithm may fit into two or more categories.

histogram cluster binarization

Histogram cluster binarization algorithms assume that the foreground can be estimated by those pixel whose gray intensity is lower than or equal to some threshold. They take as input the histogram of gray intensities. Classical examples of these algorithms are Kittler's, Otsu's, and Portes's thresholding.

The **minimum error thresholding** [40] (**Kittler's threshold**) maximizes the likelihood of the joint distribution of gray intensities assuming that foreground and background are normally distributed with different means and variances[2]. In contrast, **Otsu's threshold** [66], without assuming an a priori distribution, minimizes the sum of the variance of gray intensities of foreground and background. **Portes's threshold** [67] maximizes the **nonextensive entropy**, also called **Tsallis entropy** [88], of both foreground and background.

Statistical binarization

Statistical binarization algorithms are another class of binarization algorithms, which rely on information from statistics of gray intensities. These algorithms usually compute the mean and variance of gray intensities in the pixel neighborhood. I compared my approach specifically with four of these algorithms: **Kavallieratou's algorithm** [36], [37]; **Niblack's** [61]; **Sauvola's** [80]; and **Wolf's** [90] algorithms.

Kavallieratou's algorithm sets to white the pixels with a gray intensity above the local mean while the rest of the pixels are normalized. The process is iterated until a stopping criterion is satisfied. Sauvola's algorithm is a modified version of Niblack's algorithm; both algorithms assume that the gray intensities of the background are approximately normally distributed and select a threshold as the lower limit of an interval centered in the local mean of gray intensities. Wolf and Jolion modified the equation of Sauvola's threshold by adding the minimum gray intensity of the pixel neighborhood and the maximum standard deviation of gray intensities of all neighborhoods, which act as dynamic variance-normalization factors.

edge-contrast binarization

Edge-Contrast binarization algorithms exploit edge information and local contrast of gray intensities. These algorithms assume that there is a large dif-

[2] Sezgin and Sankur [82] present an exhaustive categorization of thresholding. They affirm that the minimum error thresholding and Sauvola's threshold are the best-scored algorithms binarizing documents uniformly illuminated and degraded with noise and blur.

ference between the gray intensity of foreground and background while the gray intensities within each set do not differ significantly. Indeed, the foreground and background may correspond to those pixels whose gray intensities are the minimum and maximum in the neighborhood, respectively, in the ideal situation. Some examples of this kind of binarization algorithms are Bersen's, Kamel's, Oh's, Li's, and Chen's algorithms.

Bersen's algorithm [3] computes a threshold which lies between the maximum and minimum gray intensity in the neighborhood. More sophisticated edge-contrast algorithms have been proposed by **Kamel** [34], and **Oh** [65]. These binarization methods use the contrast of gray intensities between small neighborhoods around the pixel of interest.

Li's algorithm [44] uses the **Laplace operator** and a covariance matrix of gray intensities to compute a threshold. **Chen's method** [11] applies the **Canny edge detector** [6] to generate the edge image. Several morphological operators subsequently help to generate an enhanced binary image. Both algorithms apply a criterion for selecting pixels with high information content.

Remark 1.1: In this thesis, I refer as "*method*" to those algorithms whose subtasks can be performed with different algorithms such that the election of any of these "*sub-algorithms*" in a step may lead to different binarization results. For example, suppose that a binarization algorithm requires edge detection. This task can be performed by Canny's, Prewitt's [47], or Robert's Cross algorithms, to mention some; since the output of these algorithms may differ from each other, the binarization results may change according to which algorithm performs the edge detection.

spatial binarization

Spatial binarization algorithms gather information from spatial relationship between gray intensities. Some edge-contrast binarization algorithms, like Kavallieratou's, Oh's and Kamel's algorithms, could be classified as spatial binarization algorithms because they analyze relationships of gray intensities between small neighborhoods. Lu's and Yanowitz's algorithms also fall within this category.

Lu's algorithm [46] computes a polynomial surface for modeling shading fluctuations of gray intensities. Likewise, Yanowitz and Bruckstein [5] (**Yanowitz's method**) proposed an adaptive threshold surface, determined by interpolation of the image gray intensities at pixels where the image gradient is high.

1.3 Overview of this thesis

Chapter 2 introduces and formalizes preliminary concepts of digital images (pixel, image and neighborhood). It also introduces some morphological operators that I will use later on. Readers who are not interested in such a meticulous formalism can skip this chapter. However, I advise not to skip Section 2.2 where the concept and notation of neighborhoods are defined.

The purpose of Chapter 3 is to examine the local implementation, assumptions and variants of several binarization algorithms which are either related to my method or considered reference algorithms in the binarization literature.

The first main contribution of my thesis is enclosed in Chapter 4, where I propose and describe the concept of **t-transition pixel** from which I derived a novel approach for binarization, edge detection and detection of region of interest. The theory of **transition set**, **transition functions**, and **transition values** is also introduced and developed in this chapter. Specifically, I describe the transition function **maxmin**.

The second main contribution of my thesis is in Chapter 5. I mathematically describe **the transition method** in gray images for binarization, and to a minor degree, for edge detection, and for detection of regions of interest. Several binarization methods based on the transition method are proposed. Additional to these binarization methods, I describe a simple method for edge detection and a simple method for detection of regions of interest.

In Chapter 6, I address the problem of parameter selection of binarization algorithms. I review several unsupervised evaluation methods to assess the quality of a segmentation, and propose a several novel measures based on the normal and lognormal distribution. I also statistically analyze each of the reviewed measures and ascertain whether a measure is suitable or not to assess a binarization algorithm.

I summarize the results of my publications [72], [73], and [70] in Chapter 7 where I propose a mechanism for systematic comparison of the the efficiency of unsupervised evaluation methods for parameter selection of binarization algorithms in optical character recognition (OCR). I also analyze and compare binarization algorithms based on the transition method with several top-ranked binarization algorithms.

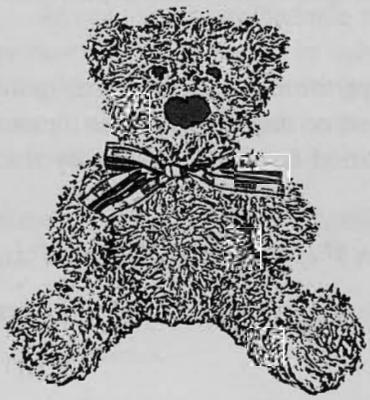
Finally, Chapter 8 introduces a novel estimator for the slope parameter in a simple linear regression. This estimator is unbiased and efficient. Moreover, I show that it has a low computational cost.

Two appendixes are to be noted: Chapter A extends the **integral image con-**

cept to efficiently compute any statistical moment in subsets of pixels in neighborhoods of radius r. This is particularly useful for the transition method, and for statistical binarization methods. In Chapter C, I develop the **uncertainty test** to compare measures based on an intuitive triad of possible results: better, worse or comparable performance whereby I ascertain that an algorithm is better than another in my experiments.

Chapter 2

Digital images



The journey of a thousand miles
begins with one step.

Lao Tse
Philosopher of ancient China

The aim of this chapter is threefold. Firstly, digital images are characterized as partitions of continuous images. Later on, the concept of neighbor and neighborhood of a pixel is introduced. Finally, the last section describes some morphological operators which will be frequently referred to in further chapters.

2.1 Digital images

Digital images are typically given as sets of discrete points due to the discrete process of image acquisition, and the discrete nature of computers from which image processing theory develops. In fact, the acquisition of a two-dimensional digital image from a camera or scanner is done through a set of sensors. How that

process is done is beyond the scope of this thesis; readers interested in pursuing the subject further may consult Gonzalez and Woods [26].

A **continuous image** $\bar{\mathcal{P}}$ within $[0, w] \times [0, h]$ can be modeled as a finite **partition** of the continuous plane within $[0, w] \times [0, h]$. The term **pixel** is used to refer to each element of this partition. A pixel represents an area in two-dimensional images.

A **partition** of a set $\mathcal{A}$ is a division of $\mathcal{A}$ into non-overlapping and non-empty regions that cover all of $\mathcal{A}$. A **finite partition** is a partition with a finite number of regions.

Definition 2.1: *A continuous image $\bar{\mathcal{P}}$ is a connected and continuous n-dimensional region.*

Remark 2.1: $\bar{\mathcal{P}}$ can have any form and, consequently, it does not necessarily take the form of an n-hypercube.

Definition 2.2: *A pixel $\boldsymbol{p}$ is an n-dimensional region of an n-dimensional partition $\mathcal{P}$ in an n-dimensional image $\bar{\mathcal{P}}$.*

Two elements $\boldsymbol{p}$ and $\boldsymbol{q}$ of a two-dimensional partition $\mathcal{P}$ strictly satisfy $\boldsymbol{p} \cap \boldsymbol{q} = \emptyset$. However, this definition is in conflict with some definitions in this thesis. Therefore I consider a *softer* definition of a partition for images given by the following definition.

Definition 2.3: *A **two-dimensional image partition** $\mathcal{P}$ of a continuous image $\bar{\mathcal{P}}$ is a set of regions $\boldsymbol{p}$ such that*

$$\iint \boldsymbol{p} \neq 0 \quad \forall \boldsymbol{p} \in \mathcal{P}, \tag{2.1}$$

$$\bigcup_{\boldsymbol{p} \in \mathcal{P}} \boldsymbol{p} = \bar{\mathcal{P}}, \tag{2.2}$$

$$\iint \boldsymbol{p} \cap \boldsymbol{q} = 0, \tag{2.3}$$

where

$$\iint \cdot \tag{2.4}$$

denotes the area of $\cdot$.

Notation:

A **grid** is a partition of non-overlapping squares with sides of constant length.

In this thesis, images are divided in a **grid** such that $\boldsymbol{p}_{0,0}$ represents the element

at the top-left corner of the partition and $\boldsymbol{p}_{h-1,w-1}$ represents the element at the bottom-right corner, where h and w denote the number of rows and columns in the rectangular partition, respectively. Note the swap of the axes. The notation of a pixel $\boldsymbol{p}_{i,j}$ is simplified to $\boldsymbol{p}$ if its spatial location is irrelevant.

Definition 2.4: *A **two-dimensional digital image**, or **image function** (these terms will be used interchangeably throughout this thesis), is a function*

$$F : \mathcal{A} \to \mathbb{Z}^d, \tag{2.5}$$

where $\mathcal{A} \subset \mathcal{P}$ and d is a positive integer.

Image values can become negative during processing or as a result of interpretation. For example, in radar images, objects moving toward a radar system often are interpreted as having negative velocities while objects moving away are interpreted as having positive velocities. Thus, a velocity image might be coded as having both positive and negative values.

Remark 2.2: Let $\mathcal{P}$ be a partition of a continuous image $\bar{\mathcal{P}}$. Then, $\boldsymbol{p} \subset \bar{\mathcal{P}}$ (not $\boldsymbol{p} \in \bar{\mathcal{P}}$) and $\boldsymbol{p} \in \mathcal{P}$.

Remark 2.3: The word image(s) refers to two dimensional digital image(s) throughout this thesis.

A digital image from cameras and scanners is typically represented by an image $C : \mathcal{P} \to \mathbb{N}^3$, such that the triplet $C(\boldsymbol{p}) = (C_{red}(\boldsymbol{p}), C_{green}(\boldsymbol{p}), C_{blue}(\boldsymbol{p}))$ represents the color intensity of $\boldsymbol{p}$, with $C_{red} : \mathcal{P} \to \mathbb{N}$ for the red intensity, $C_{green} : \mathcal{P} \to \mathbb{N}$ for the green intensity, $C_{blue} : \mathcal{P} \to \mathbb{N}$ for the blue intensity. Each component of this triplet can vary from zero to a defined maximum value, usually black for $(0, 0, 0)$ and white for $(255, 255, 255)$.

Because of the computational load, researchers usually transform a color image into a gray image, which is an image where only gray tones, including white and black, are present. Formally:

Definition 2.5: *A **gray image** is a two-dimensional digital image*

$$I : \mathcal{P} \to \mathbb{Z}_{g+1}. \tag{2.6}$$

where $\mathbb{Z}_{g+1}$ the set of congruence classes modulo $g+1$. That is $\mathbb{Z}_{g+1} = \{0, 1, \ldots, g\}$.

Gray images are commonly stored with 8 bits per pixel, which allows 256 ($g = 255$) different intensities to be recorded. The color black is then represent with zero, the color white with g and shades of gray are linearly represented with integers i such that $0 < i < g$. The precision provided by this format is sufficient to avoid visible banding artifacts[1] and convenient for programming due to a pixel occupies a Byte (8 bits).

As I point out in the introduction, most of the binarization thresholds use a gray image as input. Because of this, color images are transformed into gray images by a mapping $\gamma : \mathbb{Z} \to \mathbb{Z}_{g+1}$. For simplicity, I define I as the gray image $\gamma \circ F$, where F is a color image. Notice that while I depends both on the image F and the gray-intensity map γ, this dependency will always be clear from the context and thus will be left implicit. The transformation γ may vary according the applications and further methods. In this thesis, I use the transformation

$$\begin{aligned} C(\boldsymbol{p}) &\xrightarrow{\gamma} I(\boldsymbol{p}) \\ (C_{red}(\boldsymbol{p}), C_{green}(\boldsymbol{p}), C_{blue}(\boldsymbol{p})) &\xrightarrow{\gamma} \frac{299C_{red}(\boldsymbol{p}) + 587C_{green}(\boldsymbol{p}) + 114C_{blue}(\boldsymbol{p})}{1000}. \end{aligned} \tag{2.7}$$

As a result of the binarization process, each pixel is associated with either 1 (foreground), or 0 (background), but not both. Formally, these images can be defined as

Definition 2.6: *A **binary image** is a two-dimensional digital image*

$$B : \mathcal{P} \to \{0, 1\}. \tag{2.8}$$

Because binary images can be stored as a single bit (0 or 1), they are also called **bi-level** or **two-level image**.

2.2 Neighborhoods

Since image operators will be defined in subsets of $\mathcal{P}$, the notion of *neighbors* of a pixel is defined in a similar manner to neighbors of a vertex in graph theory. That is, pixels could be seen as vertexes in a graph and the relation "$\boldsymbol{p} \in \mathcal{P}$ neighbor of $\boldsymbol{q} \in \mathcal{P}$" as an edge between $\boldsymbol{p}$ and $\boldsymbol{q}$.

The simplest definition of a neighbor of a pixel is given as those pixels whose areas share a common edge with the pixel of interest, namely direct neighbors.

[1] See Gonzalez and Woods [26], pages 62-65

Definition 2.7: *Let $\mathcal{P}$ be a grid of an image; the* ***cross neighborhood*** *of a pixel $\boldsymbol{p}_{i,j}$ is the set*

$$\mathcal{P}_+(\boldsymbol{p}_{i,j}) = \left\{\boldsymbol{p}_{h,k} \in \mathcal{P} \mid \{h,k\} \in C_+(i,j)\right\}, \tag{2.9}$$

where

$$C_+(i,j) = \{\{i-1,j\},\{i+1,j\},\{i,j-1\},\{i,j+1\}\}. \tag{2.10}$$

Likewise, the diagonal neighborhood of a pixel is the set of pixels which share a point with the pixel in question; see Fig. 2.1.

Definition 2.8: *Let $\mathcal{P}$ be a grid of an image; the* ***diagonal neighborhood*** *of a pixel $\boldsymbol{p}_{i,j}$ is the set*

$$\mathcal{P}_\times(\boldsymbol{p}_{i,j}) = \left\{\boldsymbol{p}_{h,k} \in \mathcal{P} \mid \{h,k\} \in C_\times(i,j)\right\}, \tag{2.11}$$

where

$$C_\times(i,j) = \{\{i-1,j-1\},\{i-1,j+1\},\{i+1,j-1\},\{i+1,j+1\}\}. \tag{2.12}$$

Both cross and diagonal neighborhoods are frequently used in image operators that remove the noise of binary images. However, most of the binarization algorithms and image operators in this thesis are defined in terms of the square neighborhood centered at the pixel of interest.

Definition 2.9: *Let $\mathcal{P}$ be a grid of an image; the* ***rectangular neighborhood*** *of a pixel $\boldsymbol{p}_{i,j}$ is the set*

$$\mathcal{P}_{y,x}(\boldsymbol{p}_{i,j}) = \left\{\boldsymbol{p}_{h,k} \in \mathcal{P} \mid \boldsymbol{p}_{h,k} \neq \boldsymbol{p}_{i,j},\ |h-i| \leq y \text{ and } |k-j| \leq x\right\}, \tag{2.13}$$

where $|\cdot|$ denotes the absolute value. In particular a ***square neighborhood*** *is a rectangle neighborhood where $x = y$.*

Notation:

In the following chapters, local binarization algorithms are defined in squared neighborhoods of radius r so that the notation $\mathcal{P}_{r,r}(\boldsymbol{p}_{i,j})$ is simplified to $\mathcal{P}_r(\boldsymbol{p}_{i,j})$.

Finally, let me introduce the concept of the general position for pixels and for neighborhoods, which is needed for definitions in Section 2.3 and Section A.

Definition 2.10: *Given two integers x and y, a pixel $\boldsymbol{p}_{i,j}$ is in a* ***general position*** *if and only if $\boldsymbol{p}_{h,k} \in \mathcal{P}$ for all pair of integers h and k such that $|h - i| \leq y$ and $|k - j| < x$.*

Definition 2.11: *A rectangular neighborhood $\mathcal{P}_{y,x}(\boldsymbol{p})$ is in a general position if and only if $\boldsymbol{p}$ is in a general position.*

Notation:
For simplicity, the intersection set of $\mathcal{A}$ with $\mathcal{P}_r(\boldsymbol{p})$ is denoted as $\mathcal{A}_r(\boldsymbol{p})$. For example, $\mathcal{F}_r(\boldsymbol{p}) = \mathcal{F} \cap \mathcal{P}_r(\boldsymbol{p})$, $\mathcal{B}_r(\boldsymbol{p}) = \mathcal{B} \cap \mathcal{P}_r(\boldsymbol{p})$, and so on.

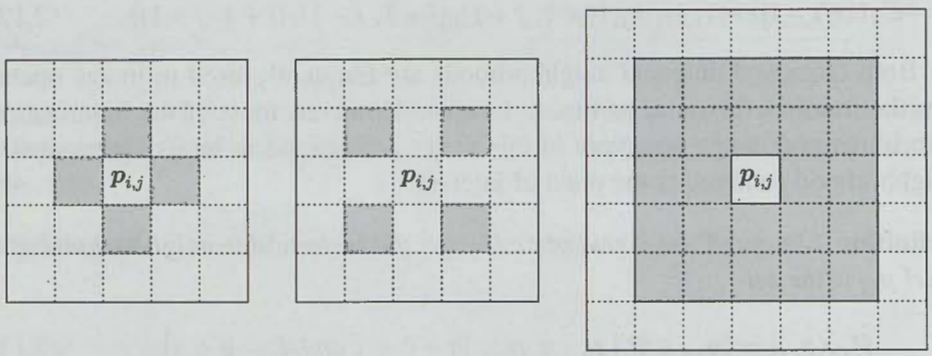


Figure 2.1 – *From left to right: cross neighborhood, diagonal neighborhood and square neighborhood. Pixels in the neighborhood of $\boldsymbol{p}_{i,j}$ are shown in gray.*

2.3 Morphological operators

The basis of mathematical morphology is given by set theory and, more specifically, by **Minkowski algebra**. I attempt to introduce some basic morphological operators which I use to remove noise in binary images. For further details on the field of mathematical morphology and a formal introduction of the operators below, the interested reader is referred to [26], [48], [61], and [81]. For the purpose

of this thesis, I describe two major morphological operations.

2.3.1 Isolate operators

Definition 2.12: *The **simple isolate operator** is defined as*

$$\mathcal{U} \boxdot \mathcal{P}(\boldsymbol{p}) = \begin{cases} \{\boldsymbol{p}\} & \text{if } \mathcal{U} \cap \mathcal{P}(\boldsymbol{p}) \neq \emptyset \text{ and } \boldsymbol{p} \in \mathcal{U} \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.14}$$

where $\mathcal{U}$ is a subset of a partition $\mathcal{P}$ of a continuous image, and $\mathcal{P}(\boldsymbol{p})$ is a neighborhood of $\boldsymbol{p}$.

Three isolate operators can be defined with the cross neighborhood, diagonal neighborhood and rectangular neighborhood. In this manner, **cross isolate operator** refers to the simple isolate operator with the cross neighborhood. The **diagonal isolate operator** and **rectangular isolate operator** are defined similarly.

I define a generalization of isolate operator as:

Definition 2.13: *The **k-isolate operator** is defined as*

$$\mathcal{U} \underset{k}{\boxdot} \mathcal{P}(\boldsymbol{p}) = \begin{cases} \{\boldsymbol{p}\} & \text{if } |\mathcal{U} \cap \mathcal{P}(\boldsymbol{p})| \geq k \text{ and } \boldsymbol{p} \in \mathcal{U}, \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.15}$$

where $\mathcal{U}$ is a subset of a partition $\mathcal{P}$ of a continuous image, and $\mathcal{P}(\boldsymbol{p})$ is a neighborhood of $\boldsymbol{p}$.

Remark 2.4: From the previous definitions: $\mathcal{U} \boxdot \mathcal{P}(\boldsymbol{p}) = \mathcal{U} \underset{1}{\boxdot} \mathcal{P}(\boldsymbol{p})$.

Applications like text recognition or fingerprint classification assume that the foreground consists of large connected components. These components are commonly larger than a particular rectangular area. Therefore, connected components which are completely contained in small rectangular neighborhoods may be removed from the foreground; see Fig. 2.2.

Definition 2.14: *The **frame isolate operator** for a pixel $\boldsymbol{p}$ is defined as*

$$\mathcal{P}(\boldsymbol{p}) \overset{\mathcal{U}}{\boxdot} \mathcal{P}'(\boldsymbol{p}) = \begin{cases} \{\boldsymbol{p}\} & \text{if } (\mathcal{P}(\boldsymbol{p}) \setminus \mathcal{P}'(\boldsymbol{p})) \cap \mathcal{U} \neq \emptyset \text{ and } \boldsymbol{p} \in \mathcal{U} \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.16}$$

where $\mathcal{U}$ is a subset of a partition $\mathcal{P}$ of a continuous image, and both $\mathcal{P}(\boldsymbol{p})$ and $\mathcal{P}'(\boldsymbol{p})$ are neighborhoods of $\boldsymbol{p}$.

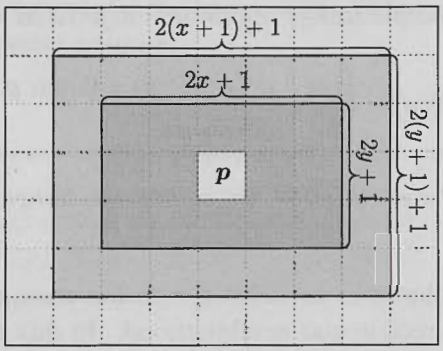


Figure 2.2 – *In this example, the gray areas denote two rectangular neighborhoods $\mathcal{P}_{y,x}(\boldsymbol{p})$ and $\mathcal{P}_{y+1,x+1}(\boldsymbol{p})$. Then, given a set $\mathcal{U}$ and these two neighborhoods, the frame isolate operator returns the empty set if none of the pixels within light-gray area belongs to $\mathcal{U}$.*

Remark 2.5: Since the cardinality of $\mathcal{P}_{y,x}(\boldsymbol{p}) \cap \mathcal{U}$ can be computed in constant time with integral images (Section A), an efficient implementation of

$$\mathcal{P}_{u,v}(\boldsymbol{p}) \overset{\mathcal{U}}{\boxdot} \mathcal{P}_{y,x}(\boldsymbol{p}) \tag{2.17}$$

can be done by comparing $|\mathcal{P}_{y,x}(\boldsymbol{p}) \cap \mathcal{U}|$ with $|\mathcal{P}_{u,v}(\boldsymbol{p}) \cap \mathcal{U}|$.

2.3.2 Expansion operators

Expansion operators add pixels to subsets of $\mathcal{P}$ unlike isolate operators, which remove pixels from a subset of $\mathcal{P}$. Section 5.3 introduces several operators by considering two particular subsets of $\mathcal{P}$. Then, I introduce the generalization of these operators.

Definition 2.15: *The **incidence operator** is defined as*

$$\mathcal{U} \underset{u,v}{\overset{\mathcal{P}(\boldsymbol{p})}{\bowtie}} \mathcal{V} = \begin{cases} \{\boldsymbol{p}\} & \text{if } |\mathcal{U} \cap \mathcal{P}(\boldsymbol{p})| \geq u, \text{ and } |\mathcal{V} \cap \mathcal{P}(\boldsymbol{p})| \geq v \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.18}$$

where $\mathcal{U}$ and $\mathcal{V}$ are two subsets of a partition $\mathcal{P}$ of a continuous image, u and v are two positive integers, and $\mathcal{P}(\boldsymbol{p})$ denotes a neighborhood of $\boldsymbol{p}$.

Definition 2.16: *The **simple expansion operator** is defined as*

$$\mathcal{U} \underset{u,v}{\overset{\mathcal{P}(p)}{\trianglerighteq}} \mathcal{V} = \begin{cases} \{p\} & \text{if } \boldsymbol{p} \notin \mathcal{U}, \mathcal{V}, \text{ and } |\mathcal{P}(\boldsymbol{p}) \cap \mathcal{U}| \geq u, \text{ and } |\mathcal{P}(\boldsymbol{p}) \cap \mathcal{V}| \leq v \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.19}$$

where $\mathcal{U}$ and $\mathcal{V}$ are two subsets of a partition $\mathcal{P}$ of a continues image, u and v are two positive integers, and $\mathcal{P}(\boldsymbol{p})$ denotes a neighborhood of $\boldsymbol{p}$.

2.4 Summary

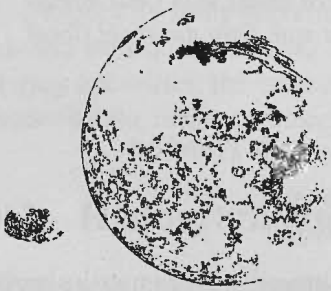
In this chapter, I introduced and formalized preliminary concepts of digital images: pixels, color images, gray images, neighborhoods, and morphological operator.

In Section 2.1, a pixel is defined as a region of a two-dimensional image (Definition 2.2), while a digital image is defined as a function from its set of pixels to $\mathbb{Z}^d$ (Definition 2.4). In the same section, I also stated a specific transformation from color to gray intensities, which is used throughout this thesis; see (2.7).

The concept and notation of both neighbors and neighborhoods were briefly given in Section 2.2. In Section 2.3.1, I proposed three morphological operators: the frame isolate operator (Definition 2.14), which removes pixels from a set; the incidence operator (Definition 2.15), which adds pixels to the set; and the simple expansion operator (Definition 2.16), which also adds pixels to a set. All three operators are efficiently computed through integral images; see Appendix A. The capability of these operators to restore sets will be shown in Section 5.3 and Chapter 7.

Chapter 3

Survey of binarization algorithms



Mars

The purpose of this chapter is to examine the local implementation of several binarization algorithms which are either related to my method or because they are considered reference algorithms in the binarization literature. They demand a small number of assumptions and are straightforward to implement. Even so, they provide an excellent basis to produce sophisticated methods which incorporate a priori information about the objects to be recognized.

3.1 Preliminaries

binarization

The task of identifying pixels with relevant information is formally known as **binarization**. It divides the set of pixels $\mathcal{P}$ into two sets $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ with the aim of estimating the **foreground** $\mathcal{F}$ and **background** $\mathcal{B}$, where $\mathcal{F}$ represents the set of pixels containing the objects of interest and $\mathcal{B}$ represents the complement of $\mathcal{F}$ in $\mathcal{P}$.

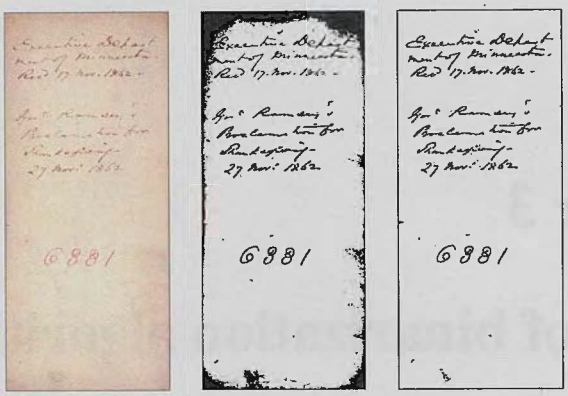


Figure 3.1 – *(Left) Original image. (Centre) Binarized image with Otsu threshold (global implementation). (Right) Binarized image with Otsu threshold (local implementation).*

Global thresholding computes a unique value $t_{opt} \in [0, g]$ and set

$$B(\boldsymbol{p}) = \begin{cases} 1 \text{ (foreground)} & \text{if } I(\boldsymbol{p}) \leq t_{opt} \\ 0 \text{ (background)} & \text{otherwise.} \end{cases} \tag{3.1}$$

However, these algorithms are unsuitable to binarize images with composite backgrounds and wide changes of illumination. Figure 3.1 (Centre), for instance, shows the binarized image of an image with a wide range of illumination. Otsu's threshold in its original implementation classifies background pixels at the image border as foreground pixels. This happens because these background pixels are darker than background pixels in the image centre. However, Fig. 3.1 (Right), computed with a local implementation of Otsu's threshold (Section 3.2.1), overcomes this problem.

Local thresholding computes a threshold surface

$$T : \{\mathcal{P}_r(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \rightarrow \{0, 1, \ldots, g\} \tag{3.2}$$

over the whole image, and sets

$$B(\boldsymbol{p}) = \begin{cases} 1 \text{ (foreground)} & \text{if } I(\boldsymbol{p}) \leq T(\boldsymbol{p}) \\ 0 \text{ (background)} & \text{otherwise.} \end{cases} \tag{3.3}$$

A local implementation of a global algorithm is such that the global analysis is restricted to $\mathcal{P}_r(\boldsymbol{p})$. Similarly, a secondary neighborhood supplies the "global information" to hybrid algorithms; see Fig. 3.2.

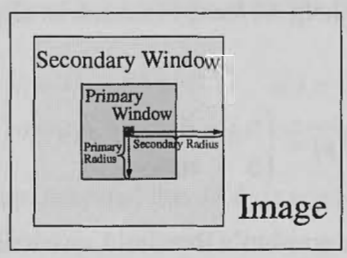


Figure 3.2 – *All algorithms gather the threshold information from a primary neighborhood, although the hybrid algorithms use a secondary neighborhood to compute any "global information".*

I especially study two kinds of binarization methods: histogram cluster methods and statistical methods. The former rely on information from the histogram of gray intensities; the latter rely on information from statistics of the gray intensities, like the mean, variance, third moment, maximum and minimum.

3.2 Histogram cluster binarization algorithms

Definition

Histogram cluster binarization algorithms assume that the foreground and background can be estimated by the cluster

$$\begin{aligned} \hat{\mathcal{F}} &= \left\{ \boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) \mid I(\boldsymbol{q}) \leq t_{opt} \right\} \text{ and} \\ \hat{\mathcal{B}} &= \left\{ \boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) \mid I(\boldsymbol{q}) > t_{opt} \right\}, \end{aligned} \tag{3.4}$$

respectively, where the optimal threshold $t_{opt} \in [0, g]$ satisfies the algorithm criterion optimally. Examples of methods to obtain t_{opt} include using entropy functions and mixture of two distributions, curvature analysis, and many more.

In images with composite background, t_{opt} may not exist such that $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ approximate $\mathcal{F}$ and $\mathcal{B}$ accurately. Therefore, its applicability may be restricted to neighborhoods where the method's assumptions are satisfied. Otherwise the local implementation of a histogram cluster method will systematically produce false positives due to neighborhoods which are completely contained in the background. To solve this problem, several techniques have been proposed by binarization researchers like Chou [13], Moghaddam and Cheriet [54], and Gupta [27]. Although the analysis of these techniques is beyond the scope of this thesis, I use a

simple restriction that may help all these binarization algorithms without favoring a particular algorithm.

$$T(\boldsymbol{p}) = \begin{cases} t_{opt} & \text{if } \hat{\mu}_{I,\hat{\mathcal{B}}} - \hat{\mu}_{I,\hat{\mathcal{F}}} < c \\ 0 & \text{otherwise,} \end{cases} \tag{3.5}$$

where t_{opt} is the optimal algorithm's threshold restricting the global analysis to $\mathcal{P}_r(\boldsymbol{p})$, and c depicts the minimum expected contrast between the foreground and background. I set $c = 15$ in all the experiments for my thesis, since the human eye can approximately distinguish contrast between two gray intensities that differ in 15 or more levels in gray images with 256 levels; see Gonzalez and Woods [26], chapter 2.

Notation:
I denote $H_{I,\mathcal{A}}$ as the **histogram of gray intensities** of a set $\mathcal{A}$. For instance, $H_{I,\mathcal{F}\cap\mathcal{P}_r(\boldsymbol{p})}$ denotes the histogram of gray intensities of foreground pixels within $\mathcal{P}_r(\boldsymbol{p})$. For the sake of brevity, I simplify $H_{I,\mathcal{P}_r(\boldsymbol{p})}(i)$ with h_i.

Readers may be interested in an efficient implementation to compute $H_{I,\mathcal{P}_r(\boldsymbol{p})}$ described in [72] and [87].

3.2.1 Otsu's algorithm and variants

Otsu's algorithm [66] is a global algorithm, which minimizes the sum of the variance of gray intensities of $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$. It assumes that the gray intensities of foreground and background form two distinguishable clusters whose overlap is small. The local Otsu's threshold uses the criterion

$$t_{opt} = \underset{t\in(0,g)}{\arg\max} \left\{ F_0(t) \cdot F_1(t) \cdot [\hat{\mu}_1(t) - \hat{\mu}_0(t)]^2 \right\}, \tag{3.6}$$

where

$$F_i(t) = \sum_{j=a}^{b} h_j \quad \text{and} \quad \hat{\mu}_i(t) = \frac{1}{F_i(t)} \sum_{j=a}^{b} j \cdot h_j \quad \text{for } i = 0, 1, \tag{3.7}$$

and the lower limit a and upper limit b depend on the index $i = 0, 1$. These limits are defined as:

$$a, b \begin{cases} a = 0 \text{ and } b = t & \text{if } i = 0, \\ a = t + 1 \text{ and } b = g & \text{otherwise.} \end{cases} \tag{3.8}$$

Liao et al. [45] have demonstrated that (3.6) is equivalent to

$$\begin{aligned} t_{opt} &= \arg\max_{t \in (0,g)} \left\{ F_0(t) \cdot \hat{\mu}_0^2(t) + F_1(t) \cdot \hat{\mu}_1^2(t) - \hat{\mu} \right\} \\ &= \arg\max_{t \in (0,g)} \left\{ F_0(t) \cdot \hat{\mu}_0^2(t) + F_1(t) \cdot \hat{\mu}_1^2(t) \right\}, \end{aligned} \tag{3.9}$$

Ng [59] derived **the valley-emphasis threshold** (Ng's algorithm) from (3.9) given by

$$t_{opt} = \arg\max_{t \in (0,g)} \left\{ [n - h_t] \left[F_0(t) \cdot \hat{\mu}_0^2(t) + F_1(t) \cdot \hat{\mu}_1^2(t) \right] \right\}. \tag{3.10}$$

Ng's algorithm attempts to ensure the selection of a value which lies at the valley or left bottom rim in the histogram of gray intensities.

Ng misinterpreted the valley-emphasis threshold as the application of a weight $(n - h_t)$ to the calculation of Otsu's threshold. Even though (3.6) is equivalent to (3.9), (3.10) is not equivalent to

$$\begin{aligned} t_{opt} &= \arg\max_{t \in (0,g)} \left\{ [n - h_t] \cdot F_0(t) \cdot F_1(t) \cdot [\hat{\mu}_1(t) - \hat{\mu}_0(t)]^2 \right\} \\ &= \arg\max_{t \in (0,g)} \left\{ [n - h_t] \left[F_0(t) \cdot \hat{\mu}_0^2(t) + F_1(t) \cdot \hat{\mu}_1^2(t) - \hat{\mu} \right] \right\}, \end{aligned} \tag{3.11}$$

which is indeed the application of the weight $(n - h_t)$ to (3.6).

Moghaddam and Cheriet [54], and Chou et al. [13] have proposed variants of Otsu's algorithm in order to discard outliers and detect neighborhoods that are completely contained in the background. For each pixel, **Moghaddam's method** compares the global Otsu's threshold (whole image) with a scaled Otsu's threshold which uses information from the pixel neighborhood. In contrast, **Chou's method** divides the image into regions and, furthermore, contextual rules decide whether the local Otsu's threshold restricted in the region of interest is applied or not. However, the analysis of such binarization methods is beyond the scope of this thesis.

3.2.2 Johannsen and Bille's algorithm

Johannsen and Bille's method (Johannsen' algorithm) [32] is a global algorithm, which minimizes the interdependence, in an information theoretic sense, between the gray intensities of the estimated foreground and background. The local Johannsen's algorithm chooses t_{opt} from the relation

$$t_{opt} = \underset{t\in(0,g)}{\arg\min} \{C_0(t) + C_1(t)\}, \tag{3.12}$$

$$C_0(t) = \ln\left(\sum_{j=0}^{t} p_j\right) - \frac{1}{\sum_{j=0}^{t} p_j}\left[p_t \cdot \ln(p_t) + \left[\sum_{j=0}^{t-1} p_j\right] \cdot \ln\left(\sum_{j=0}^{t-1} p_j\right)\right], \tag{3.13}$$

$$C_1(t) = \ln\left(\sum_{j=t}^{g} p_j\right) - \frac{1}{\sum_{j=t}^{g} p_j}\left[p_t \cdot \ln(p_t) + \left[\sum_{j=t+1}^{g} p_j\right] \cdot \ln\left(\sum_{j=t+1}^{g} p_j\right)\right], \tag{3.14}$$

where

$$p_j = \frac{h_j}{|\mathcal{P}_r(\boldsymbol{p})|} \tag{3.15}$$

denotes the empirical probability of the gray intensity at level j in $\mathcal{P}_r(\boldsymbol{p})$.

3.2.3 The minimum error thresholding

The **minimum error thresholding (Kittler's algorithm)** [40] is a global algorithm, which minimizes a criterion related to the average classification error rate assuming that the gray intensities of both background and foreground are normally distributed with different mean and variance. The local Kittler's threshold is computed as

$$t_{opt} = \underset{t\in(0,g)}{\arg\min} \left\{\sum_{i=0}^{1} F_i(t) \cdot \ln\left(\frac{\hat{\sigma}_i^2(t)}{[F_i(t)]^2}\right)\right\}, \tag{3.16}$$

where

$$\hat{\sigma}_i^2(t) = \frac{1}{F_i(t)}\left[\sum_{j=a}^{b} j^2 \cdot h_j\right] - [\hat{\mu}_i(t)]^2, \tag{3.17}$$

and $F_i(t)$, $\hat{\mu}_i(t)$, a, and b are defined as in Otsu's threshold.

Cho et al. [12] argue that Kittler's algorithm models the gray intensities of both foreground and background as normally distributed, but the parameters of such distributions are estimated with bias. Indeed, $\mu_{I,\mathcal{F}_r(p)}$ and $\sigma^2_{I,\mathcal{F}_r(p)}$ are $M_0(t)$ and $\sigma_0^2(t)$ which come from a distribution whose tails are truncated by the threshold. However, this bias becomes noticeable only when the histogram of gray intensities shows vague bimodality.

3.2.4 Kapur, Sahoo and Wong's algorithm

Kapur, Sahoo and Wong's algorithm (Kapur's algorithm) [35] is a global algorithm, which maximizes the sum of the entropy of gray intensities in $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$. The local optimal threshold is derived as

$$t_{opt} = \underset{t\in(0,g)}{\arg\min} \left\{ -\sum_{i=0}^{1} \sum_{j=a}^{b} \left[\frac{h_j}{F_i(t)} \cdot \ln\left(\frac{h_j}{F_i(t)} \right) \right] \right\}, \tag{3.18}$$

and $F_i(t)$, a, and b are defined as in Otsu's threshold.

3.2.5 Tsallis entropy's algorithm

Tsallis entropy's algorithm (Portes's algorithm) [67] is a global algorithm proposed by Portes de Albuquerque, which maximizes the information measure between background and foreground. Locally, it derives the optimal threshold from **Tsallis entropy** [88] as

$$t_{opt} = \underset{t\in(0,g)}{\arg\max} \{C_0(t) + C_1(t) + (1-\alpha) \cdot C_0(t) \cdot C_1(t)\}, \tag{3.19}$$

where

$$C_i(t) = \frac{1 - \sum_{j=a}^{b} \left[\frac{h_j}{F_i(t)} \right]^{\alpha}}{\alpha - 1}, \tag{3.20}$$

where $F_i(t)$, a, and b are defined as in Otsu's threshold, and α is a parameter whose influence on the threshold was not determined in the original publication. Notice that Tsallis entropy reduces to **Boltzmann-Gibbs** entropy if $\alpha \to 1$. That is,

$$\lim_{\alpha\to 1} \frac{1 - \sum_i x_i^{\alpha}}{\alpha - 1} = -\sum_i x_i \cdot \ln x_i, \quad \text{where} \quad \sum_i x_i = 1. \tag{3.21}$$

Therefore, **Kapur's algorithm** is a particular case of Tsallis entropy's algorithm for $\alpha = 1$.

Tsallis entropy is also used by Mello and Schuler [51], and Mello et al. [50]. They proposed a linear combination

$$t_{opt} = \arg\max_{t\in(0,g)} \left\{ \sum_{i=0}^{1} w_i \cdot C_i(t) \right\} \tag{3.22}$$

where the weights w_i's were experimentally determined for each image type. However, the parameter space is enormous, considering the weights as parameters and the fact that the their ranges were not determined. Hence, this variant of Tsallis's entropy method was excluded from my experiments.

3.3 Statistical algorithms

Statistical algorithms rely on information from statistics of gray intensities. These algorithms usually compute the mean and variance of gray intensities in $\mathcal{P}_r(p)$. Both statistics are quickly computed with **integral images** [72], so statistical algorithms have the advantage of speed over histogram cluster algorithms.

Notation:
$\hat{\mu}$ and $\hat{\sigma}^2$ denote the estimators of the **mean** and **variance** of gray intensities in $\mathcal{P}_r(\boldsymbol{p})$, respectively.

3.3.1 Niblack's algorithm

Niblack's algorithm [61] is a local algorithm, which assumes that the gray intensities of the background form a dominant peak. Niblack's threshold is computed as

$$T(\boldsymbol{p}) = \hat{\mu} - \alpha \cdot \hat{\sigma}, \tag{3.23}$$

where α is a parameter which usually is greater than zero, the higher α, the lower $T(\boldsymbol{p})$. However, α could be negative if there is not a unique dominant peak or the dominant peak is mainly formed by foreground pixels in the histogram of gray intensities. Trier and Jain [87] suggested $\alpha = 0.2$.

3.3.2 Sauvola and Pietikäinen's algorithm

Sauvola and Pietikäinen's algorithm (Sauvola's algorithm) [80] is a local algorithm, which computes a threshold similar to Niblack's threshold, but it incorporates a second parameter $\beta > 0$,

$$T(\boldsymbol{p}) = \hat{\mu} - \alpha \cdot \hat{\mu} + \alpha \frac{\hat{\sigma}}{\beta} \hat{\mu}, \tag{3.24}$$

where α behaves as in Niblack's threshold. The influence of $\hat{\sigma}$ on $T(\boldsymbol{p})$ is regulated by β so that $T(\boldsymbol{p}) \to \hat{\mu} - \alpha \cdot \hat{\mu}$ if $\hat{\sigma} \to 0$; $T(\boldsymbol{p}) \to \hat{\mu}$ if $\hat{\sigma} \to \beta$. Neighborhoods that are completely contained in the background may have a low $\hat{\sigma}$, which implies that $T(\boldsymbol{p}) \approx \hat{\mu} - \alpha \cdot \hat{\mu}$ and, consequently, $I(\boldsymbol{p}) > T(\boldsymbol{p})$ with high probability.

Sauvola and Pietikäinen suggest $\alpha = 0.5$ and $\beta = 128$ assuming that $g = 255$.

3.3.3 Wolf and Jolion's algorithm

Wolf and Jolion's algorithm (Wolf's algorithm) [90] is a hybrid algorithm, which replaces the parameter β of Sauvola's algorithm with the maximum standard deviation of gray intensities of neighborhoods of radius r so that the influence of $\hat{\sigma}$ on $T(\boldsymbol{p})$ is normalized. It also replaces the mean of gray intensities in the last two terms of (3.24) with the difference between the mean and minimum of gray intensities in the neighborhood. Wolf and Jolian reflect thus the idea that the optimal threshold should lie between such an interval. Wolf's threshold is given by

$$\begin{aligned} T(\boldsymbol{p}) &= \hat{\mu} - \alpha\,[\hat{\mu} - m] + \alpha \frac{\hat{\sigma}}{s}\,[\hat{\mu} - m]\,, \\ m = \min_{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})} \{I(\boldsymbol{q})\}\,, &\qquad s = \max_{\boldsymbol{q} \in \mathcal{P}_{r^*}(\boldsymbol{p})} \left\{\hat{\sigma}_{I,\mathcal{P}_r(\boldsymbol{q})}\right\}, \end{aligned} \tag{3.25}$$

where $\mathcal{P}_{r^*}(\boldsymbol{p})$ is a secondary neighborhood of radius $r^* \geq r$ and $\alpha \leq 1$. The higher α, the lower $T(\boldsymbol{p})$. Wolf and Jolion suggest the parameter $\alpha = 0.5$.

3.3.4 Iterative global thresholding

The **iterative global thresholding** is a hybrid and iterative method, which was originally proposed in [36] and subsequently improved in [37].

In each iteration i, the gray intensities are linearly transformed from $[m, \mu_{\mathcal{P}}^{(i)}]$ to [0,g], where m and $\mu_{\mathcal{P}}^{(i)}$ are the minimum and mean of the gray intensities at the iteration i, respectively, setting gray intensities greater than $\mu_{\mathcal{P}}^{(i)}$ to g.

I propose **Kavallieratous's algorithm**, which is a variant of the iterative global thresholding. Instead of $\hat{\mu}_{\mathcal{P}}^{(i)}$, my modified version computes the mean of gray intensities in the pixel neighborhood of interest. Thereby

$$T(\boldsymbol{p}) = \begin{cases} I(\boldsymbol{p}) & \text{if } \hat{\mu}^{(\alpha)}(\boldsymbol{p}) > I^{(\alpha)}(\boldsymbol{p}) \\ 0 & \text{otherwise,} \end{cases} \tag{3.26}$$

where α is the number of iterations, $I^{(1)}(\boldsymbol{p}) = I(\boldsymbol{p}) - \min\limits_{\boldsymbol{q}\in\mathcal{P}_r(\boldsymbol{p})} \{I(\boldsymbol{q})\}$, and

$$\hat{\mu}^{(i)}(\boldsymbol{p}) = \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|} \sum_{\boldsymbol{q}\in\mathcal{P}_r(\boldsymbol{p})} I^{(i)}(\boldsymbol{q}) \quad \text{for } i = 1, \ldots, \alpha, \tag{3.27}$$

$$I^{(i)}(\boldsymbol{p}) = \min\left(\hat{\mu}^{(i-1)}(\boldsymbol{p}), g \cdot \frac{I^{(i-1)}(\boldsymbol{p})}{\hat{\mu}^{(i-1)}(\boldsymbol{p})}\right) \quad \text{for } i = 2, \ldots, \alpha \tag{3.28}$$

Chapter 4

Transition pixels

Happiness is like ideal situations, it appears only locally.

The main contribution of my thesis is enclosed in this chapter. I describe mathematically the concept of **t-transition pixel**. I originally proposed a rough notion of t-transition pixels, **transition functions** and related concepts in [69]. However, I formalized these ideas in [72]. Therefore, this chapter is an extended version of [72].

In order to structure my approach mathematically, I first introduce the term **ideal image** assuming that both foreground and background vary smoothly, exhibiting high contrast at the boundary. Later on, the concepts of t-transition pixel and t-transition set are used as an extension of edge pixels and the **edge set,** respectively.

Properties of transition pixels are analyzed in an ideal image, providing the mathematical bases for deriving discriminant functions, which I named transition functions. Each pixel is then associated with a **transition value** (varying from negative to positive) computed by a corresponding transition function.

I propose several transition functions. In particular, I state the conditions for which maxmin function is a transition function.

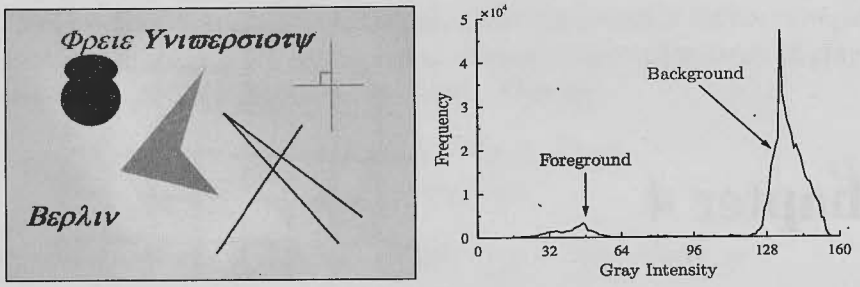


Figure 4.1 – *An ideal image (left) and its histogram of gray intensities (right).*

4.1 Ideal image

All binarization algorithms reported in and [82], [84], and [87] assume that foreground pixels can be distinguished by extracting diverse features based on their gray intensities. Under this assumption, authors like [42], [79], and [82] conjecture that both foreground and background should be uniform and homogeneous regions in terms of gray intensities (Fig. 4.1).

Although that conjecture is false for images with composite fore- or backgrounds like in Fig. 4.2, the gray intensities of both foreground and background appear to be approximately normally distributed in small neighborhoods of radius r; see Fig. 4.2. Hence, I propose characterizing the behavior of gray intensities locally.

Definition 4.1: *Given r, an image follows* ***Model 1*** *if the gray intensities of the foreground in all neighborhoods of radius r can be modeled as random variables which are independent and identically distributed (two different neighborhoods may follow different distributions). Gray intensities in the background are modeled in a similar manner.*

Notation:

Assuming Model 1, $\mu_{I,\mathcal{F}_r(\boldsymbol{p})}$ and $\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}$ denote the **mean** and **variance** of gray intensities in $\mathcal{F}_r(\boldsymbol{p})$, respectively. Analogously, $\mu_{I,\mathcal{B}_r(\boldsymbol{p})}$ and $\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}$ correspond to $\mathcal{B}_r(\boldsymbol{p})$. In general, the mean and variance of gray intensities in a set $\mathcal{A}$ is denoted by $\mu_{I,\mathcal{A}}$ and $\sigma^2_{I,\mathcal{A}}$.

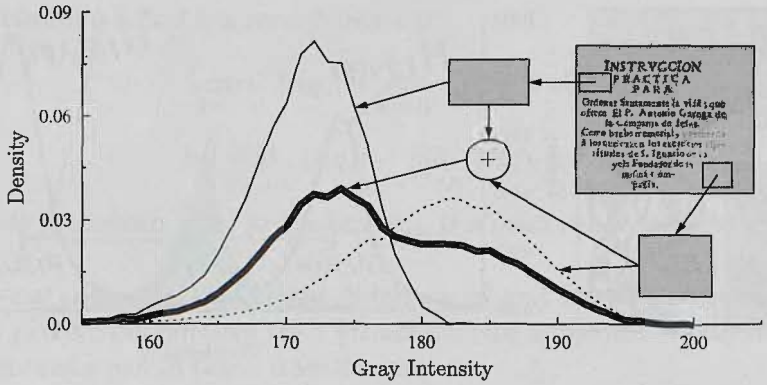


Figure 4.2 – *Two different regions form the background. Even though the gray intensities of the background are approximately normally distributed in each region, the gray intensities of the entire background are not.*

In my experience, historical documents fit Model 1 in a large percentage of neighborhoods if the background has no patterns deliberately printed.

Authors like Chow and Kaneko [14], and Kittler and Illingworth [40] pointed out that the gray intensities behave as (approximately) normally distributed and proposed a threshold based on them. The assumption of a priori distribution with Model 1 gives the mathematical basis to describe the behaviour of the histogram of gray intensities based on the probability density function. For my analysis, I assume that the gray intensities obey Model 1 and are normally distributed. Therefore, the histogram of gray intensities can be viewed as

$$H_{I,\mathcal{P}_r(p)}(i) \approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{I,\mathcal{F}_r(p)}, \sigma^2_{I,\mathcal{F}_r(p)}) + |\mathcal{B}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{I,\mathcal{B}_r(p)}, \sigma^2_{I,\mathcal{B}_r(p)}), \quad (4.1)$$

where $\phi(x; \mu, \sigma^2)$ denotes the probability density function of the normal distribution with mean μ and variance σ^2. However, I pointed out in [72] and [73] that the gray intensities are lognormally distributed rather than normally distributed. Thus,

$$H_{I,\mathcal{P}_r(p)}(i) \approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \lambda(i; \tilde{\mu}_{I,\mathcal{F}_r(p)}, \tilde{\sigma}^2_{I,\mathcal{F}_r(p)}) + |\mathcal{B}_r(\boldsymbol{p})| \cdot \lambda(i; \tilde{\mu}_{I,\mathcal{B}_r(p)}, \tilde{\sigma}^2_{I,\mathcal{B}_r(p)}), \quad (4.2)$$

where $\lambda\left(i; \tilde{\mu}, \tilde{\sigma}^2\right)$ denotes the lognormal probability density function with parameters $\tilde{\mu}$ and $\tilde{\sigma}^2$, which are the mean and variance of the variables natural logarithm, respectively.

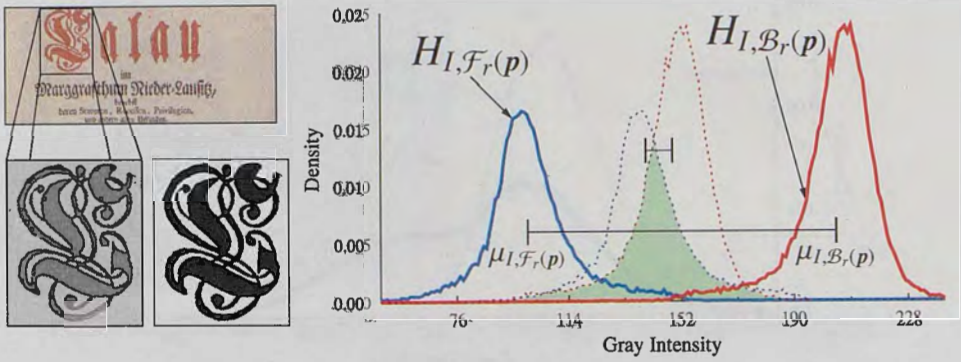


Figure 4.3 – *Example of* "good" *contrast in a neighborhood. In dotted lines, a hypothetical example of* "bad" *contrast.*

For both (4.1) and (4.2), the smoothness of the gray intensity surface depends then on the variance of gray intensities of both foreground and background. In addition to such variances, the contrast of gray intensity between the foreground and background in small neighborhoods also depends on the difference between their means of gray intensities. The closer $\mu_{\mathcal{F}_r(p)}$ to $\mu_{\mathcal{B}_r(p)}$, the higher the probability of misclassifying the pixel. The highlighted neighborhood in Fig. 4.3, for instance, has a "good" contrast because $\sigma_{\mathcal{F}_r(p)}$ and $\sigma_{\mathcal{B}_r(p)}$ are small and the difference between $\mu_{\mathcal{F}_r(p)}$ and $\mu_{\mathcal{B}_r(p)}$ is large. However, if this difference were small, as dash lines show, the probability of error is large (filled area), which may lead to misleading binarization based on gray intensities.

Figure 4.2 also shows that the larger the neighborhood, the larger the variance. In fact, the values of r for which (4.1) is true may be different for each pixel. Therefore, I postulated a generalization of (4.1) in diminutive neighborhoods, which does not depend on a particular statistical distribution.

Given $\mathcal{P}_s(\boldsymbol{p})$,

- **Foreground tendency**: Locally, the pixel's probability of being foreground increases when its gray intensity gets closer to zero. Conversely, the pixel's probability of being background tends to increase when its gray intensity gets closer to g.

- **Smoothness:** The difference of gray intensity between two pixels from the same set is close to zero in small neighborhoods $\mathcal{P}_s(\boldsymbol{p})$.

Definition 4.2: *I is a smooth image if*

$$\max_{q\in\mathcal{B}_s(p)}\{I(q)\} - \min_{q\in\mathcal{B}_s(p)}\{I(q)\} < d_{smo} \text{ and} \tag{4.3}$$

$$\max_{q\in\mathcal{F}_s(p)}\{I(q)\} - \min_{q\in\mathcal{F}_s(p)}\{I(q)\} < d_{smo} \tag{4.4}$$

with probability close to 1, where d_{smo} is a small value with respect to g.

- **Local contrast:** Locally, the difference of gray intensities between a pair of pixels from different set is greater than the difference of gray intensities between a pair of pixels from the same set.

  Definition 4.3: *I is an image with local contrast if*

$$\min_{q\in\mathcal{B}_s(p)}\{I(q)\} - \max_{q\in\mathcal{F}_s(p)}\{I(q)\} > d_{con} \tag{4.5}$$

  with probability close to 1, where d_{con} is a large positive number with respect to d_{smo}.

Remark 4.1: $\mathcal{P}_s(\boldsymbol{p})$ is a small neighborhood compared to $\mathcal{P}_r(\boldsymbol{p})$. Typically, $s < 5$ while r ranges between 15 and 100 [82], [84], [87].

The previous concepts can be expressed statistically. Let x_s be a random variable with mean μ_{x_s} and variance $\sigma^2_{x_s}$ representing the gray intensity difference between any pixels $\boldsymbol{q} \in \mathcal{B}_s(\boldsymbol{p})$ and the pixel of interest $\boldsymbol{p}$. A low mean μ_{x_s} combined with a small variance $\sigma^2_{x_s}$ represents a smooth **background surface.** Likewise, a smooth **foreground surface** is obtained when y_s has a low mean μ_{y_s} and a small variance $\sigma^2_{y_s}$, where y_s is analogously defined to x_s considering foreground pixels; see Fig. 4.4 (top).

A third random variable z_s represents the difference between the minimum gray intensity of the background and the maximum gray intensity of the foreground in the neighborhood of the pixel of interest. If z_s is negative, the histograms of both foreground and background are overlapped. Therefore, a misclassification may occur in any thresholding method (Fig. 4.4 (Bottom)). For example, the image in Fig. 4.5 contains a dark and a bright area. As a result, the histogram of gray intensities of both foreground and background are bimodal. The first and second modes of foreground (background) histogram are formed by gray intensities in the light area and dark area, respectively. Misclassifications stem from an overlapping of the second foreground peak with the background modes. Thus, any

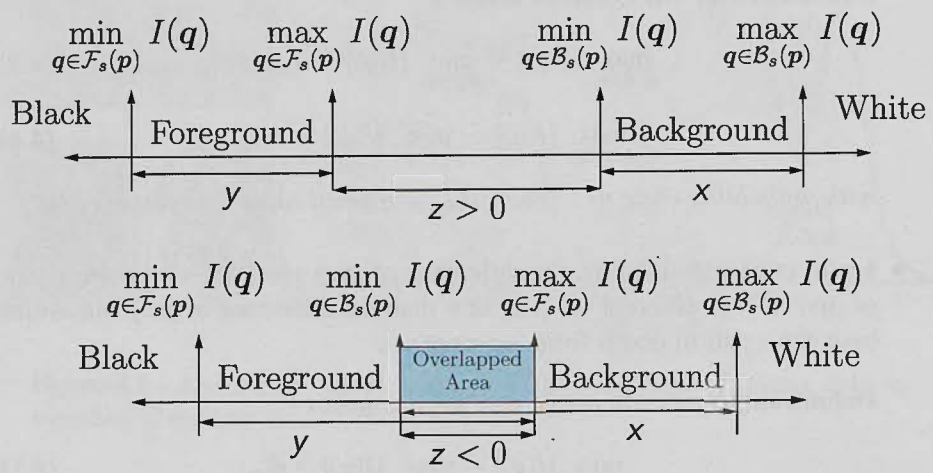


Figure 4.4 – *(Top) Representation of the random variables x , y and z into $\mathcal{P}_s$. (Bottom) A neighborhood $\mathcal{P}r$ where z is negative.*

thresholding method misclassifies either the pixels in the first background mode or the pixels in the second foreground mode Fig. 4.6.

Notation:
We will refer to x_s, y_s and z_s as the random variable of **background differences, foreground differences,** and **contrast differences,** respectively. From now on, we will omit the sub-index s in such random variables.

4.2 Transition pixel and transition set

In the following paragraph, the definition of t-transition pixel is stated as it was first introduced in [72].

Definition 4.4: *In a rectangular partition $\mathcal{P}$ of an image, a pixel $\mathbf{p}$ is a **t-transition pixel** if there exist $\mathbf{q}, \mathbf{q}' \in \mathcal{P}_I(\mathbf{p})$ such that $\mathbf{q} \in \mathcal{F}$ and $\mathbf{q}' \in \mathcal{B}$.*

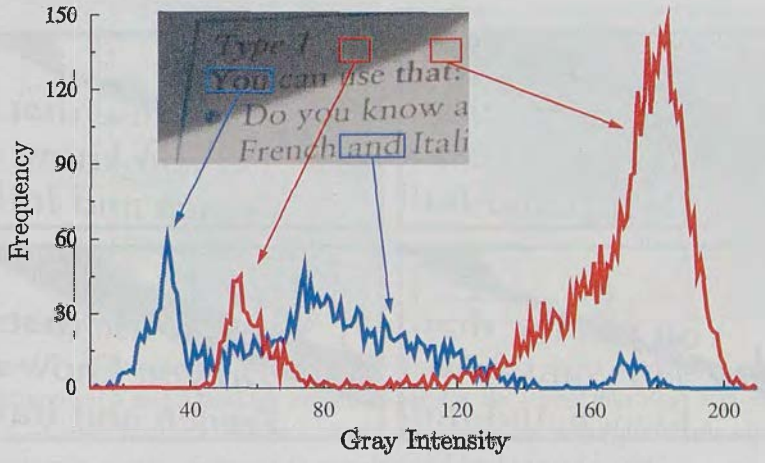


Figure 4.5 – *Histograms of gray intensities of both foreground and background are separately drawn.*

I formulated a generalization of transition pixel as:

Definition 4.5: *Let $\mathcal{P}(\boldsymbol{p})$ be the neighborhood associated with the pixel $\boldsymbol{p}$ in a partition $\mathcal{P}$ of an image, and*

$$\mathcal{P}^* = \{\mathcal{P}(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \tag{4.6}$$

*be the set of all neighborhoods; a pixel $\boldsymbol{p} \in \mathcal{P}$ is a **transition pixel** in $\mathcal{P}^*$ if there exist $\boldsymbol{q}, \boldsymbol{q}' \in \mathcal{P}(\boldsymbol{p})$ such that $\boldsymbol{q} \in \mathcal{F}$ and $\boldsymbol{q}' \in \mathcal{B}$.*

The set of t-transition pixels is denoted ${}_t\mathcal{P}$. This set extends along the whole foreground contour. In particular, a t-transition pixel is an edge pixel if $t = 1$.

A neighborhood that contains a dense subset of ${}_t\mathcal{P}$ also contains a significant subset of the **foreground contour**. Furthermore, the statistical distribution of ${}_t\mathcal{F} = \mathcal{F} \cap {}_t\mathcal{P}$ approximates the distributions of $\mathcal{F}$. since it is a large foreground sample. Analogously, the distribution of ${}_t\mathcal{B} = \mathcal{B} \cap {}_t\mathcal{P}$ approximates the distributions of $\mathcal{B}$[1].

Four neighborhood types help to characterize the transition pixels. Figure 4.7 shows neighborhoods of type 1 ($\mathcal{NT}1$) which have only background pixels. Neigh-

[1]The analysis of the sampling bias is beyond the scope of this thesis. Readers interested in pursuing the topic further are encouraged to consult the books by [10], and Kay [39] for a more thorough explanation. Useful discussion is also available in [16] and [29].

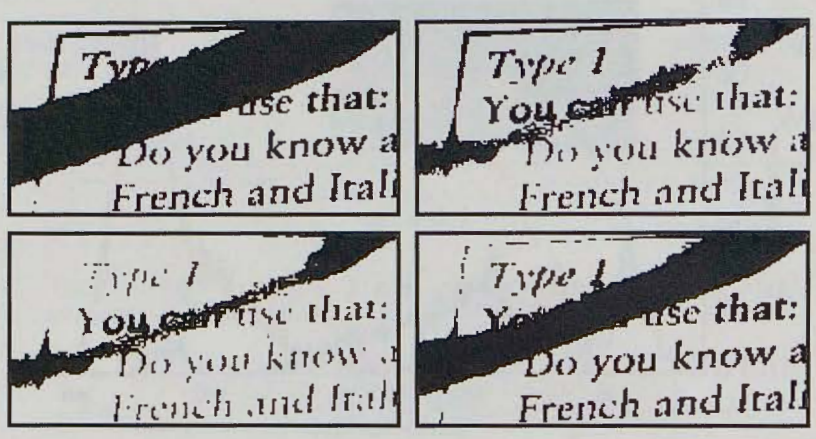


Figure 4.6 – *Binary images from the example in Fig. 4.5. All binarized imaged were computed with neighborhoods of radius* $r = 30$. *At the top,* ***Otsu's method*** *(left) with contrast* $c = 15$ *and* ***Kavallieratou's method*** *(right) with parameters* $\alpha = 5$;. *On the bottom,* ***Sauvola's method*** *(left) with parameters* $\alpha = 0.5$ *and* $\beta = 128$. ***Wolf's method*** *(right) with parameters* $\alpha = 0.5$.

borhoods of type 2 ($\mathcal{NT}2$) have their central pixels in the background and have foreground pixels. Conversely, $\mathcal{NT}3$ and $\mathcal{NT}4$ correspond to $\mathcal{NT}2$ and $\mathcal{NT}1$, respectively. Formally,

Definition 4.6: *A neighborhood* $\mathcal{P}_r(\boldsymbol{p}) \in \mathcal{NT}1$ *if for any* $\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) \Rightarrow \boldsymbol{q} \in \mathcal{B}$.

Definition 4.7: *A neighborhood* $\mathcal{P}_r(\boldsymbol{p}) \in \mathcal{NT}2$ *if* $\boldsymbol{p} \in \mathcal{B}$ *and there exists* $\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})$ *such that* $\boldsymbol{q} \in \mathcal{F}$.

Definition 4.8: *A neighborhood* $\mathcal{P}_r(\boldsymbol{p}) \in \mathcal{NT}3$ *if* $\boldsymbol{p} \in \mathcal{F}$ *and there exists* $\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})$ *such that* $\boldsymbol{q} \in \mathcal{B}$.

Definition 4.9: *A neighborhood* $\mathcal{P}_r(\boldsymbol{p}) \in \mathcal{NT}4$ *if for any* $\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) \Rightarrow \boldsymbol{q} \in \mathcal{F}$.

The most outstanding feature of transition pixels is easy to appreciate in a binary image. The difference of binary values between two pixels within neighborhoods type 1 or type 4 is always zero because the binary value of both pixels are the same, either both one, or both zero:

$$\mathcal{P}_t(\boldsymbol{p}) \in \mathcal{NT}1 \cup \mathcal{NT}4 \Rightarrow B(\boldsymbol{p}) - B(\boldsymbol{q}) = 0 \; \forall \, \boldsymbol{q} \in \mathcal{P}_t(\boldsymbol{p}). \tag{4.7}$$

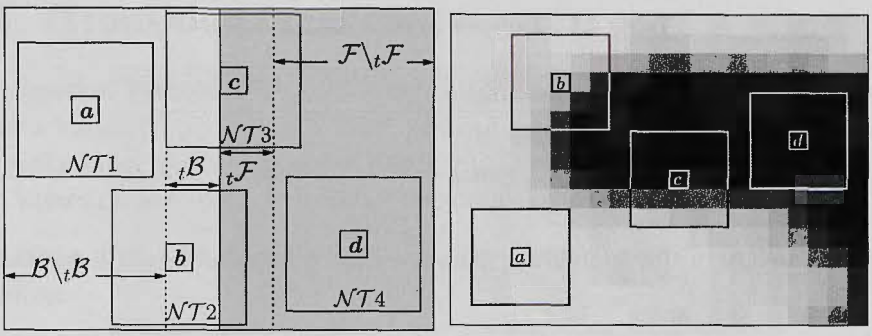


Figure 4.7 – *The schemes exemplify two no-transition pixels* $\boldsymbol{a}$ *and* $\boldsymbol{d}$*, a negative transition pixel* $\boldsymbol{b}$ *and a positive transition pixel* $\boldsymbol{c}$*. We expect that* $V(\boldsymbol{a}) \approx 0 \approx V(\boldsymbol{d})$*,* $V(\boldsymbol{b}) \leq -t_-$ *and* $V(\boldsymbol{c}) \geq t_+$*, where* t_- *and* t_+ *are approximately equal to the expected gray-intensity contrast between foreground and background pixels.*

On the other hand, neighborhoods type 2 and 3 contain, besides pairs of pixels whose difference of binary values is 0, pairs of pixels whose difference of binary values is either -1, or 1. The value -1 is reached in neighborhoods type 2 when the central pixel is compared with a foreground pixel in terms of binary value:

$$\mathcal{P}_t(\boldsymbol{p}) \in \mathcal{NT}2 \Rightarrow \exists \boldsymbol{q} \in \mathcal{F}_r(\boldsymbol{p}) \text{ such that } B(\boldsymbol{p}) - B(\boldsymbol{q}) = -1. \tag{4.8}$$

Conversaly, 1 is reached in neighborhoods type 3 when the central pixel is compared with a background pixel in terms of binary value:

$$\mathcal{P}_t(\boldsymbol{p}) \in \mathcal{NT}3 \Rightarrow \exists \boldsymbol{q} \in \mathcal{B}_r(\boldsymbol{p}) \text{ such that } B(\boldsymbol{p}) - B(\boldsymbol{q}) = 1. \tag{4.9}$$

Extending the above argument to non-binary but ideal images, neighborhoods type 1 and 4 have differences close to zero, unlike neighborhoods type 2 and 3, in which there are pairs of pixels whose difference of gray intensities is large in absolute magnitude.

Figure 4.7 exemplifies two transition pixels $\boldsymbol{b}$ and $\boldsymbol{c}$. Without considering outliers, we expect that

$$V(\boldsymbol{c}) \approx \max_{\boldsymbol{p} \in \mathcal{N}_t(\boldsymbol{c})} \{I(\boldsymbol{p})\} - I(\boldsymbol{c}) > d_{con} \tag{4.10}$$

and

$$\min_{\boldsymbol{p} \in \mathcal{N}_t(\boldsymbol{c})} \{I(\boldsymbol{p})\} - I(\boldsymbol{c}) \approx d_{smo} \approx 0 \tag{4.11}$$

Table 4.1 – *Differences in an ideal image.*

Difference	$NT1$	$NT2$	$NT3$	$NT4$
$\max_{q \in \mathcal{P}_s(p)} \{I(q)\} - I(p)$	$< d_{smo}$	$< d_{smo}$	$> d_{con}$	$< d_{smo}$
$I(p) - \min_{q \in \mathcal{P}_s(p)} \{I(q)\}$	$< d_{smo}$	$> d_{con}$	$< d_{smo}$	$< d_{smo}$

because the pixel with minimum gray intensity in $N_t(c)$ is foreground and has to be similar to the gray intensity of c (foreground smoothness). Moreover,

$$\max_{p \in N_t(c)} \{I(p)\} - I(c) \tag{4.12}$$

has to be higher than the minimum contrast expected in the image because it is the difference of gray intensities between foreground and background pixels. On the contrary, $V(d) \approx 0$ because both maximum and minimum gray intensities within $N_t(d)$ have to be similar to d.

Table 4.1 is constructed taking d_{smo} and d_{con} from the ideal image definitions and the fact that the pixel with the maximum gray intensity in neighborhoods type 2 is a background pixel while the pixel with minimum gray intensity in neighborhoods type 3 is a foreground pixel.

Notation:

I denote with ${}_t\mathcal{P}_r(p)$ the subset of t-transition pixels in $\mathcal{P}_r(p)$. Following the notation:

$${}_t\mathcal{F}_r(p) = {}_t\mathcal{P} \cap \mathcal{F} \cap \mathcal{P}_r(p) = {}_t\mathcal{P} \cap \mathcal{F}_r(p) \tag{4.13}$$

and

$${}_t\mathcal{B}_r(p) = {}_t\mathcal{P} \cap \mathcal{B} \cap \mathcal{P}_r(p) = {}_t\mathcal{P} \cap \mathcal{B}_r(p). \tag{4.14}$$

4.3 Transition function

A **transition function** F is a discriminant function taking extreme values only when a transition pixel is evaluated: positive for foreground pixels and negative for background pixels. Moreover, pixels in ${}_t\mathcal{P}^c$ (complement set of transition set) take values close to zero. In terms of conditional probabilities:

Definition 4.10: *A function F is a transition function if it satisfies the following relations:*

$$\Pr(\boldsymbol{p} \in {}_t\mathcal{F} \mid F(\boldsymbol{p}) \geq t_+) \;>\; 1-\varepsilon_+, \tag{4.15}$$

$$\Pr(\boldsymbol{p} \in {}_t\mathcal{B} \mid F(\boldsymbol{p}) \leq -t_-) \;>\; 1-\varepsilon_-, \tag{4.16}$$

$$\Pr(\boldsymbol{p} \in {}_t\mathcal{P}^c \mid -t_- < F(\boldsymbol{p}) < t_+) \;\approx\; 1-\varepsilon, \tag{4.17}$$

where $\varepsilon_+, \varepsilon_-, \varepsilon < 0.5$.

Definition 4.10 restricts ε_+, ε_-, and ε to $[0, 0.5)$, but the closer they are to zero, the better. Equations (4.15) and (4.16) mean $\boldsymbol{p}$ is pre-classified as foreground when $F(\boldsymbol{p})$ is greater than t_+ while $\boldsymbol{p}$ is pre-classified as background when $F(\boldsymbol{p})$ is lower than $-t_-$. Note that there is no information to pre-classify $\boldsymbol{p}$ if $-t_- < F(\boldsymbol{p}) < t_+$.

I suggested in [72] some functions to measure a transition value:

Maxmin

$$V(\boldsymbol{p}) = \max_{\boldsymbol{q}\in\mathcal{P}_s(\boldsymbol{p})} \{I(\boldsymbol{q})\} + \min_{\boldsymbol{q}\in\mathcal{P}_s(\boldsymbol{p})} \{I(\boldsymbol{q})\} - 2I(\boldsymbol{p}). \tag{4.18}$$

Discrete Laplace

$$L(\boldsymbol{p}_{i,j}) = \frac{1}{4}\left[I(\boldsymbol{p}_{i-1,j}) + I(\boldsymbol{p}_{i+1,j}) + I(\boldsymbol{p}_{i,j-1}) + I(\boldsymbol{p}_{i,j+1})\right] - I(\boldsymbol{p}_{i,j}). \tag{4.19}$$

Linear kernel

$$G(p) = \left[\sum_{\boldsymbol{q}\in\mathcal{P}_t(\boldsymbol{p})} w(\boldsymbol{q}) \cdot I(\boldsymbol{q})\right] - I(\boldsymbol{p}) \tag{4.20}$$

where

$$\sum_{\boldsymbol{q}\in\mathcal{P}_t(\boldsymbol{p})} w(\boldsymbol{q}) = 1. \tag{4.21}$$

Remark 4.2: Notice that

$$V: \; \{\mathcal{P}_s(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \;\rightarrow [-g, g], \tag{4.22}$$

$$L: \; \{\mathcal{P}_+(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \;\rightarrow [-g, g], \tag{4.23}$$

$$G: \; \{\mathcal{P}_s(\boldsymbol{p}) \mid \boldsymbol{p} \in \mathcal{P}\} \;\rightarrow [-g, g], \tag{4.24}$$

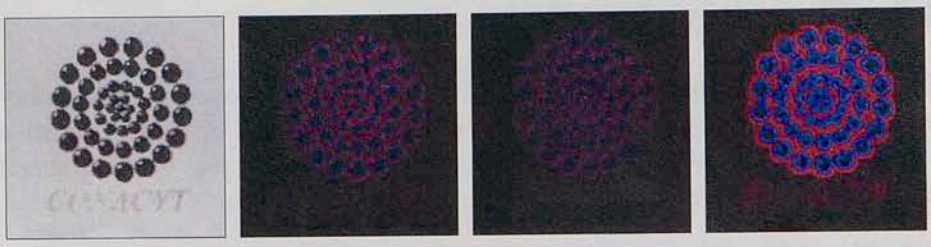


Figure 4.8 – *On the left, original image. In the center-left, $G(\boldsymbol{p})$ with Gaussian weights ($\sigma^2 = 1$ in $\mathcal{P}_2(\boldsymbol{p})$). In the center-right, Laplace operator. On the right* maxmin *with neighborhoods of radius 2.*

Table 4.2 – *Lower and upper bounds of* maxmin *function according the neighborhood type.*

Neighborhood	Bounds
$\mathcal{NT}1$	$-d_{smo} < V(\boldsymbol{p}) < d_{smo}$
$\mathcal{NT}2$	$V(\boldsymbol{p}) < -d_{con} + d_{smo}$
$\mathcal{NT}3$	$V(\boldsymbol{p}) > d_{con} - d_{smo}$
$\mathcal{NT}4$	$-d_{smo} < V(\boldsymbol{p}) < d_{smo}$

Figure 4.8 shows 3 images, each of which was computed with a different transition functions. Pixels with negative transition values are shown in red, a pixel with a $-x$ transition value is associated with a x-red intensity. The pixels with positive transition values are shown in blue.

4.4 Maxmin function

Table 4.2 was derived from Table 4.1. This table indicates that maxmin function is a transition function in ideal images, where t_+ and t_- (Definition 4.10) correspond to d_{con} and $-d_{con}$. However, Theorem 4.1 extends this result to gray images. Figure 4.9, for instance, shows how the histogram of transition values is constituted; in this example, (4.15) and (4.16) are satisfied with $t_+ = 10$ and $t_- = 10$.

Theorem 4.1. *Given a gray image I, suppose that their random variables of background differences x, foreground differences y and contrast differences z are ap-*

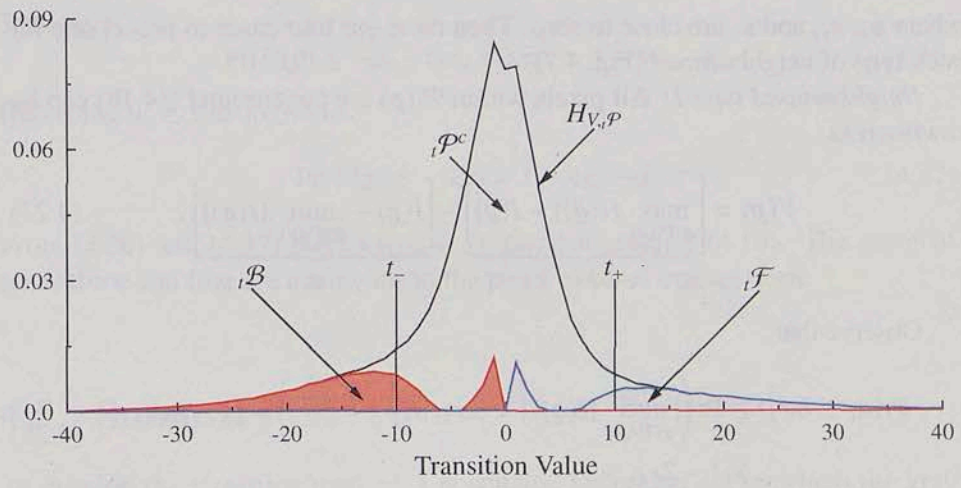


Figure 4.9 – *Histogram of transition values calculated by* maxmin *function with neighborhoods of radius 2.*

proximately Gaussian distributed in $\mathcal{P}_s(\boldsymbol{p})$ *such that* $\mu_z > 15\sigma$, *where*

$$\sigma = \max\left\{\sigma_x, \sigma_y, \sigma_z\right\}. \tag{4.25}$$

Then maxmin *function is a transition function in neighborhoods of radius* $t \leq s$.

Proof. To prove the theorem is sufficient to find t_- and t_+ such that

- $\Pr(V(\boldsymbol{p}) < -t_-) \approx 1$ if $\boldsymbol{p} \in \mathcal{NT}2$,
- $\Pr(V(\boldsymbol{p}) > -t_-) \approx 1$ if $\boldsymbol{p} \in (\mathcal{NT}2)^c$,
- $\Pr(V(\boldsymbol{p}) > t_+) \approx 1$ if $\boldsymbol{p} \in \mathcal{NT}3$ and
- $\Pr(V(\boldsymbol{p}) < t_+) \approx 1$ if $\boldsymbol{p} \in (\mathcal{NT}3)^c$.

where $(\mathcal{NT}i)^c$ represents pixels in all type of neighborhoods, except neighborhood of type $\mathcal{NT}i$.

We know that practically all the observations drawn from x are within $(\mu_x - 3\sigma_x, \mu_x + 3\sigma_x)$. Explicitly:

$$\begin{aligned} &\Pr(-3\sigma_x < x < 3\sigma_x) = 1 - \varepsilon_x, \\ &\Pr(-3\sigma_y < y < 3\sigma_y) = 1 - \varepsilon_y, \text{ and} \\ &\Pr(-3\sigma_z < z < 3\sigma_z) = 1 - \varepsilon_z, \end{aligned} \tag{4.26}$$

where ε_x, ε_y, and ε_z are close to zero. Then there are four cases to prove, one for each type of neighborhood (Fig. 4.7).

Neighborhood type 1: All pixels within $\mathcal{P}_t(\boldsymbol{p})$ are background. (4.18) can be rewritten as

$$V(\boldsymbol{p}) = \underbrace{\left[\max_{\boldsymbol{q}\in\mathcal{P}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\} - I(\boldsymbol{p})\right]}_{a_1} - \underbrace{\left[I(\boldsymbol{p}) - \min_{\boldsymbol{q}\in\mathcal{P}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\}\right]}_{a_2}. \tag{4.27}$$

Observe that

$$\begin{aligned}\Pr(a_1 \le 6\sigma_x) &\ge \Pr\left(\left|\max_{\boldsymbol{q}\in\mathcal{B}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\}\right| < 3\sigma_x, |I(\boldsymbol{p})| < 3\sigma_x\right) = [1-\varepsilon_x]^2 \\ &> 1 - 2\varepsilon_x,\end{aligned} \tag{4.28}$$

and

$$\Pr(a_2 \le 6\sigma_x) > 1 - 2\varepsilon_x. \tag{4.29}$$

Then

$$\Pr(-6\sigma_x \le V(\boldsymbol{p}) \le 6\sigma_x) \ge 1 - 4\varepsilon_x \tag{4.30}$$

Neighborhood type 2: There are both foreground and background pixels within $\mathcal{P}_t(\boldsymbol{p})$ and $\boldsymbol{p}$ is background. Regardless of outliers, we can assume that the pixel with the maximum gray intensity is background and the pixel with the minimum gray intensity is foreground. Rewriting (4.18) as:

$$\begin{aligned}V(\boldsymbol{p}) = &\underbrace{\left[\max_{\boldsymbol{q}\in\mathcal{B}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\} - I(\boldsymbol{p})\right]}_{a_1} - \underbrace{\left[I(\boldsymbol{p}) - \min_{\boldsymbol{q}\in\mathcal{B}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\}\right]}_{a_2} \\ &- \underbrace{\left[\min_{\boldsymbol{q}\in\mathcal{B}_t(\boldsymbol{p})}\{I\}(\boldsymbol{q}) - \max_{\boldsymbol{q}\in\mathcal{F}_t(\boldsymbol{p})}\{I\}(\boldsymbol{q})\right]}_{a_3} - \underbrace{\left[\max_{\boldsymbol{q}\in\mathcal{F}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\} - \min_{\boldsymbol{q}\in\mathcal{F}_t(\boldsymbol{p})}\{I(\boldsymbol{q})\}\right]}_{a_4}\end{aligned} \tag{4.31}$$

Thus,

$$\Pr(a_1 \le 6\sigma_x) > 1 - 4\varepsilon_x, \tag{4.32}$$

$$\Pr(a_2 \ge 0) = 1, \tag{4.33}$$

$$\Pr(a_3 \ge \mu_z - 3\sigma_z) > 1 - \varepsilon_z, \text{ and} \tag{4.34}$$

$$\Pr(a_4 \ge 0) = 1. \tag{4.35}$$

Therefore,

$$\Pr(V(\boldsymbol{p}) \leq -\mu_z + 6\sigma_x + 3\sigma_z) > 1 - 4\varepsilon_x - \varepsilon_z. \tag{4.36}$$

Replacing $\mu_z \geq 15\sigma$ in (4.36)

$$\Pr(V(\boldsymbol{p}) \leq -6\sigma) > 1 - 4\varepsilon_x - \varepsilon_z. \tag{4.37}$$

From (4.30) and (4.37) there exists $t_- \leq 6\sigma$ that satisfies (4.16). The proof of cases three and four are analogous to the proof of cases one and two.

□

4.5 Summary

To develop the transition method, I postulated Definition 4.1 in which the gray intensities are modeled in small neighborhoods as random variables (independent and identically distributed). The histogram of gray intensities is then modeled as a linear combination of the density functions of two normal distributions; as in [14] and [40]. However, I suggested the lognormal distribution as an alternative for the distribution of gray intensities; the strength of the lognormal model will be shown in Chapter 5 and Chapter 7.

I proposed three desirable properties that an ideal image must fulfill in binarization context: local tendency, local smoothness, and local contrast. In particular, local smoothness ensures an upper bound in the differences of gray intensities of two foreground (background) pixels; see Definition 4.2. Similarly, local contrast determines a lower bound between the differences of gray intensities of a foreground pixel and a background pixel; see Definition 4.3. Afterward, I statistically expressed these bounds for non ideal images with three random variables: background differences, foreground differences, and contrast differences.

The concept of t-transition pixel introduced in Definition 4.4 and Definition 4.5 is the first main contribution of my thesis. A pixel is a t-transition pixel if its neighborhood contains foreground and background pixels. Subsequently, the transition set (set of transition pixels) is divided into two subsets: positive transition set (intersection between foreground and transition set) and negative transition set (intersection between background and transition set).

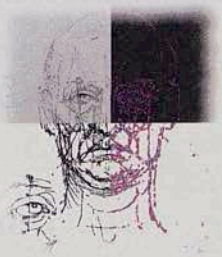
Later on, transition pixel's properties are analyzed in binary images and in ideal images, providing the mathematical foundations for deriving discriminant functions, which I named transition functions; see Section 4.3 and Definition 4.10. Transition functions are functions that take extreme values only when a transition

pixel is evaluated: positive for foreground pixels and negative for background pixels.

A minor contribution of this thesis is given in Section 4.4, where I proved that maxmin function is a transition function in ideal images. All transition values in further experiments are computed with this function using neighborhoods of radius 2.

Chapter 5

The transition method



Each life sparks changes of tone so gradually
that we believe we are in the same place.

The second main contribution of my thesis is enclosed in this chapter. I describe mathematically **the transition method** in gray images for binarization, and to a minor degree, for edge detection, and for detection of regions of interest.

The success of this novel approach depends on the definition of the **t-transition pixel,** previously defined in Chapter 4. In this chapter, I will show that the **positive transition set** (intersection of foreground and transition set) is approximated by the set of pixels with high positive transition values, and that the **negative transition set** (intersection of background and transition set) is approximated by the set of pixels with high negative transition values.

Several binarization methods based on the transition set are proposed. In addition to these binarization methods, I describe two simple methods for edge detection and detection of region of interest.

Even though the transition method has the potential to deal with uneven illumination, this chapter will focus only on images without sudden illumination changes in small neighborhoods.

5.1 Overview of the transition method

Figure 5.1 shows that the histogram of gray intensities of the highlighted neighborhood of radius r ($H_{I,\mathcal{P}_r(\boldsymbol{p})}$) is bimodal; the left peak of $H_{I,\mathcal{P}_r(\boldsymbol{p})}$ is mainly formed by foreground pixels, while the right peak is mainly formed by background pixels.

If we knew the **class-conditional density**

$$\Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in \mathcal{F}_r(\boldsymbol{p})) \text{ and } \Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in \mathcal{B}_r(\boldsymbol{p})), \tag{5.1}$$

we could consider the **maximum likelihood estimation** or **Bayesian estimation** approach to solve the binarization problem; see Fig. 5.2. Unfortunately, we rarely know the class-conditional densities. However, we can reasonably assume that the gray intensities of both foreground and background are approximately normally distributed; see (4.1). In consequence, $T(\boldsymbol{p})$ is quickly computed when there is an analytic intersection between

$$|\mathcal{F}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{I,\mathcal{F}_r(\boldsymbol{p})}, \sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}) \tag{5.2}$$

and the correspondent background function

$$|\mathcal{B}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{I,\mathcal{B}_r(\boldsymbol{p})}, \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}), \tag{5.3}$$

We can approximate $\Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in \mathcal{F}_r(\boldsymbol{p}))$ by drawing a representative sample of $\mathcal{F}_r(\boldsymbol{p})$; see Fig. 5.3. Since ${}_t\mathcal{F}_r(\boldsymbol{p})$ is a representative sample of $\mathcal{F}_r(\boldsymbol{p})$, the following equation holds in neighborhoods of radius r :

$$\Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in \mathcal{F}_r(\boldsymbol{p})) \approx \Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in {}_t\mathcal{F}_r(\boldsymbol{p})), \tag{5.4}$$

Although the transition sets are also unknown, my method provides ${}_t\hat{\mathcal{F}}_r(\boldsymbol{p})$, which is an accurate estimate of ${}_t\mathcal{F}_r(\boldsymbol{p})$, see Fig. 5.4. Thus, (5.4) changes to

$$\Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in \mathcal{F}_r(\boldsymbol{p})) \approx \Pr(I(\boldsymbol{q}) \mid \boldsymbol{q} \in {}_t\hat{\mathcal{F}}_r(\boldsymbol{p})). \tag{5.5}$$

We are now able to compute the gray threshold with the usual classification procedures. In Table 5.1, for instance, we computed a threshold with the **minimum symmetric values**; see Section 5.5.3.

The complete method consists of the following steps:

1. Compute the transition values for each pixel with a transition function. I suggest the maxmin function with neighborhoods of radius 2; see Fig. 5.5 (b).

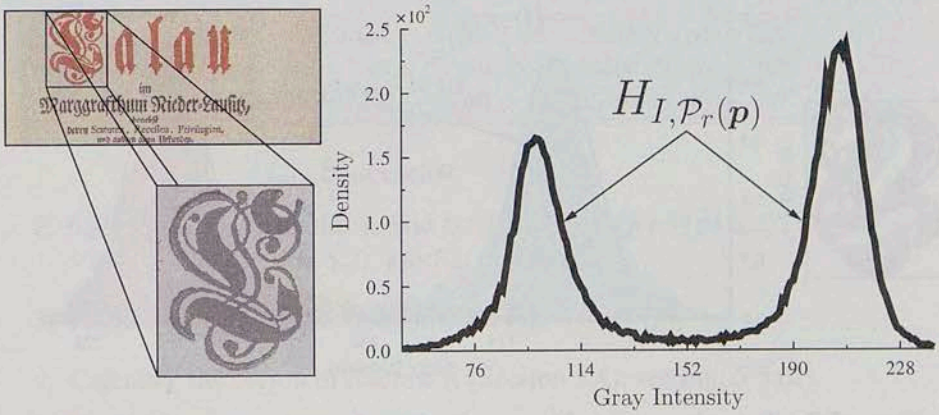


Figure 5.1 – *Histogram of gray intensities of the highlighted neighborhood.*

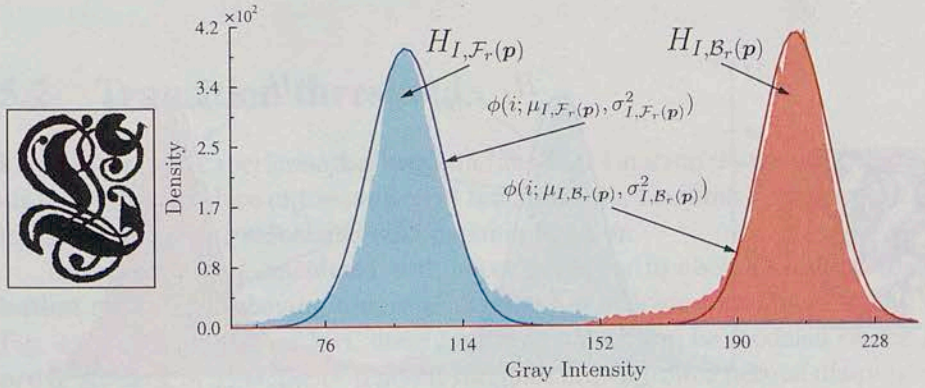


Figure 5.2 – *Binary ground truth, and histograms of gray intensities of both foreground and background. I have manually fitted a normal probability density distribution function to each histogram.*

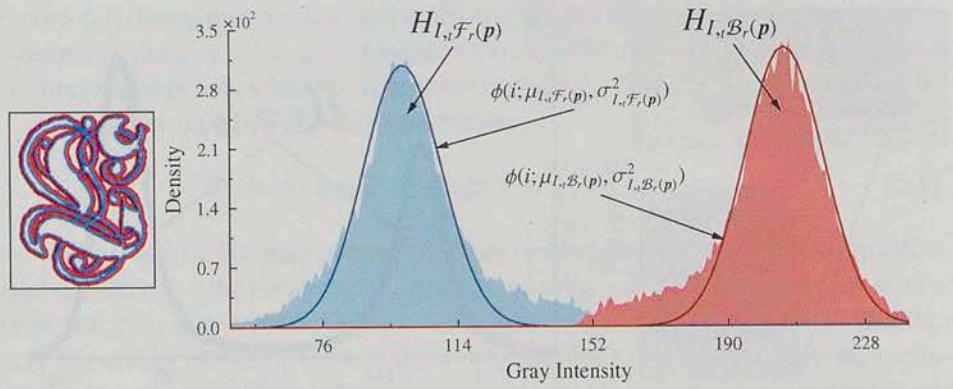


Figure 5.3 – *Transition sets. In blue, pixels within the positive transition set, and pixels within the negative transition set are shown in red. In $\mathcal{P}_r(\boldsymbol{p})$, the distribution of gray intensities in ${}_t\mathcal{F}_r(\boldsymbol{p})$ and ${}_t\mathcal{B}_r(\boldsymbol{p})$ approximate the distribution of gray intensities in $\mathcal{F}$ and $\mathcal{B}$, respectively.*

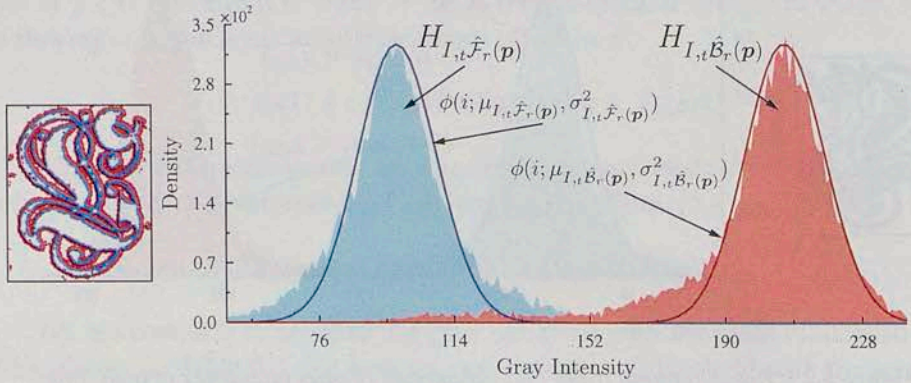


Figure 5.4 – *Approximation of the transition sets. We use the approximation of positive and negative transition sets as foreground and background samples.*

Table 5.1 – *Estimated threshold by minimum symmetric value with* $k = 10$.

	a	b	$T(\boldsymbol{p}) = \frac{a+b}{2}$
Ground truth	$\mathrm{SM}(k, \mathcal{F}_r(\boldsymbol{p})) = 98$	$\mathrm{SM}(k, \mathcal{B}_r(\boldsymbol{p})) = 207$	152.5
Transition set	$\mathrm{SM}(k, {}_t\mathcal{F}_r(\boldsymbol{p})) = 99$	$\mathrm{SM}(k, {}_t\mathcal{B}_r(\boldsymbol{p})) = 205$	152
Transition set approximation	$\mathrm{SM}(k, {}_t\hat{\mathcal{F}}_r(\boldsymbol{p})) = 98$	$\mathrm{SM}(k, {}_t\hat{\mathcal{B}}_r(\boldsymbol{p})) = 207$	152.5

2. Calculate the thresholds t_+ and t_-. Take ${}_t\hat{\mathcal{F}} = \{\boldsymbol{p} \mid V(\boldsymbol{p}) \geq t_+\}$ and ${}_t\hat{\mathcal{B}} = \{\boldsymbol{p} \mid V(\boldsymbol{p}) \leq -t_-\}$ (Section 5.2); see Fig. 5.5 (b).
3. Restore ${}_t\hat{\mathcal{F}}$ and ${}_t\hat{\mathcal{B}}$ (Section 5.3); see Fig. 5.5 (f)-(g).
4. Calculate the region of interest $\mathcal{R}$ (Section 5.4); see Fig. 5.5 (h).
5. Label $\boldsymbol{p}$ as background if $\boldsymbol{p} \notin \mathcal{R}$. Otherwise:
   - If binarization, compute $T(\boldsymbol{p})$ (Section 5.5); see Fig. 5.5 (i).
   - If edge detection, compute simple edge transition operator (Section 5.6); see Fig. 5.5 (j).
6. Restore $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ with standard algorithms.

5.2 Transition threshold

I known through experience that maxmin function characterizes the transition pixels better than Laplace or Linear Kernel functions. So, I assume in this section that transition values are calculated with maxmin function.

Transition values calculated with maxmin appear to obey a **Gumbel distribution** rather than obeying a normal distribution or lognormal distribution; see Fig. 4.9. As a manner of fact, these transition values can be modeled by the ***i*th order statistic** of a sample of random variables drawn from a normal distribution. However, I did not explore in detail this line of research.

Given a sample of n variables $a_1, \ldots, a_n$, reorder them so that $b_1 < \ldots < b_n$. Then b_i is called the ith order statistic.

I describe three methods based on histogram cluster thresholds. The aim of all three methods is to choose a threshold for either (5.6), or (5.8) such that the chosen threshold divides the histogram in question into two groups: The first group may be mostly constituted by non-transition pixels; the second group may be mostly constituted by transition; see Fig. 5.6.

Figure 5.5 – *(a) Original image. (b) Transition image by function* maxmin *with neighborhoods of radius 2. (c) Transition image. In blue, pixels with transition value higher than zero; in red, pixels with transition value lower than zero. (d) The transition image after filtering by* $t_+ = 14$ *and* $t_- = 15$. *(e) Transition image after removing isolated pixels. (f) Transition image after incidence transition operators. (g) Transition image after dilation transition operators. (h) Region-of-interest image. (i) Binary image by modeling the gray intensities as lognormally distributed. (j) Edge image.*

- **Empirical scaled density function**

$$u_i = \frac{1}{k} H_{V,\mathcal{P}}(i), \tag{5.6}$$

where

$$k = \max_{i \in [1,g]} \{H_{V,\mathcal{P}}(i)\}. \tag{5.7}$$

See Fig. 5.6 (top-right).

- **Empirical complementary cumulative distribution function** (CCD)

$$v_i = \frac{1}{t} \sum_{j=i}^{g} H_{V,\mathcal{P}}(j), \tag{5.8}$$

where

$$t = \sum_{j=1}^{g} H_{V,\mathcal{P}}(j). \tag{5.9}$$

See Fig. 5.6 (bottom-right).

Since ${}_{t}\mathcal{F}$ and ${}_{t}\mathcal{B}$ are dual sets, I will explain only the method for ${}_{t}\mathcal{F}$, leaving out the details for ${}_{t}\mathcal{B}$.

5.2.1 Quantile transition threshold

In [69] and [72], I suggested the **quantile transition threshold**, which I derived from **P-tile method** [19].

The quantile transition threshold discards the lowest α_+ percent of positive transition values in order to approximate ${}_{t}\mathcal{F}$ without considering transition values equal to zero, see Fig. 5.7. It implies that a $1 - \alpha_+$ percentage of the highest transition values remain in ${}_{t}\hat{\mathcal{F}}$.

Given a value α_+ and $H_{V,\mathcal{P}}$, t^+ is chosen as the minimum value that satisfies

$$\frac{1}{k} \sum_{i=1}^{t_+} H_{V,\mathcal{P}}(i) \geq \alpha_+, \tag{5.10}$$

where

$$k = \sum_{i=1}^{g} H_{V,\mathcal{P}}(i). \tag{5.11}$$

Unfortunately, the main drawback of this method is the necessity of two parameters (α_+ and α_-).

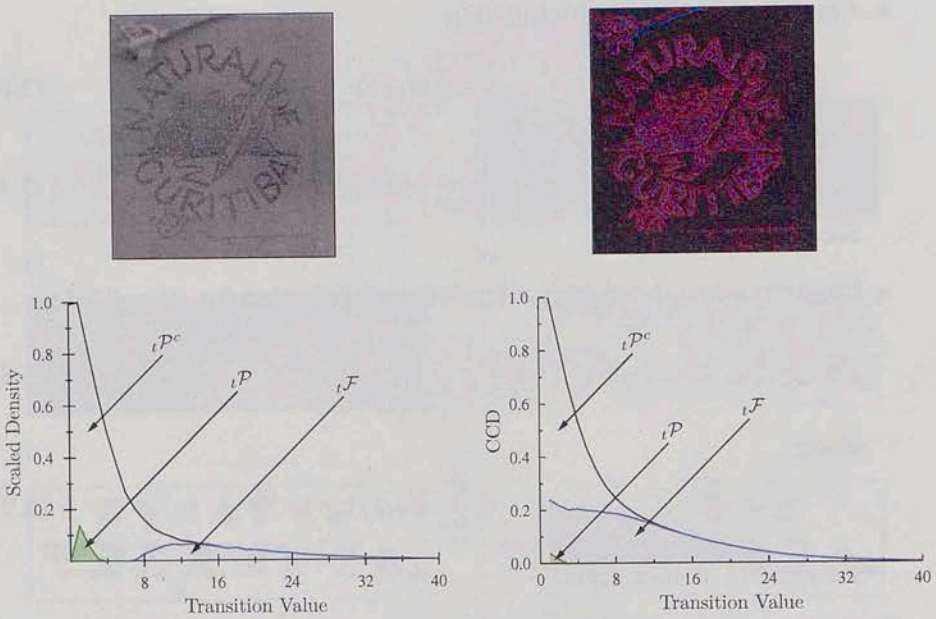


Figure 5.6 – *On the left-top, a gray image and on the right-top, its corresponding transition image (equalized image) by* maxmin *function with neighborhoods or radius 2. On the bottom, the empirical scaled density function of positive transition values (left) and empirical complementary cumulative distribution function of positive transition values (right).*

5.2.2 Rosin's threshold for transition values

I point out in [71] that the behavior of (5.6) and (5.8) is ideal for Rosin's threshold [76], which proposes a threshold for unimodal histograms.

Rosin's method [76] (Rosin's threshold) is a global algorithm, which assumes that one of the two classes produces one dominant peak located at one of the sides of the histogram. The non-dominant class may or may not produce a discernible peak, but needs to be reasonably well separated from the large peak to avoid being swamped by it.

Let w_i be values computed either with (5.6), or with (5.8). A straight line L is drawn from the peak to the high end of w_i's graph. Then, the threshold point is selected as the histogram index i which maximizes the perpendicular distance between L and the point (i, w_i); see Fig. 5.8.

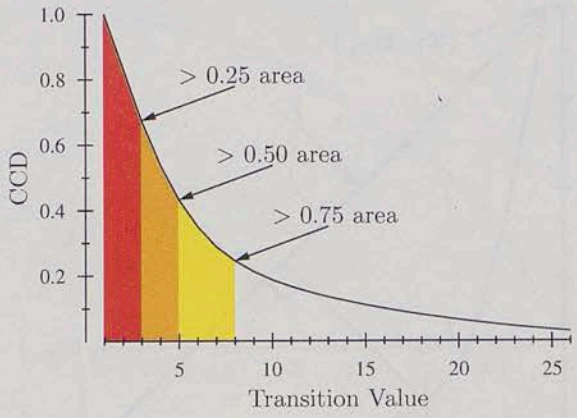


Figure 5.7 – *The positive transition threshold is calculated as the α_+ quantile of the empirical complementary cumulative distribution (CCD) function of positive transition values.*

Let $0 < x_1, x_2 < g$ be two indexes such that $w_{x_1} > w_i$ for $i = 1, \ldots, g$, and

$$\frac{w_{x_2}}{w_{x_1}} \geq \delta > \frac{w_i}{w_{x_1}} \quad \text{for } i > x_2, \tag{5.12}$$

where $\delta > 0$ is a parameter; I suggest $\delta = 0.01$. The line L is defined by the points (x_1, w_{x_1}) and (x_2, w_{x_2}). The distance function and threshold are defined as

$$D(i) = \frac{\left|(x_2 - x_1)(w_{x_1} - w_i) - (x_1 - i)(w_{x_2} - w_{x_1})\right|}{\sqrt{(x_2 - x_1)^2 + (w_{x_2} - w_{x_1})^2}}, \tag{5.13}$$

and the threshold is given by

$$t_+ = \underset{i \in [x_1, x_2]}{\arg\max} \{D(i)\}. \tag{5.14}$$

5.2.3 Double-linear threshold for transition values

The behavior of the positive transition values, see Fig. 5.6 (bottom-left), will appear to have a heavy right tail. The **power law distribution** has been discarded because the **log-log plot** (Fig. 5.9) of the empirical complementary cumulative distribution function does not follow the characteristic straight-line form of the power law distribution [57].

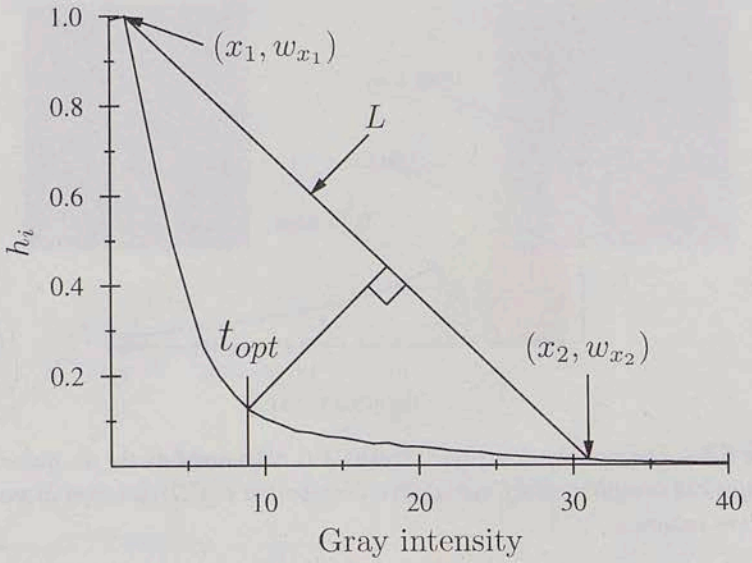


Figure 5.8 – *Rosin's threshold for positive transition values.*

A close look at Fig. 5.6 (bottom-left) shows two linear zones. The first linear relation mostly corresponds to non-transition set ${}_t\mathcal{P}^c$ having positive transition value. The second linear part is mainly formed by transition pixels. Indeed, the histogram of positive transition values is a combination of three histograms, as is shown in Fig. 5.6 (bottom-left). Thus, a criterion to select the transition threshold t_+ is to take the value t that divides the graph, into approximately two lines, using linear-linear or linear-log scales.

The **double-linear threshold** approximates the positive side of the transition graphs (i, w_i) by joining two linear functions; see Fig. 5.10 (left), where w_i is computed either with (5.6), or with (5.8). However, the transition graph is truncated between the bounds x_{min} and x_{max} in order to reduce noise in the first and last values of the graph. The value x_{min} is the minimum index i that satisfies

$$w_i > w_{i+1} \geq w_{i+2} \geq, \ldots, \geq w_g, \tag{5.15}$$

and x_{max} is the maximum index i that satisfies

$$\frac{w_i}{w_{x_{min}}} > \delta \tag{5.16}$$

such that $\delta > 0$ is small (I suggest $\delta = 0.01$).

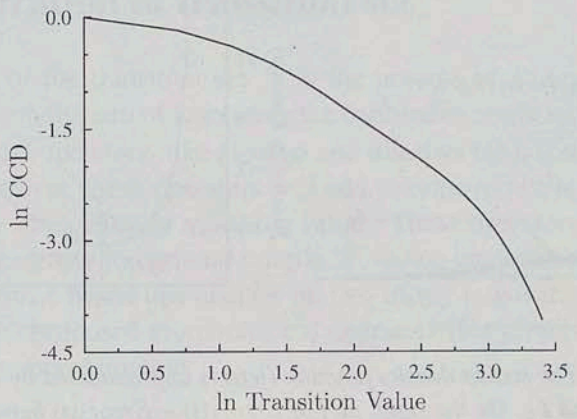


Figure 5.9 – *The log-log plot of the empirical complementary cumulative distribution functions of the positive transition pixels does not follow the characteristic straight-line form of the power-law distribution.*

For mathematical convenience, I re-label w_i as

$$y_i = w_{i+x_{min}} \text{ for } i = 0, 1, \ldots, x_{max} - x_{min} = n \tag{5.17}$$

and postulate that y_i satisfies (5.18) and (5.19).

$$y_i \approx m_1 \cdot i + b_1 \quad \text{if } i = 0, 1, 2, \ldots, t \tag{5.18}$$

$$y_i \approx m_2 \cdot i + b_2 \quad \text{if } i = t, t+1, \ldots, n. \tag{5.19}$$

I use the **differences-rate estimator** (Section 8.3) to compute $\hat{m}_1$. For this particular problem, it is simplified to

$$\hat{m}_1 = \frac{6}{t(t+1)(t+2)} \sum_{i=0}^{t} (2i - t) y_i. \tag{5.20}$$

However, the slope can be computed by **regression methods** [1] and [75].

Unfortunately, there is no differences-rate estimator for the intercept term b; therefore I use the **least-square estimator**

$$\hat{b}_1 = \frac{1}{t+1} \sum_{i=0}^{t} (y_i - \hat{m}_1 \cdot i). \tag{5.21}$$

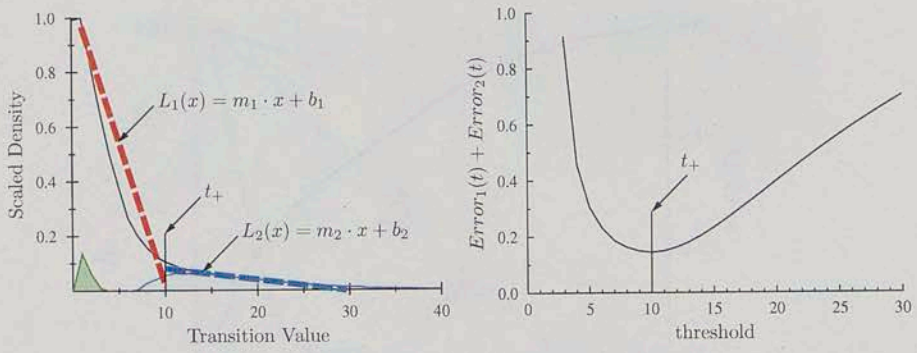


Figure 5.10 – *The scaled density function (left) is approximated by the joining of two lines L_1 and L_2. On the right, plot of $Error_1(t) + Error_3(t)$ between $x_{min} = 3$ and $x_{max} = 30$.*

A natural error function for (5.18) can be defined as

$$Error_1(t) = \sum_{i=0}^{t} \left(y_i - \hat{m}_1 \cdot i - \hat{b}_1\right)^2, \tag{5.22}$$

In the same way, an error function for (5.19) is defined as

$$Error_2(t) = \sum_{i=t}^{n} \left(y_i - \hat{m}_2 - x_i \cdot i - \hat{b}_2\right)^2, \tag{5.23}$$

where

$$\hat{m}_2 = \frac{6}{(n-t)(n-t+1)(n-t+2)} \sum_{i=0}^{n-t} (2i - n + t) y_i \tag{5.24}$$

and

$$\hat{b}_2 = \frac{1}{n-t+1} \sum_{i=t}^{n} (y_i - m_2 \cdot i) \tag{5.25}$$

Finally, t_+ is computed as

$$t_+ = \underset{t \in [1,n]}{\arg\min} \{Error_1(t) + Error_2(t)\} + x_{min} + 2. \tag{5.26}$$

Figure 5.10 (right) is a plot of (5.26).

5.3 Restoration of transition set

The restoration of the transition set ${}_t\hat{\mathcal{P}}$ is the process of adding and removing pixels from ${}_t\hat{\mathcal{P}}$ with the aim of increasing the cardinality while reducing the noise.

Morphological operators, like **erosion** and **dilation** [48], could be adapted to enhance ${}_t\hat{\mathcal{P}}$. However, these operators will add or remove pixels without considering either gray intensities, or transition values. These operators in their original form will alter the trusty foreground sample ${}_t\hat{\mathcal{F}}$, losing confidence in the transition set approximation. I based this chapter on two of my publications, namely [72] and [73], where I proposed morphological operators that preserve confidence in the transition set approximation.

5.3.1 Isolation transition operator

Isolate transition operators are derived from isolate operators in Section 2.3.1. In particular, the **cross isolate operator** and **diagonal isolate operator** were successfully used in [71], [72], and [73] for removing false positives of transition set approximations.

Cross isolate transition operator

$$ {}_t\hat{\mathcal{F}} \leftarrow \bigcup_{\boldsymbol{p}\in {}_t\hat{\mathcal{F}}} {}_t\hat{\mathcal{F}} \boxdot \mathcal{P}_{+}(\boldsymbol{p}) \tag{5.27}$$

See Definition 2.12.

Diagonal isolate transition operator

$$ {}_t\hat{\mathcal{F}} \leftarrow \bigcup_{\boldsymbol{p}\in {}_t\hat{\mathcal{F}}} {}_t\hat{\mathcal{F}} \boxdot \mathcal{P}_{\times}(\boldsymbol{p}) \tag{5.28}$$

See Definition 2.12.

Rectangular isolate transition operator

$$ {}_t\hat{\mathcal{F}} \leftarrow \bigcup_{\boldsymbol{p}\in {}_t\hat{\mathcal{F}}} \mathcal{P}_{u,v}(\boldsymbol{p}) \overset{{}_t\hat{\mathcal{F}}}{\boxdot} \mathcal{P}_{y,x}(\boldsymbol{p}), \tag{5.29}$$

See Definition 2.14.

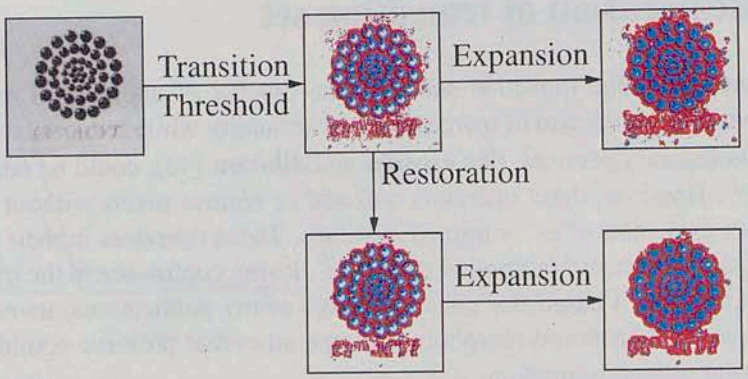


Figure 5.11 – *Two different transition set approximations. Above, an accurate transition set approximation which was previously filtered by transition operators. Below, a raw transition set approximation ($t_+ = t_- = 10$). Blue pixels depict pixels in the positive transition set. In red, those pixels in the negative transition set.*

5.3.2 Simple expansion transition operator

The cardinality of ${}_t\hat{\mathcal{F}}$ can be incremented by adding those pixels that are surrounded by positive transition pixels to ${}_t\hat{\mathcal{F}}$. Assume, for instance, that $\boldsymbol{p} \in {}_t\hat{\mathcal{P}}^c$ is a pixel such that $u = |{}_t\hat{\mathcal{F}}_t(\boldsymbol{p})|$ is a large number and $v = |{}_t\hat{\mathcal{B}}_t(\boldsymbol{p})|$ is small or zero. Then, intuitively, $\boldsymbol{p}$ may belong to ${}_t\hat{\mathcal{F}}$ with high probability. Extending this idea to neighborhoods of radius k :

Definition 5.1: *The **simple expansion transition operator***

$$ {}_t\hat{\mathcal{F}} \leftarrow {}_t\hat{\mathcal{F}} \cup \left\{ \boldsymbol{p} \in {}_t\hat{\mathcal{P}}^c \;\middle|\; |{}_t\hat{\mathcal{F}}_k(\boldsymbol{p})| \geq u \quad \text{and} \quad |{}_t\hat{\mathcal{B}}_k(\boldsymbol{p})| \leq v \right\}, \tag{5.30} $$

which is equivalent to

$$ {}_t\hat{\mathcal{F}} \leftarrow \bigcup_{\boldsymbol{p} \in {}_t\hat{\mathcal{P}}^c} {}_t\hat{\mathcal{F}} \overset{\mathcal{P}_k(\boldsymbol{p})}{\underset{u,v}{\geqslant}} {}_t\hat{\mathcal{B}}. \tag{5.31} $$

See Definition 2.16.

The simple expansion transition operator is sensitive to noise, and it can easily lead to mistrustful approximations because it does not consider either gray intensities or transition values. Nonetheless, it is useful when the boundaries between ${}_t\hat{\mathcal{F}}$ and ${}_t\hat{\mathcal{B}}$ are well defined and there are no scattered noise spots; see Fig. 5.11.

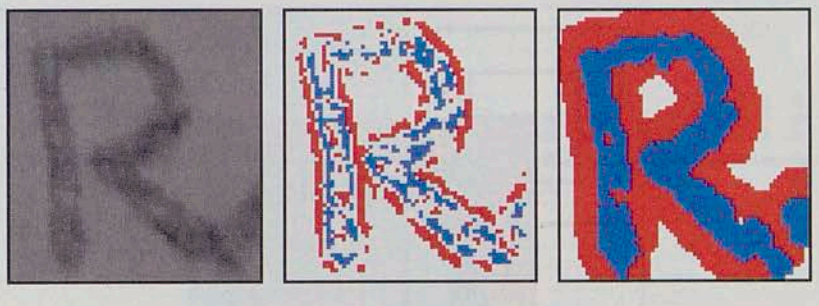


Figure 5.12 – *On the left, original image. In the center, filtered image using* $t_+ = 14$ *and* $t_- = 16$. *On the right, Restored image by isolation transition operators and expansion transition operators. Blue pixels depict pixels in the positive transition set. In red, those pixels in the negative transition set below* t_-.

I did not determine a practical rule to *"tune"* the simple expansion transition operator. In most of the cases, this operator is only helpful through no-trivial combinations of transition operators. For example, Figure 5.12 (right) was computed with seven transition operators in the following order:

1. Expansion transition operator ($k = 2$ and $u = v = 3$).
2. Cross isolate transition operator.
3. Expansion transition operator ($k = 2$ and $u = 3$, $v = 13$).
4. Diagonal isolate transition operator.
5. Cross isolate transition operator.
6. Expansion transition operator ($k = 1$, $u = 5$, and $v = 5$).
7. Expansion transition operator ($k = 2$, $u = 13$, and $v = 2$).

5.3.3 Incidence transition operator

The blue pixels in Fig. 5.13 depict pixels with high positive transition values. In red, those pixels with high negative transition values. In the same figure, whereas the isolated blue pixel $\boldsymbol{q}$ (right bottom corner) is an outlier and easily removed by cross, diagonal, or rectangular transition operators, the red pixels around $\boldsymbol{q}$ form a

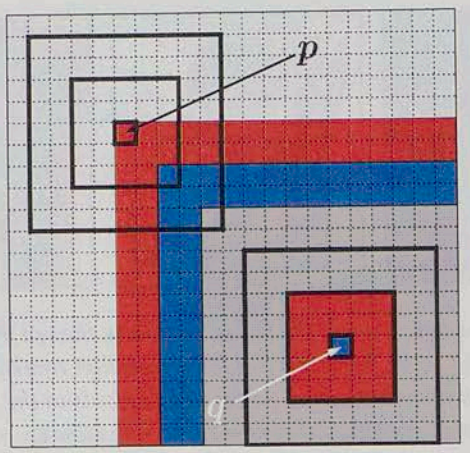


Figure 5.13 – *The transition values were computed using* maxmin *width radius 2. The pixels with high positive values are shown in blue, in red the pixels with high negative values.*

large "isolated" connected component (24 pixels) that cannot be removed by those operators.

By definition, a background t-transition pixel $\boldsymbol{p}$ contains at least one foreground t-transition pixel in $\mathcal{P}_t(\boldsymbol{p})$. That is $|{}_t\mathcal{F}_t(\boldsymbol{p})| \geq 1$, if $\boldsymbol{p} \in {}_t\mathcal{P}$. Moreover, $|{}_t\mathcal{F}_{2t}(\boldsymbol{p})| > |{}_t\mathcal{F}_t(\boldsymbol{p})|$ in most of the transition pixels. Thus, the neighborhood $\mathcal{P}_{2t}(\boldsymbol{p})$ of a pixel with a high positive transition value may contain several pixels with high positive transition values. For example, Fig. 5.13 depicts $|{}_2\hat{\mathcal{F}}_2(\boldsymbol{p})| = 1$ and $|{}_2\hat{\mathcal{F}}_4(\boldsymbol{p})| = 8$. In opposition to $\boldsymbol{p}$, the pixel $\boldsymbol{q}$ and all the red pixels around it only contain one blue pixel in $\mathcal{P}_4(\boldsymbol{q})$.

To deal with pixels like $\boldsymbol{q}$, I proposed in [73] the following definition:

Definition 5.2: *A pixel* $\boldsymbol{p}$ *is an* ***isolated transition pixel*** *if*

$$|{}_t\mathcal{F} \cap \mathcal{P}_k(\boldsymbol{p})| < f \quad or \quad |{}_t\mathcal{B} \cap \mathcal{P}_k(\boldsymbol{p})| < b \tag{5.32}$$

where f *and* b *are two positive integers. An alternative form of (5.32) is*

$$\left({}_t\mathcal{F} \underset{f}{\boxdot} \mathcal{P}_k(\boldsymbol{p})\right) \cap \left({}_t\mathcal{B} \underset{b}{\boxdot} \mathcal{P}_k(\boldsymbol{p})\right) = \emptyset. \tag{5.33}$$

This alternative expression is helpful to calculate this operator with integral images; see Definition 2.15 and Chapter A.

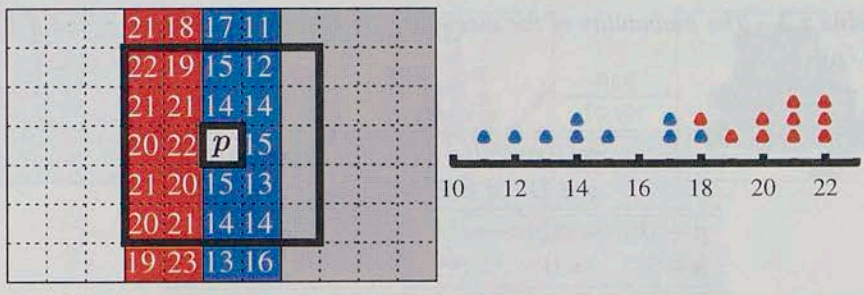


Figure 5.14 – *On the left, the gray-intensity of e blue pixel in $\mathcal{P}_2(\boldsymbol{p})$ is lower or equal to $I(\boldsymbol{p})$. On the right, the approximation of the transition balance of $\boldsymbol{p}$.*

Therefore, isolated transition pixels can be removed from the transition set approximation with the **incidence transition operator** (Definition 2.15) as follows:

$$ {}_t\hat{\mathcal{F}} \leftarrow \bigcup_{\boldsymbol{p}\in {}_t\hat{\mathcal{F}}} {}_t\hat{\mathcal{F}} \underset{f,b}{\overset{\mathcal{P}_k(\boldsymbol{p})}{\bowtie}} {}_t\hat{\mathcal{B}}, \tag{5.34} $$

where k is a positive integer. I recommend setting $k = 2t$, $f = b = 1 + t$.

Remark 5.1: The incidence operator does not remove dense **random-valued noise.** Thus, it has to be applied after isolate transition operators.

Given a partition $\mathcal{P}$ and an image function F, **random-valued noise** is a set of pixels $\mathcal{A} \subset \mathcal{P}$ whose spatial position are uniformly distributed, and whose values can take any random value of F [9], [23].

5.3.4 Dilation transition operator

Suppose that $\boldsymbol{p}$ and $\boldsymbol{q} \in \mathcal{P}_t(\boldsymbol{p})$ are two foreground pixels such that $\boldsymbol{q} \in {}_t\mathcal{F}$ and $\boldsymbol{p} \notin {}_t\mathcal{F}$, thus $V(\boldsymbol{q}) \geq t_+$ and $V(\boldsymbol{p}) < t_+$. This implies that $\boldsymbol{p}$ is excluded from ${}_t\hat{\mathcal{F}}$. However, we can assume

$$ \Pr(I(\boldsymbol{q}) \geq I(\boldsymbol{p})) \approx \Pr(I(\boldsymbol{q}) \leq I(\boldsymbol{p})) \quad \text{if } \boldsymbol{p}, \boldsymbol{q} \in \mathcal{F} \cap \mathcal{P}_t(\boldsymbol{p}) \tag{5.35} $$

because

$$ I(\boldsymbol{q}) \approx I(\boldsymbol{p}) \quad \text{for all } \boldsymbol{q} \in \mathcal{F} \cap \mathcal{P}_t(\boldsymbol{p}). \tag{5.36} $$

So,

$$ \Pr(I(\boldsymbol{q}) \geq I(\boldsymbol{p})) \approx \Pr(I(\boldsymbol{q}) \leq I(\boldsymbol{p})) \quad \text{if } \boldsymbol{p}, \boldsymbol{q} \in {}_t\mathcal{F}_t(\boldsymbol{p}). \tag{5.37} $$

In other words, about half of the pixels in ${}_t\hat{\mathcal{F}}_t(\boldsymbol{p})$ have a gray intensity equal or lower than $I(\boldsymbol{p})$; see Fig. 5.14. In addition, the gray intensities of the background are strictly higher than $I(\boldsymbol{p})$ in the ideal case. Therefore, the number of pixels

Table 5.2 – *The probability of the inequality are approximated given* $\boldsymbol{p}$ *and* $\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})$

	$\Pr(I(\boldsymbol{p}) \geq I(\boldsymbol{q}))$		$\Pr(I(\boldsymbol{p}) \leq I(\boldsymbol{q}))$	
	$\boldsymbol{q} \in \mathcal{B}$	$\boldsymbol{q} \in \mathcal{F}$	$\boldsymbol{q} \in \mathcal{B}$	$\boldsymbol{q} \in \mathcal{F}$
$\boldsymbol{p} \in \mathcal{B}$	≈ 0.5	≈ 1	≈ 0.5	≈ 0
$\boldsymbol{p} \in \mathcal{F}$	≈ 0	≈ 0.5	≈ 1	≈ 0.5

that are equal or lower in gray intensity than $I(\boldsymbol{p})$ may be zero or close to zero. Table 5.2 is constructed following the same reasoning, although a formal proof of the probabilities is beyond the scope of this thesis.

Using the conditional probabilities of Table 5.2, a large number of pixels in ${}_t\mathcal{F}_t(\boldsymbol{p})$ that are equal or lower in gray intensity than $I(\boldsymbol{p})$ is strong evidence that $\boldsymbol{p}$ belongs to the foreground. We derived a similar argument for background pixels. To measure these conditional probabilities, we define:

Definition 5.3: *The* ***t-transition balance****:*

$$TB_t(\boldsymbol{p}) = |\{\boldsymbol{q} \in {}_t\mathcal{F}_t(\boldsymbol{p}) \mid I(\boldsymbol{q}) \geq I(\boldsymbol{p})\}| - |\{\boldsymbol{q} \in {}_t\mathcal{B}_t(\boldsymbol{p}) \mid I(\boldsymbol{q}) \leq I(\boldsymbol{p})\}|. \tag{5.38}$$

So, $TB_t(\boldsymbol{p}) \approx \frac{1}{2}|{}_t\mathcal{F}_t(\boldsymbol{p})|$ if $\boldsymbol{p}$ is foreground, and $TB_t(\boldsymbol{p}) \approx -\frac{1}{2}|{}_t\mathcal{B}_t(\boldsymbol{p})|$ if $\boldsymbol{p}$ is background. Hence, $TB_t(\boldsymbol{p})$ is approximated with $\widehat{TB}_t(\boldsymbol{p})$, which uses ${}_t\hat{\mathcal{P}}_t(\boldsymbol{p})$ instead of ${}_t\mathcal{P}_t(\boldsymbol{p})$.

Definition 5.4: *Given* $\boldsymbol{p} \notin {}_t\hat{\mathcal{P}}$, *the* ***dilation transition operator*** *set*

$${}_t\hat{\mathcal{F}} \leftarrow {}_t\hat{\mathcal{F}} \cup \{\boldsymbol{p}\} \quad \textit{if} \quad \widehat{TB}(\boldsymbol{p}) \geq f, \tag{5.39}$$

and

$${}_t\hat{\mathcal{B}} \leftarrow {}_t\hat{\mathcal{B}} \cup \{\boldsymbol{p}\} \quad \textit{if} \quad \widehat{TB}(\boldsymbol{p}) \leq -b, \tag{5.40}$$

where f *and* b *are two positive integers.*

I recommend setting $f = b = 1 + t$.

5.4 Detection of regions of interest

For a human observer, detecting a perceptually important region in an image is a natural task which is done instantaneously, but for a machine it is far more

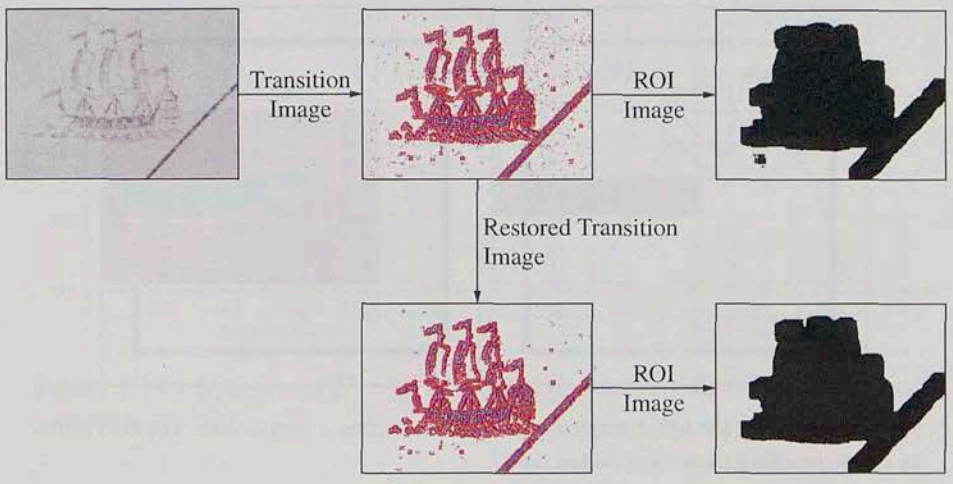


Figure 5.15 – *Detection of region of interest by transition set. The ROI image on the top is computed with the rough transition set approximation (without transition operators). The ROI image on the bottom is computed with the restored transition set.*

difficult. The machine lacks the cultural references and knowledge to identify the content of the scene.

One of the causes for this difficulty is the subjective nature of the notion of **region of interest** (ROI). In the most general sense, a region of interest is a part of the image for which the observer of the image shows interest. For example, in medical images, a definition of region of interest is based on anatomical markers [62]; in computer vision, Caron et al. [7] assume that the region of interest to be detected is a single connected region in the image; it must be both significant in size and different from the background in structural complexity.

The interest shown by the observer in viewing the image is determined not only by the image itself, but also by the observer's own sensitivity. For a given image, different people could find different regions of interest. However, regions of interest generally have distinctive features (contrast, color, region size and shape, distribution of contours or texture pattern) which make it possible to distinguish regions of interest from the rest of the images. Then, these structural characteristics can be used to detect regions of interest of an image without making hypotheses about the semantic content of the picture.

In document analysis context, a region of interest can be defined as the set *region of interest*

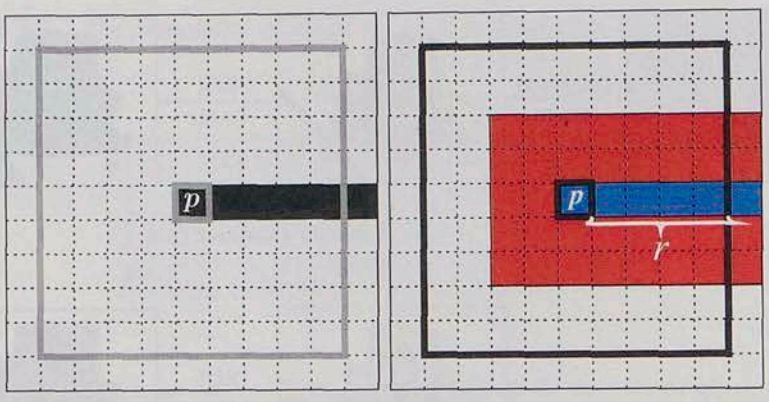


Figure 5.16 – *On the left, binary image which contains a simple line. On the right, its corresponding transition image (t=2).*

of pixels $\mathcal{R}$ such that the neighborhood of radius r of each pixel contains both foreground and background. Indeed, this is the definition of t-transition set for $t = r$. Therefore, under this definition, ${}_r\mathcal{P} = \mathcal{R}$.

The properties of smoothness (Definition 4.2) and local contrast (Definition 4.3) do not hold for the radius $r > s$ (recalling s from Definition 4.2 and Definition 4.3) so that transition values cannot characterize the r-transition pixels. Nevertheless, ${}_r\mathcal{P}$ is fairly estimated by

$$\mathcal{R} \approx \hat{\mathcal{R}} = {}_r\hat{\mathcal{P}} = \left\{ \boldsymbol{p} \mid |{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq n_+ \text{ and } |{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq n_- \right\} \tag{5.41}$$

where n_+ and n_- are two positive integers. An alternative expression is given by

$$\mathcal{R} \approx \bigcup_{\boldsymbol{p} \in \mathcal{P}} {}_t\hat{\mathcal{F}} \underset{n_+, n_-}{\overset{\mathcal{P}_r(\boldsymbol{p})}{\bowtie}} {}_t\hat{\mathcal{B}} \tag{5.42}$$

The values n_+ and n_- depend on r and objects of interest: the larger n_+ and n_-, the larger the objects that can be removed from the foreground. Figure 5.16 (left), for instance, depicts a simple horizontal line with height 1 as foreground. The line extremes are evaluated if $n_+ \leq r + 1$. Otherwise, the line extremes are labeled as background without even computing $T(\boldsymbol{p})$; see Figure 5.16 (right). In [72] and [73] I suggested $n_+ = n_- = 5$ for detecting small foreground objects.

A second criterion to discard outliers uses the difference between the mean of gray intensities of transition sets. The pixel $\boldsymbol{p}$ is labeled as background if

$$\mu_{I_{\cdot t}\hat{\mathcal{B}}_r(\boldsymbol{p})} - \mu_{I_{\cdot t}\hat{\mathcal{F}}_r(\boldsymbol{p})} < c, \tag{5.43}$$

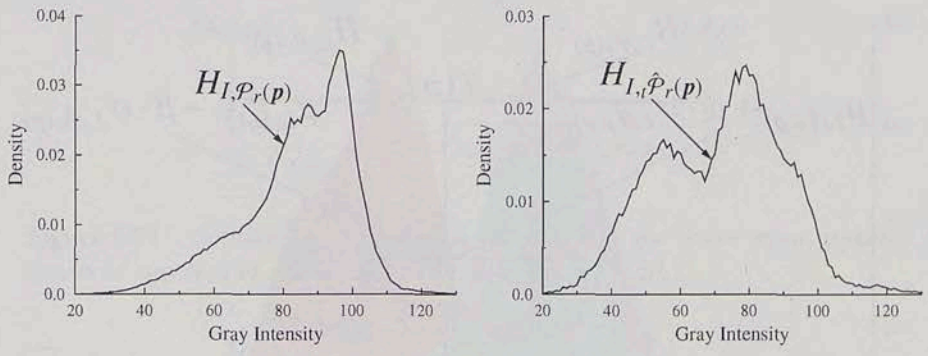


Figure 5.17 – $H_{I,\mathcal{P}_r(\boldsymbol{p})}$ *and* $H_{I,\hat{\mathcal{P}}_r(\boldsymbol{p})}$ *of Fig. 5.6 (top-left) on the left and right, respectively.*

where c is an integer, which depicts the minimum contrast expected between the foreground and background. In [72] and [73], I suggested $c = 15$.

5.5 Binarization by transition sets

At this point, I assume that $\boldsymbol{p} \in \mathcal{R}$ (region of interest). Otherwise, the pixel is directly classified as background.

For some algorithms, like Otsu's and Kittler's thresholds, the better the histogram of gray intensities approximates a bimodal curve, the better their accuracy. Those algorithms compute $T(\boldsymbol{p})$ with data from $H_{I,\mathcal{P}_r(\boldsymbol{p})}$. I propose $H_{I,\hat{\mathcal{P}}_r(\boldsymbol{p})}$ instead; see Fig. 5.17. Moreover, keeping track of $H_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $H_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}$, I propose several classification functions.

5.5.1 Linear mean-variance threshold

I introduced the **linear mean-variance threshold** in [69]; it follows the same idea of **Niblack's threshold** because it resorts to intervals based on mean and variance of gray intensities. It assumes that the gray intensities of the foreground are clustered such that most of them are contained in the interval $\mu_{I,\mathcal{F}_r(\boldsymbol{p})} \pm \alpha \cdot \sigma_{I,\mathcal{F}_r(\boldsymbol{p})}$ (**foreground interval**). In a similar manner, most of the gray intensities of the background are within $\mu_{I,\mathcal{B}_r(\boldsymbol{p})} \pm \beta \cdot \sigma_{I,\mathcal{B}_r(\boldsymbol{p})}$ (**background interval**) and, as a consequence, the optimal threshold must lie between $\mu_{I,\mathcal{F}_r(\boldsymbol{p})} + \alpha \cdot \sigma_{I,\mathcal{F}_r(\boldsymbol{p})}$ and $\mu_{I,\mathcal{B}_r(\boldsymbol{p})} - \beta \cdot \sigma_{I,\mathcal{B}_r(\boldsymbol{p})}$ in an ideal image. Hence, the linear mean-variance threshold

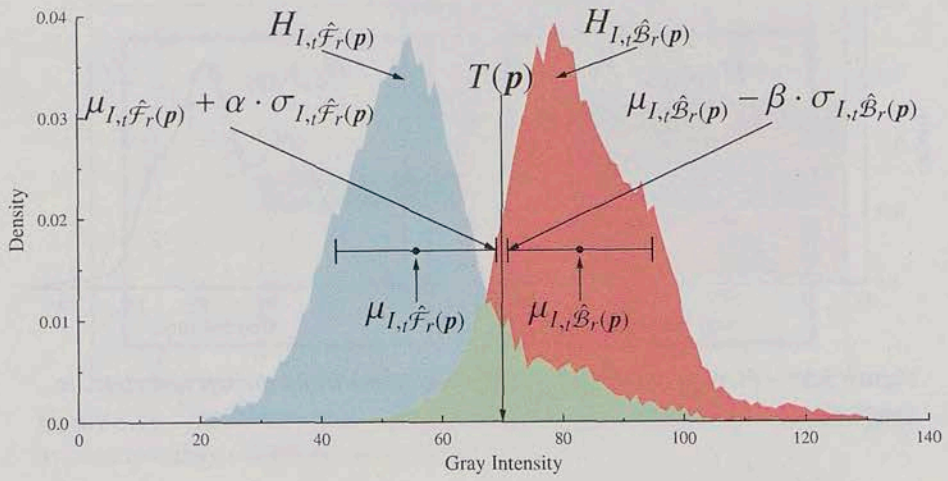


Figure 5.18 – $H_{I,t\hat{\mathcal{F}}_r(p)}$ *and* $H_{I,t\hat{\mathcal{B}}_r(p)}$ *of Fig. 5.6 (top-left) filtered by* $t_+ = 9$ *and* $t_- = -9$.

is given by

$$T(\boldsymbol{p}) = \frac{\mu_{I,t\hat{\mathcal{F}}_r(p)} + \alpha \cdot \sigma_{I,t\hat{\mathcal{F}}_r(p)} + \mu_{I,t\hat{\mathcal{B}}_r(p)} - \beta \cdot \sigma_{I,t\hat{\mathcal{B}}_r(p)}}{2}. \tag{5.44}$$

Figure 5.18, for instance, shows the optimal threshold with $\alpha = \beta = 1$.

The main disadvantage of the linear mean-variance threshold is that suitable parameters may change significantly between two different images.

A second disadvantage is that both foreground and background intervals may be overlapped, in which case $T(\boldsymbol{p})$ may be lower than $\mu_{I,t\hat{\mathcal{F}}_r(p)}$ or greater than $\mu_{I,t\hat{\mathcal{B}}_r(p)}$, which contradicts the assumptions of smoothness (Definition 4.2) and local contrast (Definition 4.3). Figure 5.19 exemplifies this problem with

$$\begin{aligned} \mu_{I,t\hat{\mathcal{B}}_r(p)} &= \mu_{I,t\hat{\mathcal{F}}_r(p)} + \sigma_{I,t\hat{\mathcal{F}}_r(p)} + \sigma_{I,t\hat{\mathcal{B}}_r(p)} \\ \sigma_{I,t\hat{\mathcal{B}}_r(p)} &= 2 \cdot \sigma_{I,t\hat{\mathcal{F}}_r(p)}. \end{aligned} \tag{5.45}$$

Then

$$T(\boldsymbol{p}) = \mu_{I,t\hat{\mathcal{F}}_r(p)} + \frac{\sigma_{I,t\hat{\mathcal{F}}_r(p)} \cdot [\alpha + 3 - 2\beta]}{2} \tag{5.46}$$

Therefore, $T(\boldsymbol{p}) < \mu_{I,t\hat{\mathcal{F}}_r(p)}$ if $\alpha + 3 < 2\beta$.

In [69], the linear mean-variance threshold yielded good binarization results with $\alpha = \beta = 1$.

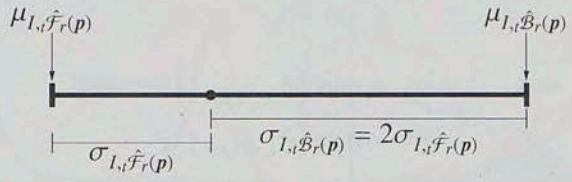


Figure 5.19 – *Considering the values in this diagram, the linear mean-variance threshold may lead to an unsuitable threshold if* $\alpha + 3 < 2\beta$.

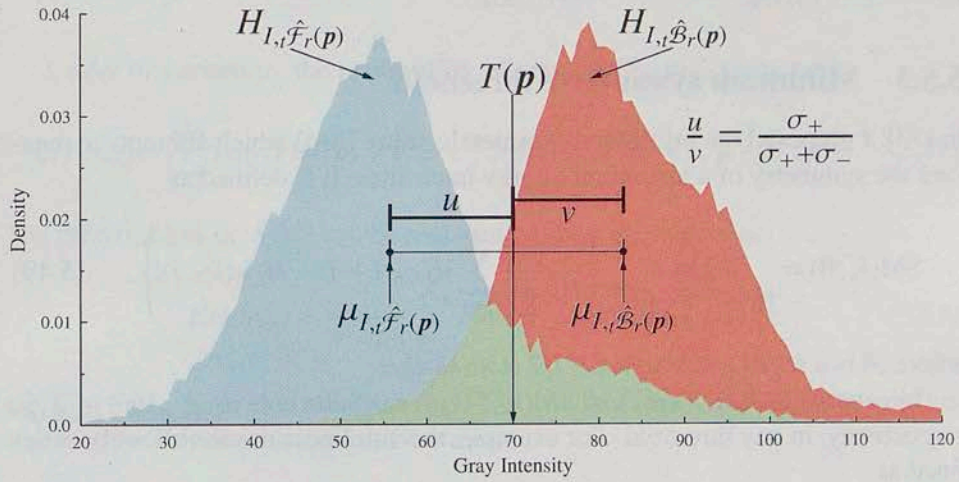


Figure 5.20 – *The autolinear threshold is a point between the segment with extremes* $\mu_{I,t\hat{\mathcal{F}}_r(p)}$ *and* $\mu_{I,t\hat{\mathcal{B}}_r(p)}$ *which divides in a proportion related to the standard deviation of gray intensities.*

5.5.2 Autolinear threshold

I introduced the **autolinear threshold** in [72] to overcome the shortcoming of the **linear mean-variance threshold** [69], which needs two parameters.

As I point out in Section 5.5.1, the optimal threshold must lie between the interval $\hat{\mu}_{I,\mathcal{F}_r(p)}$ and $\hat{\mu}_{I,\mathcal{B}_r(p)}$. With this assumption, the autolinear threshold chooses a threshold between such means as

$$T(\boldsymbol{p}) = \mu_{I,t\hat{\mathcal{F}}_r(p)} + \frac{\sigma_+}{\sigma_+ + \sigma_-}\left[\hat{\mu}_{I,t\hat{\mathcal{B}}_r(p)} - \hat{\mu}_{I,t\hat{\mathcal{F}}_r(p)}\right]. \tag{5.47}$$

where

$$\begin{aligned}\sigma_+ &= \max\left(\hat{\sigma}_{I,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})}, 1\right)\\ \sigma_- &= \max\left(\hat{\sigma}_{I,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})}, 1\right),\end{aligned} \tag{5.48}$$

see Fig. 5.20.

In this manner, if $\sigma_+ = \sigma_-$, then the threshold is chosen as the middle point between the means of gray intensities. Furthermore, it ensure that $T(\boldsymbol{p})$ is always greater than $\hat{\mu}_{I,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and lower than $\hat{\mu}_{I,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})}$.

5.5.3 Minimum symmetric threshold

In [73], I proposed the **minimum symmetric value** (SM) which attempts to measure the symmetry of a histogram of gray intensities. It is defined as

$$\mathrm{SM}(k,\mathcal{A}) = \underset{i\in[k,g-k],\, H_{I,\mathcal{A}}(i)>0}{\arg\min} \left\{ \frac{1}{H_{I,\mathcal{A}}(i)} \sum_{j=1}^{k} |H_{I,\mathcal{A}}(i+j) - H_{I,\mathcal{A}}(i-j)| \right\}, \tag{5.49}$$

where $\mathcal{A}$ is a set of pixels and $k \leq l/2$ is an integer.

In general, $\mathrm{SM}(k,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p}))$ and $\mathrm{SM}(k,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p}))$ can substitute $\mu_{I,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\mu_{I,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})}$, respectively, in any threshold. For example, the autolinear threshold can be redefined as

$$T(\boldsymbol{p}) = \mathrm{SM}(k,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})) + \frac{\sigma_+}{\sigma_+ + \sigma_-}\left[\mathrm{SM}(k,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})) - \mathrm{SM}(k,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p}))\right] \tag{5.50}$$

where σ_+ and σ_- are computed as (5.48).

5.5.4 Minimum-error-rate

According to **Bayesian decision theory,** the probability of misclassifying a pixel is minimized with the **Bayes decision rule**:

$$\text{classify } \boldsymbol{p} \text{ as } \begin{cases}\text{foreground} & \text{if } \Pr(\boldsymbol{p}\in\mathcal{F}_r(\boldsymbol{p}) \mid I(\boldsymbol{p})=i) \geq \Pr(\boldsymbol{p}\in\mathcal{B}_r(\boldsymbol{p}) \mid I(\boldsymbol{p})=i)\\ \text{background} & \text{if } \Pr(\boldsymbol{p}\in\mathcal{F}_r(\boldsymbol{p}) \mid I(\boldsymbol{p})=i) < \Pr(\boldsymbol{p}\in\mathcal{B}_r(\boldsymbol{p}) \mid I(\boldsymbol{p})=i)\end{cases} \tag{5.51}$$

where the notation $\Pr(\boldsymbol{p}\in\mathcal{A} \mid I(\boldsymbol{p})=i)$ denotes $\Pr(\boldsymbol{p}\in\mathcal{A})$ given that the gray intensity of $\boldsymbol{p}$ is i.

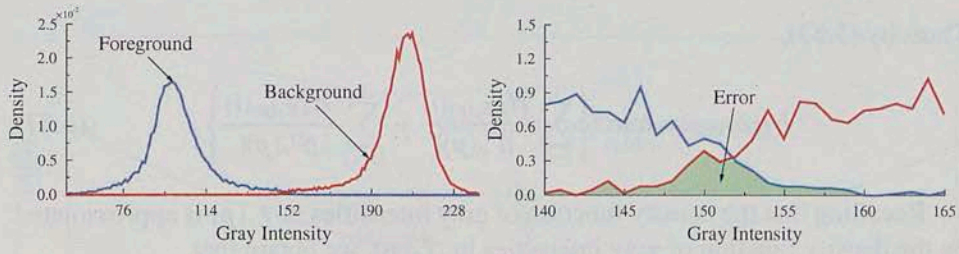


Figure 5.21 – *Empirical density functions of gray intensities from Fig. 5.1. The probability of error in light green.*

Under this criterion, the **probability of error** in $\mathcal{P}_r(\boldsymbol{p})$ is given by

$$Error_{min} = \sum_{i=0}^{g} \min\left\{\Pr(I(\boldsymbol{p}) = i\,,\ \boldsymbol{p} \in \mathcal{F}_r(\boldsymbol{p})),\ \Pr(I(\boldsymbol{p}) = i\,,\ \boldsymbol{p} \in \mathcal{B}_r(\boldsymbol{p}))\right\}. \quad (5.52)$$

The probabilities in (5.52) can be replaced by their estimators as

$$Error_{min} \approx \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|} \sum_{i=0}^{g} \min\left\{H_{I,\mathcal{F}_r(\boldsymbol{p})}(i)\,,\ H_{I,\mathcal{B}_r(\boldsymbol{p})}(i)\right\}; \quad (5.53)$$

see Fig. 5.21. Note that the factor $\frac{1}{|\mathcal{P}_r(\boldsymbol{p})|}$ is a scale factor. Therefore, given that $I(\boldsymbol{p}) = i$, the Bayes decision rule becomes:

$$\text{classify } \boldsymbol{p} \text{ as } \begin{cases} \text{foreground} & \text{if } w \cdot H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) \geq w \cdot H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) \\ \text{background} & \text{if } w \cdot H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) < w \cdot H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) \end{cases} \quad (5.54)$$

where $w > 0$ is a scale factor.

According to Section 4.1, the gray intensities in $\mathcal{F}_r(\boldsymbol{p})$ are approximately normally (lognormally) distributed. Therefore, there must exist $t_{opt} \in [0, g]$ such that

$$\begin{aligned} H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) &\geq H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) \quad \text{if } i \leq t_{opt}, \\ H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) &\leq H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) \quad \text{if } i > t_{opt}. \end{aligned} \quad (5.55)$$

However, the frequency of gray intensities randomly fluctuate and, as a consequence, a value that satisfies (5.55) may not exist. Nevertheless,

$$\frac{1}{|\mathcal{P}_r(\boldsymbol{p})|} \sum_{i=0}^{g} \min\left\{H_{I,\mathcal{F}_r(\boldsymbol{p})}(i)\,,\ H_{I,\mathcal{B}_r(\boldsymbol{p})}(i)\right\} \approx \min_{t \in [0,g]} \left\{ \sum_{i=0}^{t} \frac{H_{I,\mathcal{B}_r(\boldsymbol{p})}(i)}{|\mathcal{P}_r(\boldsymbol{p})|} + \sum_{i=t+1}^{g} \frac{H_{I,\mathcal{F}_r(\boldsymbol{p})}(i)}{|\mathcal{P}_r(\boldsymbol{p})|} \right\}. \quad (5.56)$$

Thus, by (5.53),

$$Error_{min} \approx \min_{t\in[0,g]} \left\{ \sum_{i=0}^{t} \frac{H_{I,\mathcal{B}_r(\boldsymbol{p})}(i)}{|\mathcal{P}_r(\boldsymbol{p})|} + \sum_{i=t+1}^{g} \frac{H_{I,\mathcal{F}_r(\boldsymbol{p})}(i)}{|\mathcal{P}_r(\boldsymbol{p})|} \right\}. \tag{5.57}$$

Recalling that the density function of gray intensities in $\mathcal{F}_r(\boldsymbol{p})$ is approximated by the density function of gray intensities in ${}_t\hat{\mathcal{F}}_r(\boldsymbol{p})$, we obtain that

$$\begin{aligned} H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) &\approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \Pr(I(\boldsymbol{p}) = i \mid \boldsymbol{p} \in \mathcal{F}_r(\boldsymbol{p})) \\ &\approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \Pr(I(\boldsymbol{p}) = i \mid \boldsymbol{p} \in {}_t\mathcal{F}_r(\boldsymbol{p})) \\ &\approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \Pr(I(\boldsymbol{p}) = i \mid \boldsymbol{p} \in {}_t\hat{\mathcal{F}}_r(\boldsymbol{p})) = |\mathcal{F}_r(\boldsymbol{p})| \cdot \frac{H_{I,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})}(i)}{|{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})|} \end{aligned} \tag{5.58}$$

This implies that

$$Error_{min} \approx \min_{t\in[0,g]} \{Error(t)\} = Error(\hat{t}_{opt}) \tag{5.59}$$

where

$$Error(t) = [1 - w_f] \sum_{i=0}^{t} \frac{H_{I,{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})}(i)}{|{}_t\hat{\mathcal{B}}_r(\boldsymbol{p})|} + w_f \sum_{i=t+1}^{g} \frac{H_{I,{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})}(i)}{|{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})|} \tag{5.60}$$

and

$$w_f = \frac{|\mathcal{F}_r(\boldsymbol{p})|}{|\mathcal{P}_r(\boldsymbol{p})|}. \tag{5.61}$$

The value w_f is known as the **foreground proportion** in $\mathcal{P}_r(\boldsymbol{p})$.

Figure 5.22 shows that the error function ($Error(t)$) for the "*true foreground and background*" (ℓ) is similar to the error function for both the transition set approximation ($\hat{\ell}$) and the "*true transition set*" ($\hat{\ell}^*$). In this example, t_{opt} exists and coincides with the minimum value of $\hat{\ell}^*$. Furthermore, the minimum probability of error is ≈ 0.0038 while the probability of error at level $\hat{t}_{opt} = 144$ is ≈ 0.0066. So, $\hat{t}_{opt}$ is an accurate estimator of the minimum error.

In our previous example, all MER graphs were computed assuming the true value of w_f because $|\mathcal{F}_r(\boldsymbol{p})|$ and $|\mathcal{P}_r(\boldsymbol{p})|$ are known for this example. However, usually w_f is unknown and may be estimated in some manner. Unfortunately, $|{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})|$ cannot be taken as a proportional estimator of $|\mathcal{F}_r(\boldsymbol{p})|$ since the w_f is usually different to the ratio

$$\frac{|{}_t\hat{\mathcal{F}}_r(\boldsymbol{p})|}{|{}_t\hat{\mathcal{P}}_r(\boldsymbol{p})|}. \tag{5.62}$$

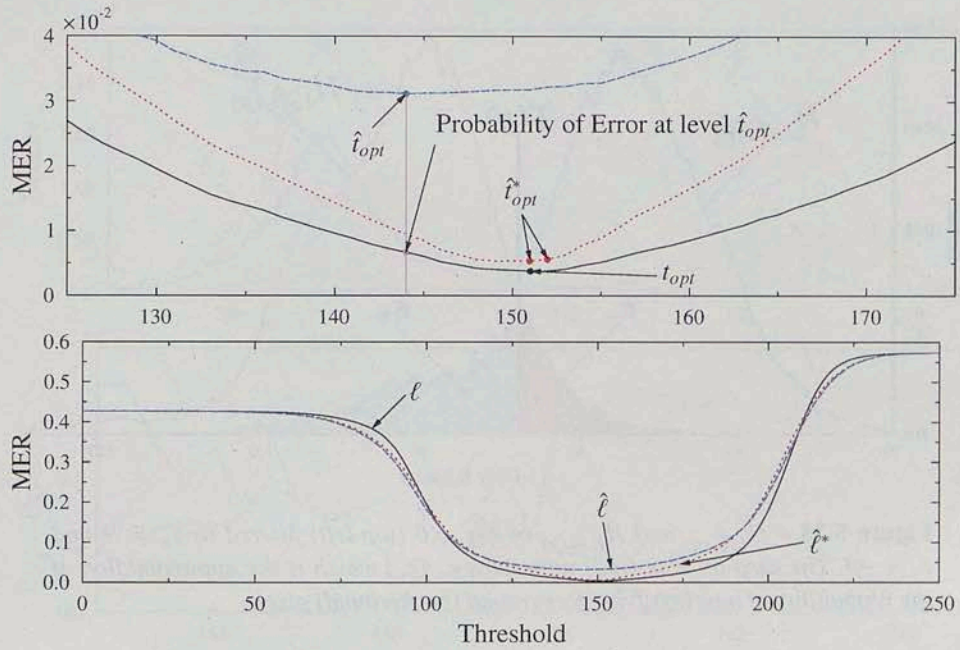


Figure 5.22 – *MER functions of Fig. 5.1: ℓ (thick black solid line) is computed from $H_{I,\mathcal{F}_r(\boldsymbol{p})}$ and $H_{I,\mathcal{B}_r(\boldsymbol{p})}$; $\hat{\ell}^*$ (red dotted line) is computed from $H_{I,\mathcal{F}_r(\boldsymbol{p})}$ and $H_{I,\mathcal{B}_r(\boldsymbol{p})}$; and $\hat{\ell}$ (blue dashed line) is computed from $H_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $H_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}$. In this example, $t_{opt} = 151$, $\hat{t}^*_{opt}$ has two minimum values in 151 and 152, and $\hat{t}_{opt} = 144$.*

This is because positive and negative transition pixels customarily come in pairs while w_f depends on r and the spatial position of $\boldsymbol{p}$.

The **minimum-error-rate threshold** based on transition sets is then defined as

$$\hat{t} = \arg\min_{t\in[0,g]} \left\{ \underbrace{[1-\hat{w}_f] \sum_{i=0}^{t} \frac{H_{I,t\hat{\mathcal{B}}_r(\boldsymbol{p})}(i)}{|_t\hat{\mathcal{B}}_r(\boldsymbol{p})|}}_{\epsilon_-} + \underbrace{\hat{w}_f \sum_{i=t+1}^{g} \frac{H_{I,t\hat{\mathcal{F}}_r(\boldsymbol{p})}(i)}{|_t\hat{\mathcal{F}}_r(\boldsymbol{p})|}}_{\epsilon_+} \right\}. \tag{5.63}$$

where $\hat{w}_f$ denotes an estimate of w_f (either given as parameter, or calculated by some method).

If $w_f < \hat{w}_f$, the minimum-error-rate threshold tends to overestimate $\hat{t}_{opt}$. Conversely, if $w_f > \hat{w}_f$, the minimum-error-rate threshold tends to underestimate $\hat{t}_{opt}$. However, $\Pr(I(\boldsymbol{p}) = i \mid \boldsymbol{p} \in \mathcal{F}_r(\boldsymbol{p}))$ decreases exponentially so that the difference

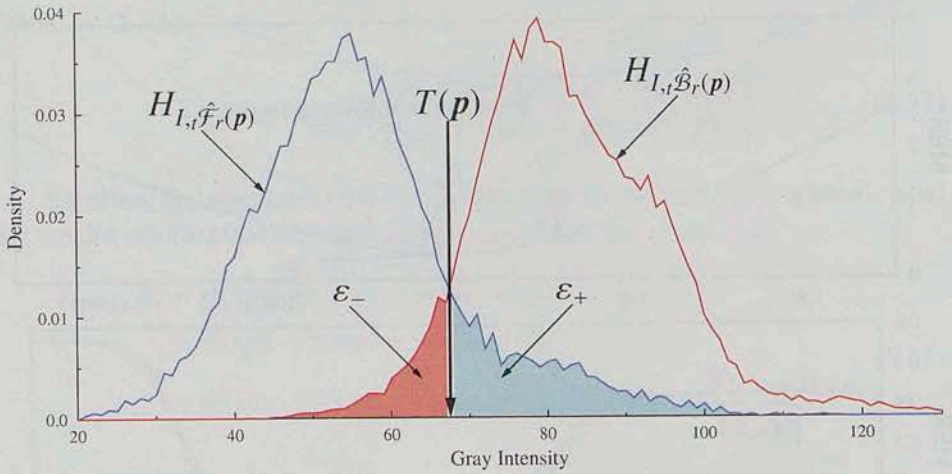


Figure 5.23 – $H_{I,t\hat{\mathcal{F}}_r(p)}$ *and* $H_{I,t\hat{\mathcal{B}}_r(p)}$ *of Fig. 5.6 (top-left) filtered by* $t_+ = 9$ *and* $t_- = -9$. *The area on blue (red) represents* ϵ_+ (ϵ_-) *which is the approximation of the probability of misclassifying foreground (background) pixels.*

between $\hat{t}_{opt}$ and $\hat{t}$ is approximately logarithmically proportional to the ratio w_f to $\hat{w}_f$.[1] Taking advantage of this property, $\hat{w}_f$ can be chosen as the upper bound of w_f without losing confidence that $\hat{t}$ approximates t^*_{opt}. Figure 5.24, for instance, shows that, even when the positive transition set is considerably overestimated and the negative transition set is considerably underestimated, MER estimates a similar threshold to $\hat{t}_{opt}$ (threshold by MER taking a complete form).

Remark 5.2: We say that the minimum-error-rate takes a complete form when $\hat{w}_f \approx \frac{|\mathcal{F}_r(p)|}{|\mathcal{P}_r(p)|}$. We say that the minimum-error-rate takes a **simple form** when $\hat{w}_f = 0.5$. Figure 5.23, for instance, shows the minimum-error-rate threshold (simple form) of Fig. 5.6.

In historical documents, r is usually chosen such that any character is completely contained in one or more neighborhood of radius r because, intuitively, the neighborhood of a character may preserve smoothness and high contrast. With such a radius, the foreground proportion is almost always less than 0.5 because letters, symbols, and lines are commonly printed with fine strokes. For example, Fig. 5.25 shows the cumulative distribution of the foreground proportion of

[1] See Section 5.5.5 for details of this argument.

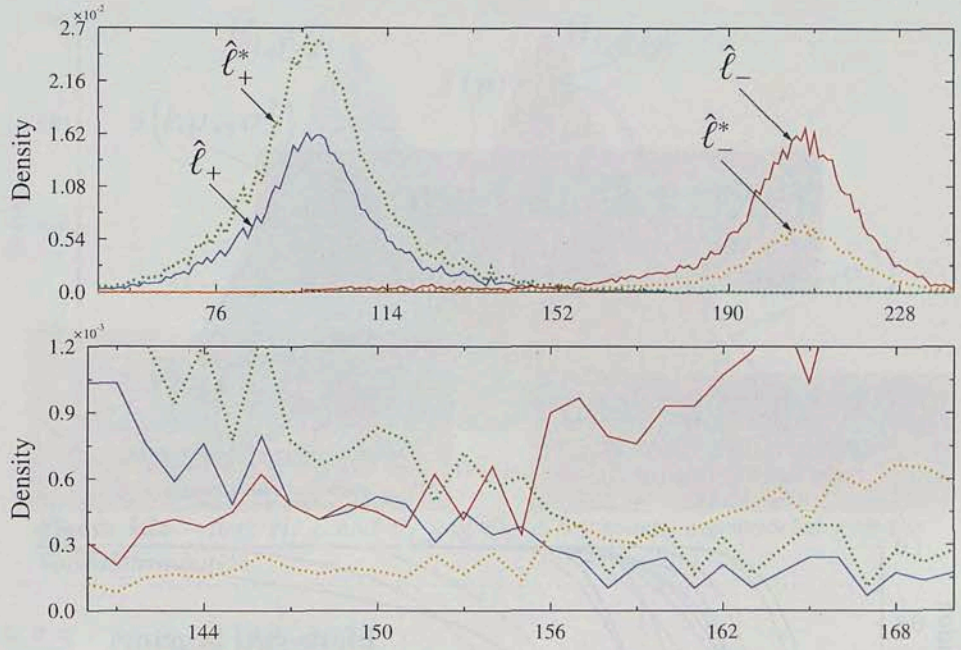


Figure 5.24 – *$\hat{\ell}_+$ and $\hat{\ell}_-$ correspond to the densities of ${}_t\mathcal{F}_r(\boldsymbol{p})$ and ${}_t\mathcal{B}_r(\boldsymbol{p})$, respectively. The line $\hat{\ell}^*_+$ overestimates $\hat{\ell}_+$ in the proportion 2:1. Conversely, $\hat{\ell}^*_-$ underestimates $\hat{\ell}_-$ in the proportion 1:2.*

14 images from historical documents; according to Gatos et al. [24], such images contain representative degradations which appear frequently (variable background intensity, shadows, smears, smudges, low contrast, and bleed-through) in historical documents.[2] In eleven images, any letter is completely contained in neighborhoods of radius $r = 50$, and 99.9% of those neighborhoods have a foreground proportion less than 0.5. In three images, however, there are letters which are only completely contained in neighborhoods of radius $r > 140$. Nevertheless, all three have foreground of proportions less than 0.5 in neighborhoods of radius $r > 140$. Hence, $w_f^+ = 0.50$ can be considered as the upper bound of w_f in historical documents for neighborhoods of radius r such that any letter in the image is completely contained in at least one neighborhood of radius r.

[2] This benchmark along with its groundtruth images can be found in http://users.iit.demokritos.gr/ bgat/DIBCO2009/benchmark/

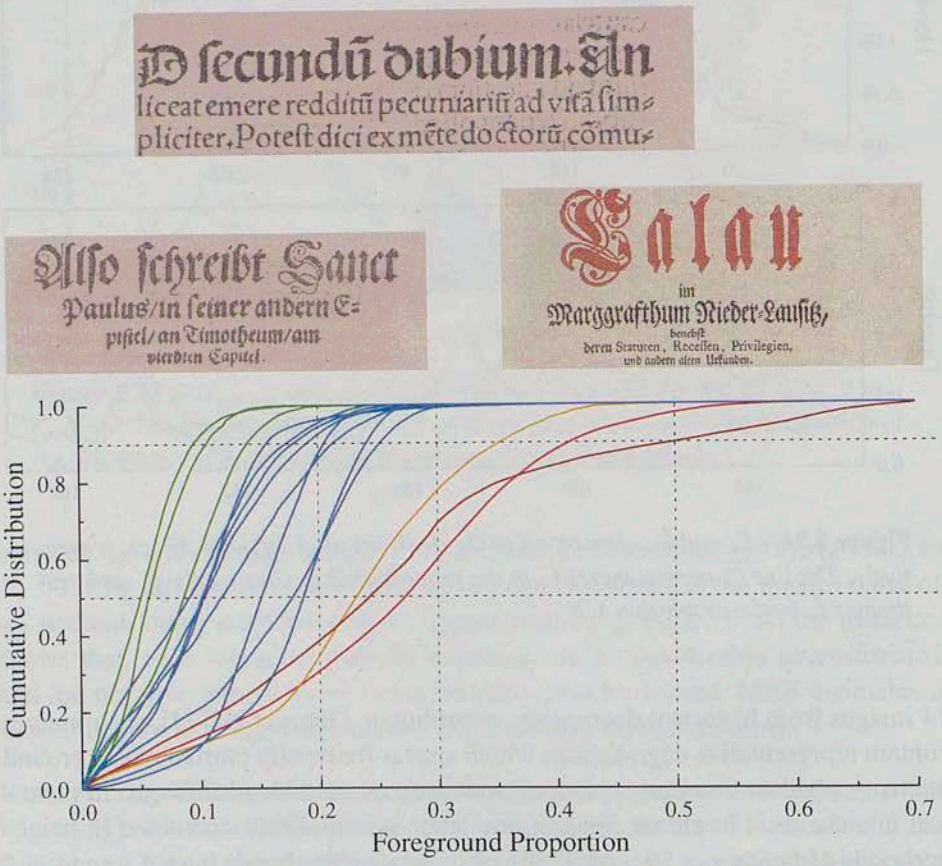


Figure 5.25 – *At the top, some images from the DIBCO 2009 benchmark. Foreground proportion ($r = 50$) of 14 images from historical documents. Curves of those cumulative distributed functions $F(x)$ such that $F(0.2) > 0.999$ are in green; those such that $F(0.3) > 0.999$ are in blue. Lines in orange, red and dark red correspond to the cumulative distribution functions of top, middle-left, and middle-right images, respectively.*

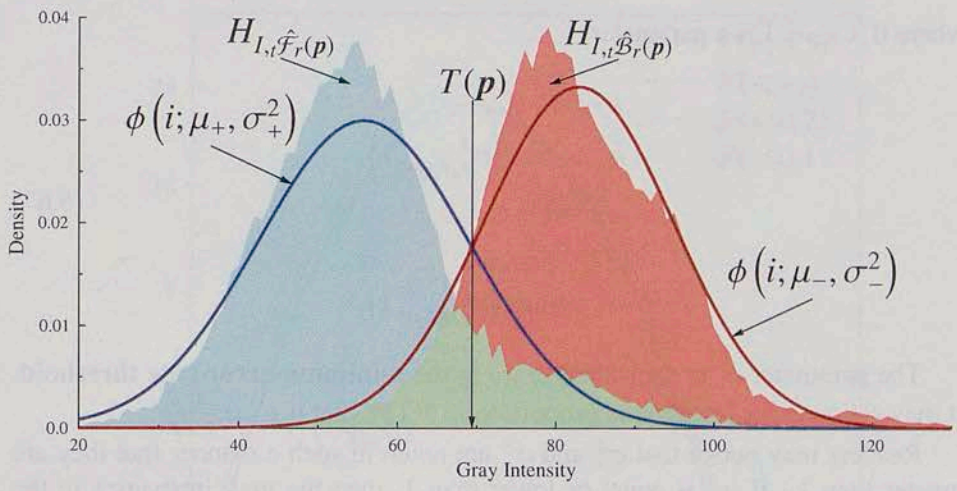


Figure 5.26 – *Both $H_{I,t\hat{\mathcal{F}}}$ and $H_{I,t\hat{\mathcal{B}}}$ of Fig. 5.6 (top-left) are modeled with the normal distribution.*

5.5.5 Normal threshold

I proposed the **normal threshold** in [72]; it assumes that the gray intensities of foreground obey a normal distribution; see Fig. 5.26. Thus,

$$\begin{aligned} H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) &\propto c_+\phi\left(i;\mu_+,\sigma_+^2\right) \\ H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) &\propto c_-\phi\left(i;\mu_-,\sigma_-^2\right) \end{aligned} \tag{5.64}$$

where $\phi(x;\mu,\sigma^2)$ denotes the probability density function of the normal distribution with mean μ and variance σ^2. Therefore, the intersection of these curves is given by the solution of the system

$$c_+\phi\left(i;\mu_+,\sigma_+^2\right) = c_-\phi\left(i;\mu_-,\sigma_-^2\right). \tag{5.65}$$

In the general case, (5.65) is a quadratic equation, and the threshold is the root $\mu_+ < \hat{t} < \mu_-$ of the quadratic equation with coefficients a, b and c given by

$$\begin{aligned} a &= \frac{1}{\sigma_+^2} - \frac{1}{\sigma_-^2} \\ b &= \frac{2\mu_-}{\sigma_-^2} - \frac{2\mu_+}{\sigma_+^2} \\ c &= \frac{\mu_+^2}{\sigma_+^2} - \frac{\mu_-^2}{\sigma_-^2} - 2\ln\left(\frac{\sigma_-\cdot c_+}{\sigma_+\cdot c_-}\right) \end{aligned} \tag{5.66}$$

where $0 < c_+ < 1$ is a parameter,

$$\begin{aligned}
\mu_+ &= \mu_{I_t \hat{\mathcal{F}}_r(\boldsymbol{p})}, \\
\sigma_+^2 &= \max\left(\sigma^2_{I_t \hat{\mathcal{F}}_r(\boldsymbol{p})}, 1\right), \\
c_- &= 1 - c_+, \\
\mu_- &= \mu_{I_t \hat{\mathcal{B}}_r(\boldsymbol{p})}, \\
\sigma_-^2 &= \max\left(\sigma^2_{I_t \hat{\mathcal{B}}_r(\boldsymbol{p})}, 1\right).
\end{aligned} \tag{5.67}$$

The parameter c_+ is equivalent to $\hat{w}_f$ in the **minimum-error-rate threshold**. It may estimate the foreground proportion in $\mathcal{P}_r(\boldsymbol{p})$, that is $c_+ \approx \frac{|\mathcal{F}_r(\boldsymbol{p})|}{|\mathcal{P}_r(\boldsymbol{p})|}$.

Readers may notice that σ_+^2 and σ_-^2 are taken in such a manner that they are greater than 1. If σ_+^2 is equal or lower than 1, then the gray intensities in the foreground are within $\mu_+ \pm 4$ since 99.99% of the values of the normal standard are within $[-4, 4]$. Then, the optimal threshold is $\mu_+ + 4$. A similar argument is given for σ_-^2.

Remark 5.3: We say that the normal threshold takes a **complete form** when $c_+ \approx \frac{|\mathcal{F}_r(\boldsymbol{p})|}{|\mathcal{P}_r(\boldsymbol{p})|}$; we say that the normal threshold takes a **simple form** when $c_+ = 0.5$.

Besides the general case in (5.65), there is a special case to solve when $\sigma_+ = \sigma_- = \sigma > 0$, which implies that $a = 0$. Thus, (5.65) has a unique solution given by

$$\hat{t} = \frac{\mu_+ + \mu_-}{2} - \frac{\sigma^2 \cdot \ln\left(\frac{c_-}{c_+}\right)}{\mu_- - \mu_+}. \tag{5.68}$$

Numerical error can arise if $\sigma_+ \approx \sigma_-$. Therefore, I also use (5.68) if $|\sigma_+ - \sigma_-| < 1$.

Assuming $a \neq 0$, the influence of c_+ on $T(\boldsymbol{p})$ can be analyzed with the symmetry of the **quadratic equation**. Let

$$\begin{aligned}
F(x; c_+) &= a \cdot x^2 + b \cdot x + \underbrace{\frac{\mu_+^2}{\sigma_+^2} - \frac{\mu_-^2}{\sigma_-^2} - 2\ln\left(\frac{\sigma_-}{\sigma_+}\right)}_{h} \underbrace{- 2\ln\left(\frac{c_+}{1 - c_+}\right)}_{k} \\
F(x; c_+) &= a \cdot x^2 + b \cdot x + h - k
\end{aligned} \tag{5.69}$$

be the quadratic equation for the normal threshold with parameter c_+. Thus, $F(x; c_+)$ has a vertical symmetry axis in $x = -\frac{b}{2a}$ for all c_+, as Fig. 5.27 shows.

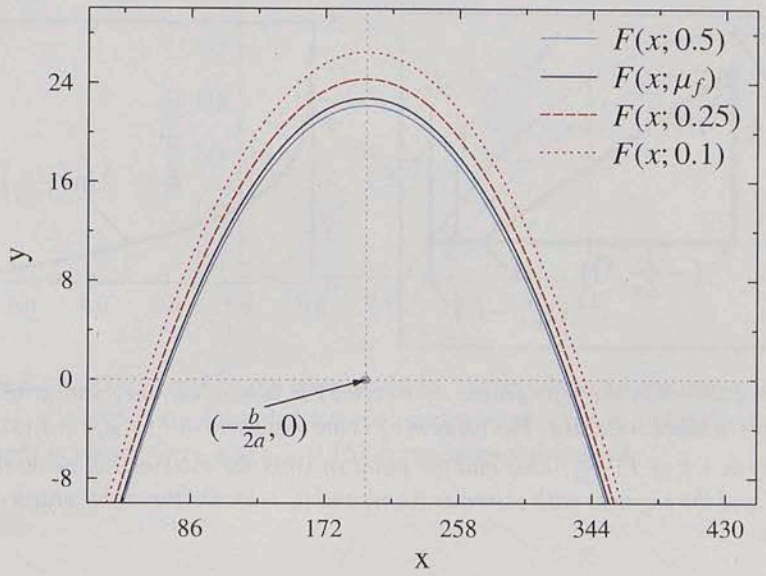


Figure 5.27 – *The normal threshold with parameter $\hat{\mu}_f$ is one root of $F(x;\hat{\mu}_f)$. In blue solid line, $F(x;\mu_f^+)$ which is the graph for the upper bound of $\hat{\mu}_f = \hat{\mu}_f^+ = 0.5$; in black solid line, $F(x;\mu_f)$ which is the graph for $\hat{\mu}_f$ equal to foreground proportion. $F(x;0.25)$ and $F(x;0.1)$ are show in dark-red dashed and red dotted lines, respectively.*

Without loss of generality, assume $c_+ < 0.5$, and $\sigma_+ > \sigma_- > 1$ such that $\sigma_+ - \sigma_- > 1$.

In Fig. 5.28 (left), t_c is the normal threshold in its complete form, t_s the normal threshold in its simple form, and the point $(0, n)$ is the intersection between the axis x and the segment with extremes $(0, m)$ and $(t_c, -k)$. The convexity of $F(x; c_+)$, guarantees that $n < t_s$. Then, by similarity of triangles

$$\begin{aligned} &\frac{m}{k} = \frac{n}{t_c - n} \\ \Rightarrow &\frac{m}{k} < \frac{t_s}{t_c - t_s} \\ \Rightarrow &t_c - t_s < \frac{k}{m} \cdot t_s \end{aligned} \tag{5.70}$$

where

$$m = F(-\frac{b}{2a};\mu_f^+) = -\frac{b^2}{4a} + h = \frac{[\mu_- - \mu_+]^2}{\sigma_-^2 - \sigma_+^2} - 2\ln\left(\frac{\sigma_-}{\sigma_+}\right) \tag{5.71}$$

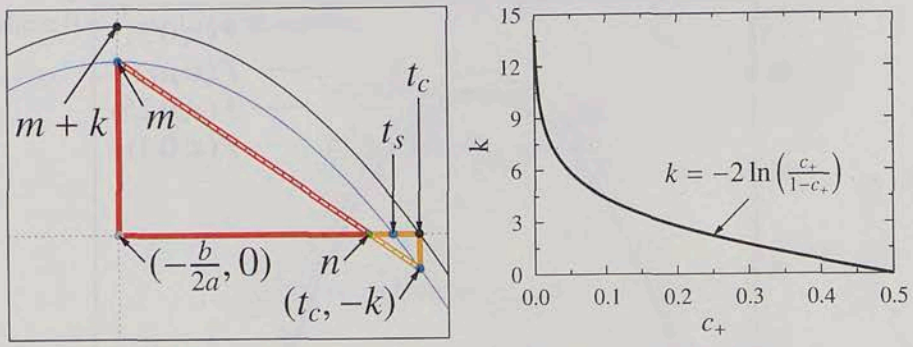


Figure 5.28 – *On the left, graphs of $F(x; 0.5)$ in blue solid line, and graphs of $F(x; \mu_f)$ in black solid line. The value $m+k$ is the maximum of $F(x; \mu_f^+) = F(x; 0.5)$, namely $m + k = F(-\frac{b}{2a}; 0.5)$; and the point $(n, 0)$ is the intersection between the axis x and the segment with extremes $(0, m)$ and $(t_c, -k)$. On the right, graph of k's values.*

is the maximum of $F(x; \mu_f^+) = F(x; 0.5)$.

The magnitude of k can be seen in Fig. 5.28 (right), which shows that $k < 15$ for $c_+ = 0.01$ and $k < 5$ for $c_+ = 0.1$. The magnitude of m, however, depends on the contrast of the image, and it can be calculated only if a, b, and h are known. Figure 5.29 shows $F(x; c_+)$ for the histogram of Fig. 5.21, where $m \approx 595$, and $|t_s - t_c| < 1.1$ for all $c_+ \in [0.1, 0.5]$.

Note that the difference between $\hat{t}_{opt}$ and t_c (or t_s) cannot be known since it depends on how well the distribution of gray intensities in ${}_t\hat{\mathcal{F}}_r(\boldsymbol{p})$ and ${}_t\hat{\mathcal{B}}_r(\boldsymbol{p})$ approximate the distribution of gray intensities of $\mathcal{F}_r(\boldsymbol{p})$ and $\mathcal{B}_r(\boldsymbol{p})$, respectively. In our example of Fig. 5.1, this difference is less than 8 gray levels which represents that the probability of error is ≈ 0.0066 at level t_s (the minimum probability of error is ≈ 0.0038). In fact, in this example, the minimum-error-rate threshold in simple form coincides with the normal threshold in simple form.

5.5.6 Lognormal threshold

I proposed the **lognormal threshold** in [72]; it assumes that the gray intensities of both foreground and background obey a lognormal distribution; see Fig. 5.30.

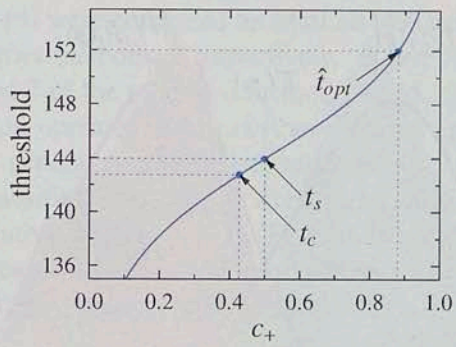


Figure 5.29 – *Graph of $F(x; c_+)$ computed from the transition set approximation of Fig. 5.1. $t_c \approx 145$ is the normal threshold in complete form, $t_s \approx 144$ is the normal threshold in simple form, and $t_{opt} \approx 152$ is the optimal threshold.*

That is,

$$\begin{aligned} H_{I,\mathcal{F}_r(\boldsymbol{p})}(i) &\propto c_+ \lambda\left(i; \tilde{\mu}_+, \tilde{\sigma}_+^2\right) \\ H_{I,\mathcal{B}_r(\boldsymbol{p})}(i) &\propto c_- \lambda\left(i; \tilde{\mu}_-, \tilde{\sigma}_-^2\right) \end{aligned} \tag{5.72}$$

where $\lambda\left(i; \tilde{\mu}, \tilde{\sigma}^2\right)$ denotes the lognormal probability density function with parameters μ and σ^2 which are the mean and variance of the variables natural logarithm, respectively.

The intersection of these curves is $\exp(\hat{t})$, where $\hat{t}$ is the root of the quadratic equation with coefficients given by (5.66), but replacing μ_+ and σ_+^2 with $\tilde{\mu}_+$ and $\tilde{\sigma}_+^2$ which are estimated using the relations:

$$\tilde{\mu}_+ = \ln\left(\mu_{I,i\hat{\mathcal{F}}_r(\boldsymbol{p})}\right) - \frac{1}{2}\tilde{\sigma}_+^2 \text{ and } \tilde{\sigma}_+^2 = \ln\left(1 + \frac{\sigma^2_{I,i\hat{\mathcal{F}}_r(\boldsymbol{p})}}{\left[\mu_{I,i\hat{\mathcal{F}}_r(\boldsymbol{p})}\right]^2}\right). \tag{5.73}$$

Likewise, $\hat{\mu}_-$ and $\hat{\sigma}_-^2$ are estimated.

5.6 Edge detection

In a binarization context, an **edge pixel** $\boldsymbol{p}$ can be defined as a foreground pixel that contains background pixels within $\mathcal{P}_1(\boldsymbol{p})$. Therefore, ${}_1\mathcal{F}$ is the set of edge pixels. Notice that an edge pixel $\boldsymbol{p}$ can be defined as the pixel that contains both

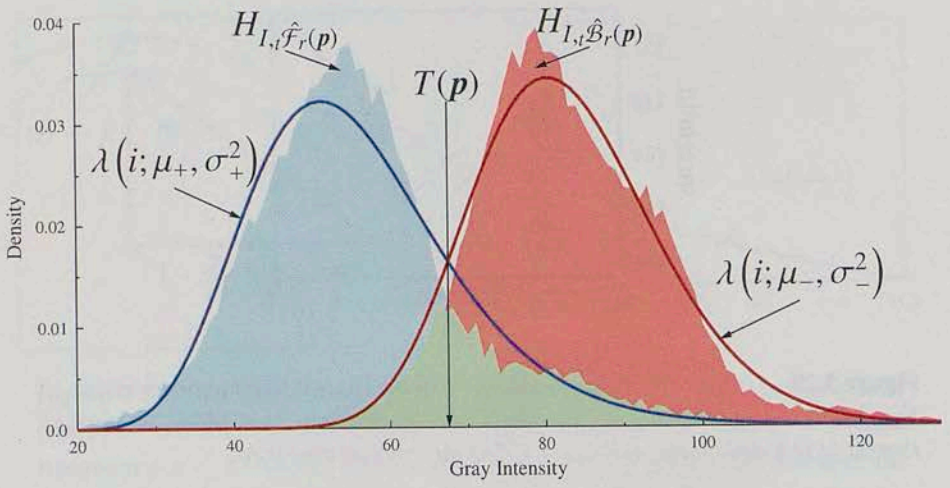


Figure 5.30 – *Both $H_{I,{}_t\hat{\mathcal{F}}}$ and $H_{I,{}_t\hat{\mathcal{B}}}$ of Fig. 5.6 (top-left) are modeled with the lognormal distribution.*

foreground and background pixels within $\mathcal{P}_1(\boldsymbol{p})$, or as those background pixels that contain foreground pixels in the neighborhood of radius 1. Nevertheless, I will use the former definition.

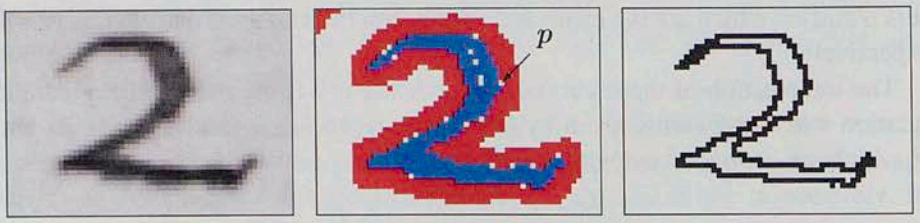


Figure 5.31 – *On the left, original image; in the center, transition set approximation; on the right, edge image by transition operator.*

We can approximate ${}_1\mathcal{F}$ by

$$ {}_1\hat{\mathcal{F}} = \{\boldsymbol{p} \mid \boldsymbol{p} \in {}_t\hat{\mathcal{F}} \text{ and } |{}_t\hat{\mathcal{B}}_1(\boldsymbol{p})| > 0\}. \tag{5.74}$$

The pixel $\boldsymbol{p}$ in Fig. 5.31 (Center), which belongs to ${}_t\hat{\mathcal{P}}^c$, can be considered an edge pixel since it is exactly between pixels in ${}_t\hat{\mathcal{F}}$ and ${}_t\hat{\mathcal{B}}$. Hence, I defined in [73] the **simple edge transition operator** as

$$ {}_1\hat{\mathcal{F}} = \{\boldsymbol{p} \mid 0 < |{}_t\hat{\mathcal{F}}_1(\boldsymbol{p})| \text{ and } |{}_t\hat{\mathcal{B}}_1(\boldsymbol{p})| > 0\}. \tag{5.75}$$

Figures 5.32 (b)-(d) were computed on MatLab [49] using Canny [6], Prewitt [47] and Roberts Cross methods [3], respectively. Figure 5.32 (f) was computed following steps 1 and 2 of the transition method (Fig. 5.32 (e)) and applying the simple edge transition operator. The raw transition set approximation (without restoration process) generates many false positives. In contrast, Fig. 5.32 (h), which follows the transition method with a restored transition set, reports a lower number of false negatives than Fig. 5.32 (f). Unfortunately, the combination of transition operators used in Fig. 5.32 (e) includes more than one cross, diagonal, and incidence transition operator in a non-trivial order:

- isolation transition operator (cross neighborhood),
- isolation transition operator (diagonal neighborhood),
- isolation transition operator (cross neighborhood),
- incidence transition operator ($k = 2$, $a = b = 2$),
- dilation transition operator ($a = b = 3$),
- isolation transition operator (cross neighborhood),
- isolation transition operator (diagonal neighborhood),
- isolation transition operator (cross neighborhood),
- rectangular isolation transition operator ($x = y = 2$), and
- incidence transition operator ($k = 2$, $a = b = 2$).

5.7 Summary

The second main contribution of my thesis is enclosed in this chapter. I described mathematically the transition method for binarization, and to a minor degree, for edge detection, and for detection of regions of interest.

Section 5.1 presents an overview of the transition method, where I pointed out that the positive transition set (intersection of foreground and transition set) and

[3]The default parameters of MatLab are chosen heuristically in a way that depends on the input data

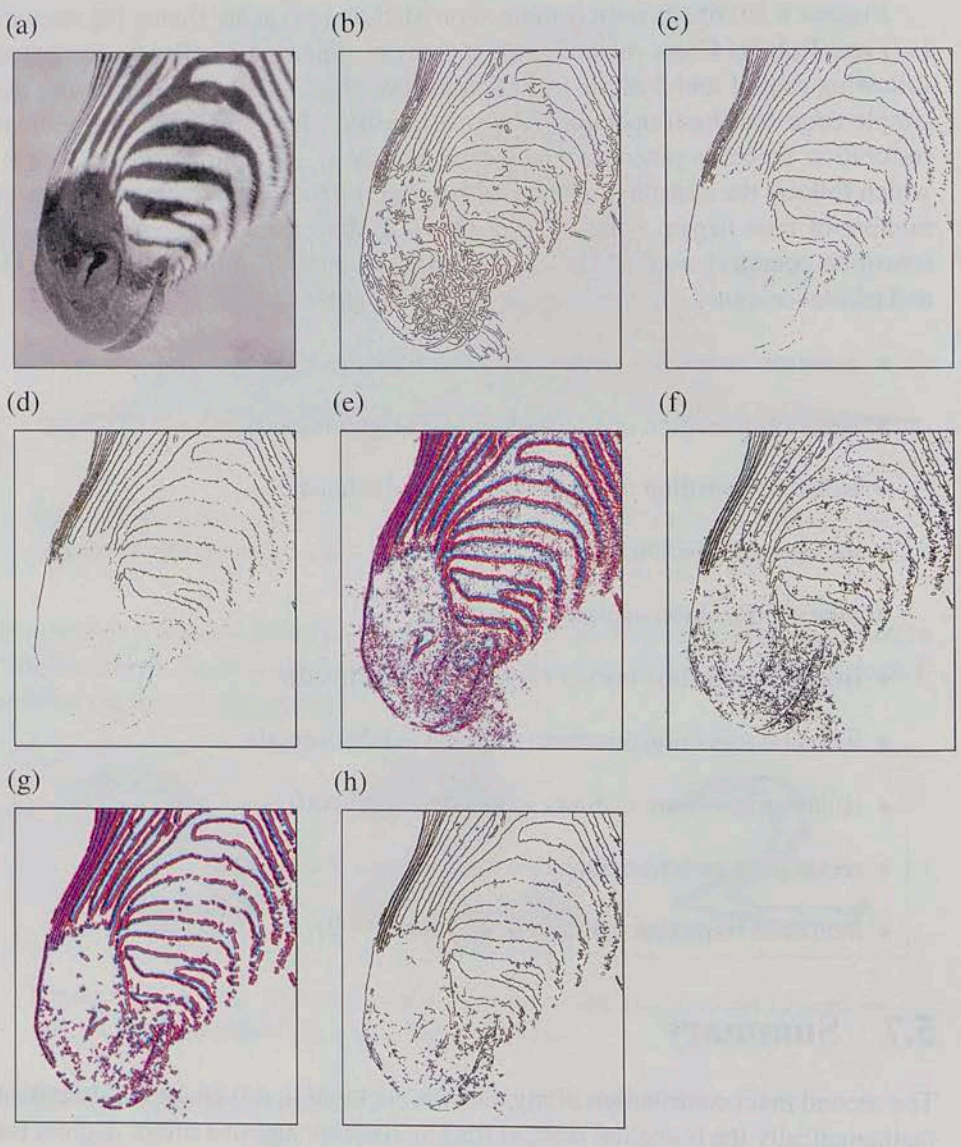


Figure 5.32 – *(a) Original image; (b) Edge image by Canny method. (c) Edge image by Prewitt method. (d) Edge image by Roberts method. (e) Raw transition image. (f) Edge image of (e) computed by the simple edge transition operator. (g) Restored transition image of (a). (h) Edge image of (g) computed by the simple edge transition operator.*

negative transition set (intersection of background and transition set) are representative samples of the foreground and background, respectively. Furthermore, I proposed that the transition set can be accurately approximated from pixels with high positive and negative transition values (transition values are computed with maxmin function).

The transition method is roughly divided into five parts: calculation of transition values, calculation of transition thresholds, restoration of transition sets, detection of regions of interest, and binarization (or edge detection).

In Section 5.2, I proposed three methods to compute transition thresholds based on the empirical complementary cumulative function of transition values: quantile transition threshold (Section 5.2.1), Rosin's transition threshold (Section 5.2.2), and double-linear transition threshold (Section 5.2.3). While the quantile transition threshold requires setting a parameter, both Rosin's and double-linear transition threshold have no parameters to set. In particular, the performance of both double-linear and Rosin's transition threshold are tested in Section 7.6 and in [71], respectively, showing comparable performance.

The restoration of the transition set is addressed in Section 5.3. It is defined as the process of adding and removing pixels from the transition set with the aim of increasing the cardinality while reducing the noise. Besides well-known morphological operators detailed in that section, I proposed two novel operators for restoring transition sets: incidence and dilation transition operators. The former removes pixels from the transition set, which cannot be removed with well-known morphological operators; see Section 5.3.3. The latter adds pixels without losing confidence in the transition set approximation, unlike the standard dilation morphological operator, which decreases confidence; see Section 5.3.4.

In Section 5.4, I proposed two simple criteria to detect regions of interest. The first criterion is based on the cardinality of the positive and negative transition set. Pixels whose neighborhood contains few positive and negative transition pixels are classified as background. The second criterion to discard outliers uses the difference between the means of gray intensities of the positive and negative transition set approximations.

I proposed five novel thresholdings based on transition sets in Section 5.5: linear mean-variance threshold (Section 5.5.1), autolinear threshold (Section 5.5.2), minimum-error-rate (Section 5.5.4), normal threshold (Section 5.5.5), and lognormal threshold (Section 5.5.6).

Although the lognormal, normal, and autolinear threshold outperform top-ranked algorithms, the lognormal threshold has performed the best; see Section 7.5, Section 7.6, and [71]. Such results strongly suggest that the positive and negative

transition approximations are lognormally distributed rather than normally distributed.

In Section 5.5.3, I proposed the minimum symmetric value, which attempts to measure the symmetry of a histogram. Minimum symmetric values can substitute the means in any threshold. Unfortunately, I did not explore the performance of binarization algorithms using this alternative technique.

The potential of the transition method for edge detection is shown in Section 5.6. In this section, I proposed a simple algorithm for edge detection based on pairs of transition pixels (one positive and one negative). The performance of this edge detector is closely related with the performance of the process of restoration of transition sets: The better the transition set approximation, the better the performance of the edge detector.

Chapter 6

Unsupervised evaluation measures

I do not know it for sure, I suppose it.

Jaimes Sabines Gutiérrez
Mexican poet (1926-1999)

Historical documents usually present several challenges and kinds of degradations, such as non-standard fonts, ink stains, weak ink strokes and wide variations in the background, to mention some. Because of this, the parameters of binarization algorithms have to be tuned for each kind of degradation. For a large set of images, however, the manual tuning of parameters is time-consuming and costly, and the use of *general parameters* may lead to a low binarization performance. Hence, the selection of binarization algorithms and their parameters play the most important role in the accuracy of recognition.

To address the problem of parameter selection in segmentation, unsupervised evaluation methods have been proposed to assess the quality of a segmentation [91], [92]. Such methods allow for evaluation of many algorithms over large parameter spaces and on diverse images without the need for human intervention.

Consequently, they enable an objective comparison of both different segmentation methods and the different parameters of a single method. Moreover, they can be used for automatic parameter choice of binarization algorithms.

Evaluation measures based on the variance of gray intensities have been used to assess binarization performance [72], [73], [79], [82]. Specially in document images, both foreground and background are intuitively thought of as uniform and homogeneous regions. Unfortunately, few authors have analyzed the mathematical and experimental behavior of these measures [8], [91], hence my interest to address the interaction between binarization methods and these evaluation measures. This interaction is analyzed under my proposed model of **simple images**, which are images where the contrast of gray intensities between foreground and background pixels is bounded in small neighborhoods. Ideal images provide the mathematical basis to prove whether the optimal value of each evaluation measure leads to the estimation of an accurate foreground.

6.1 Simple images

In general, the probability that a pixel with a certain intensity belongs to the foreground or background depends on their distributions, as I pointed out in Section 5, especially stressed by the pixels with intensities between $\mu_{I,\mathcal{F}_r(\boldsymbol{p})}$ and $\mu_{I,\mathcal{B}_r(\boldsymbol{p})}$. Thus, to minimize misclassification when using a threshold, it is better when these means are far apart and their variances are small, that is, when the contrast between foreground and background is large. This is illustrated in Fig. 6.1, where an image with good contrast is shown, and its histogram is compared with a hypothetical histogram that only differs in contrast (distance between means). The shaded region in green represents the probability of misclassified pixels according the Bayes rule.

In Section 4.1, I introduced the concept of **ideal image**. I also indicated that the contrast between the foreground and background in a neighborhood of interest not only depends on the variances, but also depends on the means of the gray intensities, more specifically, in the difference between such means. The smaller this difference, the higher the **minimum probability of error**; see Section 5.5.4.

Given that contrast is crucial for an accurate segmentation, certain bounds are required for it. I formalized this requirement with the following definition.

Definition 6.1: *Assuming Model 1 (Definition 4.1), an image is an* r***-simple image*** *if all neighborhoods with radius* r *such that* $|\mathcal{F}_r(\boldsymbol{p})| > 1$ *and* $|\mathcal{B}_r(\boldsymbol{p})| > 1$ *satisfy*

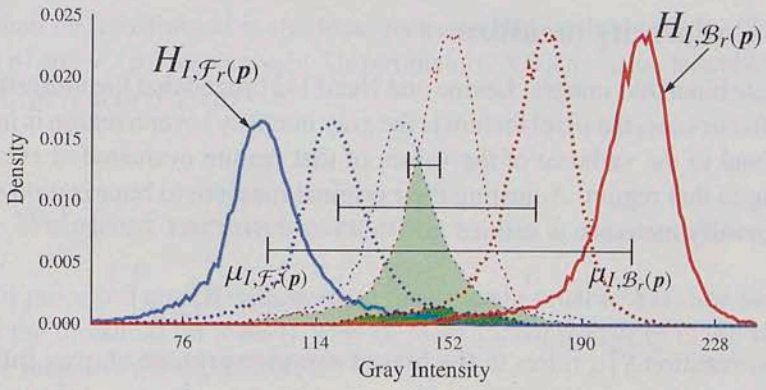


Figure 6.1 – *Example of "good" contrast in a neighborhood. In dash lines, hypothetical examples of "bad" contrast. Area filled in light (dark) green represents the minimum error given the dotted (dashed) histograms.*

the inequality:

$$\|\mu_{I,\mathcal{B}_r(\boldsymbol{p})} - \mu_{I,\mathcal{F}_r(\boldsymbol{p})}\| > \sqrt{2} \cdot \max(\sigma_{I,\mathcal{B}_r(\boldsymbol{p})}, \sigma_{I,\mathcal{F}_r(\boldsymbol{p})}) \tag{6.1}$$

where $\|\cdot\|$ *denotes the absolute value.*

I consider that the gray intensities of the foreground are darker than those in the background. That is, $\mu_{I,\mathcal{B}_r(\boldsymbol{p})} > \mu_{I,\mathcal{F}_r(\boldsymbol{p})}$.

6.2 Unsupervised binarization measures

A measure is useful if the better the binarization obtained, the smaller (larger) the measure on to which the segmented image evaluates. In particular, we would desire the minimum (maximum) of the measure to be attained only at the perfect segmentation $\hat{\mathcal{F}} = \mathcal{F}$.

In the following subsections, I will introduce local implementations of unsupervised measures. Because of that, the binarization performance over a whole image is the accumulation of the binarization performances over all neighborhoods with radius r in terms of a measure M_r. I denote this evaluation by $Eval(M_r, \hat{\mathcal{F}})$. That is,

$$Eval(M_r, \hat{\mathcal{F}}) = \sum_{\boldsymbol{p} \in \mathcal{P}} M_r(\boldsymbol{p}) \tag{6.2}$$

6.2.1 Uniformity measure

To evaluate binarized images, Levine and Nazif [42] stated that the uniformity of a feature (in our case, the pixel feature is the gray intensity) over a region is inversely proportional to the variance of the values of that feature evaluated at every pixel belonging to that region. Adjusting their original measure to binarization context, the **uniformity measure** is defined as

$$U = 1 - \frac{1}{w}\left[w_{\hat{\mathcal{F}}_r(\boldsymbol{p})} \cdot S^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})} + w_{\hat{\mathcal{B}}_r(\boldsymbol{p})} \cdot S^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}\right] \tag{6.3}$$

where the notation $S^2_{I,\mathcal{A}}$ refers to **the biased sample variance of gray intensities** (Appendix B), w_f and w_b are the weights associated to $\hat{\mathcal{F}}_r(\boldsymbol{p})$ and $\hat{\mathcal{B}}_r(\boldsymbol{p})$, respectively, and w is a normalization factor designed to limit the maximum value of the measure to one

$$w = \left[w_f + w_b\right] \cdot \frac{[I_{max} + I_{min}]^2}{2} \tag{6.4}$$

where I_{max} and I_{min} are the maximum and minimum gray intensities in $\mathcal{P}$.

Sahoo et al. [79] used a particular case of U with $w_f = w_b = 1$ to evaluate binarization methods. I simplified this particular case of U with the **gray-intensity uniformity measure** (GU)

$$GU_r = S^2_{I,\hat{\mathcal{F}}} + S^2_{I,\hat{\mathcal{B}}} \tag{6.5}$$

which is linearly equivalent to Sahoo et. al.'s evaluation measure.

Proposition 6.1. *Let $\mathcal{P}$ be an r-simple image. Then, the minimum of the expected value of $GU_r(\boldsymbol{p})$ is not necessarily reached for $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$ or $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$ (proof in Section 6.3.1).*

Proposition 6.1 indicates that GU_r does not lead to the best binarization for all r-simple images. What is more, if one wanted to minimize the expected value of $GU_r(\boldsymbol{p})$, then it could happen that the estimated background would *swallow* the foreground.

6.2.2 Region non-uniformity measure

Another measure derived from U is the **region non-uniformity measure** (NU), which was proposed by Sezgin and Sankur [82] as

$$NU = \frac{|\hat{\mathcal{F}}| \cdot S^2_{I,\hat{\mathcal{F}}}}{|\mathcal{P}| \cdot S^2_{I,\mathcal{P}}} \tag{6.6}$$

NU can be transformed in the local measure $NU_r(\boldsymbol{p})$ by replacing $\mathcal{P}$, and $\hat{\mathcal{F}}$, with $\mathcal{P}_r(\boldsymbol{p})$ and $\hat{\mathcal{F}}_r(\boldsymbol{p})$, respectively. Unfortunately, $NU_r(\boldsymbol{p})$ lacks desirable properties: $NU_r(\boldsymbol{p})$ is zero if $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \emptyset$.

6.2.3 Weighted variance measure

Otsu [66] proposed several discriminant measures in order to evaluate the "*goodness*" of the threshold (at level t). One of these global measures is the **weighted variance measure** (WV), defined as

$$WV = \frac{1}{|\mathcal{P}|}\left[|\hat{\mathcal{B}}| \cdot S^2_{I,\hat{\mathcal{B}}} + |\hat{\mathcal{F}}| \cdot S^2_{I,\hat{\mathcal{F}}}\right] \tag{6.7}$$

Remark 6.1: Ng and Lee [60] proved that WV is equivalent to U if $w_f = |\hat{\mathcal{F}}|$, $w_b = |\hat{\mathcal{B}}|$, and $w = |\mathcal{P}|$.

Let WV_r be the measure which replaces $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ with $\hat{\mathcal{F}}_r(\boldsymbol{p})$ and $\hat{\mathcal{B}}_r(\boldsymbol{p})$ in WV. Then,

Proposition 6.2. *In an r-simple image, the minimum of the expected value of WV_r is not necessarily reached for $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$ or $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$ (proof in Section 6.3.2).*

6.2.4 Uniform variance measure

I proposed the **uniform variance measure** (UV) in [72], which is defined with the local standard deviation of gray intensities as

$$UV_r(\boldsymbol{p}) = \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|}\left[|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \hat{\sigma}_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \hat{\sigma}_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}\right] \tag{6.8}$$

where the notation $\hat{\sigma}_{I,\mathcal{A}}$ refers to the **sample standard error of gray intensities**; see Appendix B.

6.2.5 Unbiased measures

To overcome the statistical bias of WV_r, I propose the **unbiased weighted variance measure** in Ramírez-Ortegón et al. [70], which is defined as

$$\widehat{WV}_r(\boldsymbol{p}) = \begin{cases} \dfrac{|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}}{|\mathcal{P}_r(\boldsymbol{p})|} & \text{if } |\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2 \text{ and } |\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2. \\ \hat{\sigma}^2_{I,\mathcal{P}_r(\boldsymbol{p})} & \text{otherwise} \end{cases} \tag{6.9}$$

Theorem 6.1. *In an r-simple image, the expected value of the unbiased weighted variance measure is minimal if $\hat{\mathcal{F}} = \mathcal{F}$ or $\hat{\mathcal{F}} = \mathcal{B}$; see proof in Section 6.3.3.*

Corollary 6.1. *In an r-simple image, if r is such that $|\mathcal{B}_r(\boldsymbol{p})|$, $|\mathcal{F}_r(\boldsymbol{p})| \geq 1$ and $\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}$, $\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} > 0$ for all $\boldsymbol{p} \in \mathcal{P}$, then*

$$Eval\left(\widehat{WV}_r, \mathcal{F}\right) < Eval\left(\widehat{WV}_r, \hat{\mathcal{F}}\right) \tag{6.10}$$

for all $\hat{\mathcal{F}} \neq \mathcal{B}$ and $\hat{\mathcal{F}} \neq \mathcal{F}$; see proof in Section 6.3.4.

6.2.6 Measures based on logarithms

Assuming that the gray intensities of both foreground and background are lognormally distributed, we derived the measures $\widetilde{WV}_r(\boldsymbol{p})$ and $\widetilde{UV}_r(\boldsymbol{p})$ from $\widehat{WV}_r(\boldsymbol{p})$ and $\widehat{UV}_r(\boldsymbol{p})$. These measures replace $\hat{\sigma}^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\hat{\sigma}^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}$ with $\tilde{\sigma}^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\tilde{\sigma}^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}$, respectively, which are the **unbiased sample variance of gray-intensity logarithm** of the foreground and background, see Appendix B.

6.3 Proof of theorems and propositions

I list some basic propositions that are useful in the subsequent discussion. Some of these proofs use standard techniques and are omitted.

Proposition 6.3. *Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be a sample of n independent and identically distributed random variables with finite variance σ^2. Then the estimators $S^2_{\mathcal{A}}$ and $\hat{\sigma}^2_{\mathcal{A}}$ satisfy*

$$E(S^2_{\mathcal{A}}) = \frac{|\mathcal{A}| - 1}{|\mathcal{A}|}\sigma^2 \tag{6.11}$$

and

$$E(\hat{\sigma}^2_{\mathcal{A}}) = \sigma^2. \tag{6.12}$$

Proposition 6.4. *A random variable a with finite expected value and variance satisfies*

$$E(a^2) = Var(a) + [E(a)]^2 . \tag{6.13}$$

Proposition 6.5. *Let* $x_i \sim N(\mu_x, \sigma_x^2)$ *for* $i = 1, \ldots, n-m$ *and* $y_i \sim N(\mu_y, \sigma_y^2)$ *for* $i = 1, \ldots, m$. *be independent random variables. Consider* $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, *where* $\mathcal{X} = \{x_i\}_{i=1}^{n-m}$ *and* $\mathcal{Y} = \{y_i\}_{i=1}^{m}$. *Then,*

$$E(\hat{\mu}_{\mathcal{Z}}) = \mu_x + \frac{m}{n}\left[\mu_y - \mu_x\right], \tag{6.14}$$

and

$$Var(\hat{\mu}_{\mathcal{Z}}) = \frac{1}{n}\sigma_x^2 + \frac{m}{n^2}\left[\sigma_y^2 - \sigma_x^2\right]. \tag{6.15}$$

Lemma 6.1. *In an r-simple image, if* $|\mathcal{A}_r(\boldsymbol{p})| = n > 1$, $|\mathcal{A}_r(\boldsymbol{p}) \cap \mathcal{F}| = h$ *and* $n \geq 2h$, *then*

$$E(\hat{\sigma}^2_{I,\mathcal{A}_r(\boldsymbol{p})}) \geq \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{h}{n}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} \tag{6.16}$$

Likewise, if $|\mathcal{A}_r(\boldsymbol{p})| = m > 1$, $|\mathcal{A}_r(\boldsymbol{p}) \cap \mathcal{B}| = k$ *and* $m \geq 2k$, *then*

$$E(\hat{\sigma}^2_{I,\mathcal{A}_r(\boldsymbol{p})}) \geq \sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} + \frac{k}{m}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} \tag{6.17}$$

Proof. By mathematical convenience, we prove Lemma 6.1 for (6.17). Denote $\mathcal{X} = \mathcal{A}_r(\boldsymbol{p}) \cap \mathcal{F}$ and $\mathcal{Y} = \mathcal{A}_r(\boldsymbol{p}) \cap \mathcal{B}$, then

$$\begin{aligned} E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(\boldsymbol{p})}\right) &= E\left(\frac{1}{m-1}\left[\sum_{\boldsymbol{p}\in\mathcal{X}} I^2(\boldsymbol{p}) + \sum_{\boldsymbol{p}\in\mathcal{Y}} I^2(\boldsymbol{p}) - m\cdot\hat{\mu}^2_{I,\mathcal{A}_r(\boldsymbol{p})}\right]\right) \\ &= \frac{1}{m-1}\left[\sum_{\boldsymbol{p}\in\mathcal{X}} E\left(I^2(\boldsymbol{p})\right) + \sum_{\boldsymbol{p}\in\mathcal{Y}} E\left(I^2(\boldsymbol{p})\right) - m\cdot E\left(\hat{\mu}^2_{I,\mathcal{A}_r(\boldsymbol{p})}\right)\right]. \end{aligned} \tag{6.18}$$

Due Proposition 6.4 and Proposition 6.5

$$\begin{aligned}
E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(p)}\right) &= \frac{[m-k]}{m-1}\left[\sigma^2_{I,\mathcal{F}_r(p)} + \mu^2_{I,\mathcal{F}_r(p)}\right] + \frac{k}{m-1}\left[\sigma^2_{I,\mathcal{B}_r(p)} + \mu^2_{I,\mathcal{B}_r(p)}\right] \\
&\quad - \frac{m}{m-1}\left[Var\left(\hat{\mu}_{I,\mathcal{A}_r(p)}\right) + \left[E\left(\hat{\mu}_{I,\mathcal{A}_r(p)}\right)\right]^2\right] \\
E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(p)}\right) &= \frac{m}{m-1}\left[\sigma^2_{I,\mathcal{F}_r(p)} + \mu^2_{I,\mathcal{F}_r(p)}\right] \\
&\quad + \frac{k}{m-1}\left[\sigma^2_{I,\mathcal{B}_r(p)} - \sigma^2_{I,\mathcal{F}_r(p)} + \mu^2_{I,\mathcal{B}_r(p)} - \mu^2_{I,\mathcal{F}_r(p)}\right] \\
&\quad - \frac{m}{m-1}\left[\frac{1}{m}\sigma^2_{I,\mathcal{F}_r(p)} + \frac{k}{m^2}\left[\sigma^2_{I,\mathcal{B}_r(p)} - \sigma^2_{I,\mathcal{F}_r(p)}\right]\right] \\
&\quad - \frac{m}{m-1}\left[\mu_{I,\mathcal{F}_r(p)} + \frac{k}{m}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]\right]^2 .
\end{aligned} \tag{6.19}$$

Reducing terms, we yield

$$\begin{aligned}
E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(p)}\right) &= \sigma^2_{I,\mathcal{F}_r(p)} + \frac{k}{m}\left[\sigma^2_{I,\mathcal{B}_r(p)} - \sigma^2_{I,\mathcal{F}_r(p)}\right] + \frac{k}{m-1}\left[\mu^2_{I,\mathcal{B}_r(p)} - \mu^2_{I,\mathcal{F}_r(p)}\right] \\
&\quad - \frac{2k\cdot\mu_{I,\mathcal{F}_r(p)}}{m-1}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right] - \frac{k^2}{m[m-1]}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]^2
\end{aligned} \tag{6.20}$$

Observe that

$$\begin{aligned}
\mu^2_{I,\mathcal{B}_r(p)} - \mu^2_{I,\mathcal{F}_r(p)} &= \left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]\left[\mu_{I,\mathcal{B}_r(p)} + \mu_{I,\mathcal{F}_r(p)}\right] \\
&= \left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)} + 2\mu_{I,\mathcal{F}_r(p)}\right] \\
&= \left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]^2 + 2\mu_{I,\mathcal{F}_r(p)}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right],
\end{aligned} \tag{6.21}$$

replacing (6.21) in (6.20)

$$E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(p)}\right) = \sigma^2_{I,\mathcal{F}_r(p)} + \frac{k}{m}\sigma^2_{I,\mathcal{B}_r(p)} - \frac{k}{m}\sigma^2_{I,\mathcal{F}_r(p)} + \frac{k[m-k]}{m[m-1]}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]^2, \tag{6.22}$$

We have the following inequality using (6.1) and $\frac{m-k}{m-1} \geq \frac{1}{2}$:

$$\frac{k[m-k]}{m[m-1]}\left[\mu_{I,\mathcal{B}_r(p)} - \mu_{I,\mathcal{F}_r(p)}\right]^2 \geq \frac{k}{m}\cdot\frac{1}{2}\left[\sqrt{2}\cdot\sigma_{I,\mathcal{F}_r(p)}\right]^2 \geq \frac{k}{m}\sigma^2_{I,\mathcal{F}_r(p)} \tag{6.23}$$

We conclude our proof by replacing (6.23) in (6.22)

$$E\left(\hat{\sigma}^2_{I,\mathcal{A}_r(p)}\right) \geq \sigma^2_{I,\mathcal{F}_r(p)} + \frac{k}{m}\sigma^2_{I,\mathcal{B}_r(p)} - \frac{k}{m}\sigma^2_{I,\mathcal{F}_r(p)} + \frac{k}{m}\sigma^2_{I,\mathcal{F}_r(p)}. \tag{6.24}$$

□

6.3.1 Proof of Proposition 6.1

By mathematical convenience, I prove Proposition 6.1 using unbiased variance instead of the biased variance (these proofs differ only by factors); I will show an example where the background "*swallows*" the foreground.

Assume that there are more background pixels than foreground pixels in $\mathcal{P}_r(\boldsymbol{p})$. That is, $h = |\mathcal{F}_r(\boldsymbol{p})|$, $n = |\mathcal{P}_r(\boldsymbol{p})|$, and $2h < n$. Also assume that $\hat{\sigma}^2_{I,\mathcal{F}_r(\boldsymbol{p})} \geq \hat{\sigma}^2_{I,\mathcal{B}_r(\boldsymbol{p})}$, and that

$$\left[\mu_{I,\mathcal{B}_r(\boldsymbol{p})} - \mu_{I,\mathcal{F}_r(\boldsymbol{p})}\right]^2 = 2\frac{n-1}{n-h}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}, \tag{6.25}$$

Then, we can derive from Lemma 6.1

$$E\left(\hat{\sigma}^2_{I,\mathcal{P}_r(\boldsymbol{p})}\right) = \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{h}{n}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} + \frac{h}{n}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} \tag{6.26}$$

$$E\left(\hat{\sigma}^2_{I,\mathcal{P}_r(\boldsymbol{p})}\right) = \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{h}{n}[\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} + \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}] < \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} \tag{6.27}$$

Therefore,

$$E\left(\hat{\sigma}^2_{I,\mathcal{P}_r(\boldsymbol{p})}\right) < E\left(\hat{\sigma}^2_{I,\mathcal{F}_r(\boldsymbol{p})} + \hat{\sigma}^2_{I,\mathcal{B}_r(\boldsymbol{p})}\right), \tag{6.28}$$

which means that GU_r evaluates better $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \emptyset$ than $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$ when $\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} \geq \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}$, half or fewer pixels are foreground in the neighborhood of interest and (6.25) holds.

6.3.2 Proof of Proposition 6.2

To prove Proposition 6.2, it is enough to show a counterexample. Assume that $\mathcal{B}_r(\boldsymbol{p}) = \mathcal{P}_r(\boldsymbol{p})$. If $\hat{\mathcal{B}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$, by Proposition 6.3 we yield

$$E\left(WV_r\right) = \frac{|\mathcal{B}_r(\boldsymbol{p})| \cdot E\left(S^2_{I,\mathcal{B}_r(\boldsymbol{p})}\right)}{|\mathcal{B}_r(\boldsymbol{p})|} = \frac{|\mathcal{B}_r(\boldsymbol{p})| - 1}{|\mathcal{B}_r(\boldsymbol{p})|}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}, \tag{6.29}$$

but if $\hat{\mathcal{B}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})\backslash\{\boldsymbol{q}\}$ where $\boldsymbol{q} \in \mathcal{B}_r(\boldsymbol{p})$, then

$$\begin{aligned} E\left(WV_r\right) &= \frac{(|\mathcal{B}_r(\boldsymbol{p})| - 1)\, E\left(S^2_{I,\mathcal{B}_r(\boldsymbol{p})}\right)}{|\mathcal{B}_r(\boldsymbol{p})|} = \left[\frac{|\mathcal{B}_r(\boldsymbol{p})| - 1}{|\mathcal{B}_r(\boldsymbol{p})|}\right]^2 \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} \\ &< \frac{|\mathcal{B}_r(\boldsymbol{p})| - 1}{|\mathcal{B}_r(\boldsymbol{p})|}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}. \end{aligned} \tag{6.30}$$

6.3.3 Proof of Theorem 6.1

We need to prove

$$|\mathcal{B}_r(\boldsymbol{p})| \cdot \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})} + |\mathcal{F}_r(\boldsymbol{p})| \cdot \sigma^2_{l,\mathcal{F}_r(\boldsymbol{p})} \leq E\left(|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{l,\hat{\mathcal{B}}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{l,\hat{\mathcal{F}}_r(\boldsymbol{p})}\right) \quad (6.31)$$

for all $\boldsymbol{p}$.

The proof is divided into several cases which depend on how the neighborhood $\mathcal{P}_r(\boldsymbol{p})$, $\hat{\mathcal{F}}_r(\boldsymbol{p})$ and $\hat{\mathcal{B}}_r(\boldsymbol{p})$ are constituted:

- ■ Case A: $|\mathcal{F}_r(\boldsymbol{p})| = 0$.
  - • A.I: $|\hat{\mathcal{F}}_r(\boldsymbol{p})| = 1$.
    - ◦ Symmetric case of A.I: $|\hat{\mathcal{B}}_r(\boldsymbol{p})| = 1$.
  - • A.II: $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2$ and $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2$.
- □ Symmetric case of A: $|\mathcal{B}_r(\boldsymbol{p})| = 0$.
- ■ Case B: $0 < |\mathcal{F}_r(\boldsymbol{p})| \leq |\mathcal{B}_r(\boldsymbol{p})|$.
  - • B.I: $|\hat{\mathcal{F}}_r(\boldsymbol{p})| = 1$.
    - ◦ Symmetric case of B.I: $|\hat{\mathcal{B}}_r(\boldsymbol{p})| = 1$.
  - • B.II: $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2$ and $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2$.
    - – Case B.II.1: $|\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{F}| \geq |\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{B}|$ and $|\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{F}| \leq |\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{B}|$.
    - - Symmetric case of B.II.1.
    - – Case B.II.2: $|\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{F}| \leq |\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{B}|$ and $|\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{F}| \leq |\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{B}|$.
- □ Symmetric case of B: $0 < |\mathcal{B}_r(\boldsymbol{p})| \leq |\mathcal{F}_r(\boldsymbol{p})|$.

Case A: $|\mathcal{F}_r(\boldsymbol{p})| = 0$. I will prove that $E\left(\widehat{WV}_r(\boldsymbol{p})\right) = \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})}$ for any partition $\hat{\mathcal{B}}_r(\boldsymbol{p})$ and $\hat{\mathcal{F}}_r(\boldsymbol{p})$.

Case A.I: $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2$ and $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2$. Thence,

$$E\left(\widehat{WV}_r(\boldsymbol{p})\right) = \frac{|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})}}{|\mathcal{P}_r(\boldsymbol{p})|} = \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})} \quad (6.32)$$

Case A.II: If $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \leq 1$ (or $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \leq 1$), then (6.9) is defined as

$$E\left(\widehat{WV}_r(\boldsymbol{p})\right) = \sigma^2_{l,\mathcal{P}_r(\boldsymbol{p})} = \sigma^2_{l,\mathcal{B}_r(\boldsymbol{p})} \quad (6.33)$$

Case B: $|\mathcal{B}_r(\boldsymbol{p})|$, $|\mathcal{F}_r(\boldsymbol{p})| \geq 1$. There are two symmetrical cases: $|\mathcal{B}_r(\boldsymbol{p})| \geq |\mathcal{F}_r(\boldsymbol{p})|$ and $|\mathcal{B}_r(\boldsymbol{p})| \geq |\mathcal{F}_r(\boldsymbol{p})|$. We will only prove the former case.

Case B.I: $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \leq 1$ (or $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \leq 1$). Based on Lemma 6.1, direct calculus

Table 6.1 – *The case B.II is divided into three sub-cases according to m, n, k and h, where $|\hat{\mathcal{F}}_r(\boldsymbol{p})| = m$, $|\hat{\mathcal{B}}_r(\boldsymbol{p})| = n$, $|\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{B}| = k$, and $|\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{F}| = h$.*

	$m \geq 2k$	$m < 2k$
$n \geq 2h$	Case B.II.1	Case B.II.2
$n < 2h$	Case B.II.3	-

yields

$$\begin{aligned} E\left(\widehat{WV}_r(\boldsymbol{p})\right) = \sigma^2_{I,\mathcal{P}_r(\boldsymbol{p})} &\geq \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{|\mathcal{F}_r(\boldsymbol{p})|}{|\mathcal{P}_r(\boldsymbol{p})|}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} \\ &\geq \frac{|\mathcal{B}_r(\boldsymbol{p})| \cdot \sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + |\mathcal{F}_r(\boldsymbol{p})| \cdot \sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}}{|\mathcal{P}_r(\boldsymbol{p})|} \end{aligned} \tag{6.34}$$

Case B.II: $|\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2$ and $|\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2$. We have three sub-cases summarized in Table 6.1. Observe that case B.II.3 is the symmetrical case of B.II.2.

Case B.II.1: $n \geq 2 \cdot h$ and $m \geq 2 \cdot k$. It follows that $E\left(\hat{\sigma}^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}\right)$ satisfies (6.16), while $E\left(\hat{\sigma}^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}\right)$ satisfies (6.17). Therefore,

$$\begin{aligned} E(\widehat{WV}_r(\boldsymbol{p})) \geq& \frac{n}{n+m}\left[\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{h}{n}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}\right] + \frac{m}{n+m}\left[\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} + \frac{k}{m}\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}\right] \\ \geq& \frac{[n+k]\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + [m+h]\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}}{n+m} \end{aligned} \tag{6.35}$$

The number of background and foreground pixels can be computed in terms of n, h, m and k as: $|\mathcal{B}_r(\boldsymbol{p})| = n - h + k$ and $|\mathcal{F}_r(\boldsymbol{p})| = m - k + h$. Then,

$$\begin{aligned} [n+k]\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} &\geq [n+k-h]\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} \\ [m+h]\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} &\geq [m+h-k]\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} \end{aligned} \tag{6.36}$$

Case B.II.2: $n \geq 2h$ and $m < 2k$. Hence, both $E\left(\hat{\sigma}^2_{I,\hat{\mathcal{B}}_r(\boldsymbol{p})}\right)$ and $E\left(\hat{\sigma}^2_{I,\hat{\mathcal{F}}_r(\boldsymbol{p})}\right)$ satisfy (6.16).

$$\begin{aligned} E(\widehat{WV}_r(\boldsymbol{p})) \geq& \frac{n}{n+m}\left[\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{h}{n}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}\right] + \frac{m}{n+m}\left[\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + \frac{k}{m}\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}\right] \\ \geq& \frac{[n+m]\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + [k+h]\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}}{n+m} \\ \geq& \frac{[n+m-h-k]\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})} + [k+h]\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})}}{n+m} \end{aligned} \tag{6.37}$$

We conclude (6.31) holds because, in this case, $|\mathcal{B}_r(\boldsymbol{p})| = n + m - h - k$ and $|\mathcal{F}_r(\boldsymbol{p})| = k + h$.

6.3.4 Proof of Corollary 6.1

The premises $|\mathcal{B}_r(\boldsymbol{p})|$, $|\mathcal{F}_r(\boldsymbol{p})| \geq 1$ restrict our analysis to case B of Theorem 6.1's proof (without considering permutations). Moreover, (6.34), (6.35) and (6.37) are strict inequalities if $|\hat{\mathcal{B}}_r(\boldsymbol{p}) \cap \mathcal{F}| = h > 0$ or $|\hat{\mathcal{F}}_r(\boldsymbol{p}) \cap \mathcal{B}| = k > 0$ because Corollary 6.1 assumes $\sigma^2_{I,\mathcal{B}_r(\boldsymbol{p})}$, $\sigma^2_{I,\mathcal{F}_r(\boldsymbol{p})} > 0$.

6.4 Summary

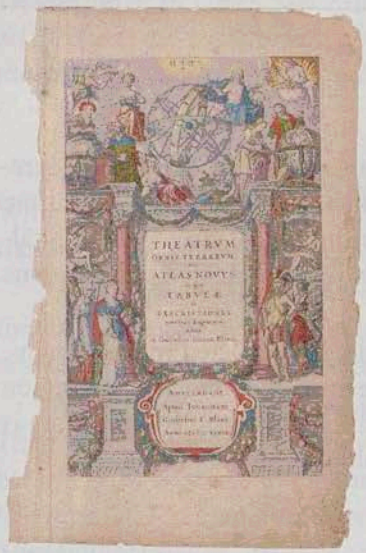
The third main contribution of my thesis is the mathematical analysis for all unsupervised measures described in this chapter. Given that contrast is crucial for an accurate segmentation, I introduced in Section 6.1 the concept of simple images (Definition 6.1). Such images satisfy a certain lower inequality between contrast and variance of gray intensities. Simple images are used throughout this chapter to analyze the optimality of unsupervised measures based on gray variances.

In Section 6.2, local implementations of three well-known unsupervised measures are discussed and analyzed: uniformity measure (Section 6.2.1), region non uniformity measure (Section 6.2.2), and weighted variance measure (Section 6.2.3). Later on, I proposed four novel unsupervised measures: the uniform variance measure (Section 6.2.4), based on the standard deviation of gray intensities; the unbiased weighted variance measure (Section 6.2.5), which overcomes the statistical bias of the weighted variance measure; and two measures based on logarithms of gray intensities (Section 6.2.6).

Theorem 6.1 is to be noted because it ensures that the expected value of the unbiased weighted variance measure is minimum in a perfect binarization, unlike the rest of the examined measures, which lack this property.

Chapter 7

Experimental comparison studies



The good Christian should beware of mathematicians, and all those who make empty prophecies. The danger already exists that the mathematicians have made a covenant with the devil to darken the spirit and to confine man in the bonds of Hell.

DeGenesi ad Litteram, Book II, xviii, 37 by Aurelius Augustinus Hipponensis (St. Augustine)
Bishop of Hippo Regius (354 – 430)

In this chapter, I summarize the results of my experiments, in which I used the same test images. Most of my conclusions are based on **pairwise comparisons** since the **uncertainty test** can ascertain which binarization algorithm is better given an intuitive triad of possible results: better, worse or comparable performance. A full explanation of the use of pairwise comparison is given in Appendix C.

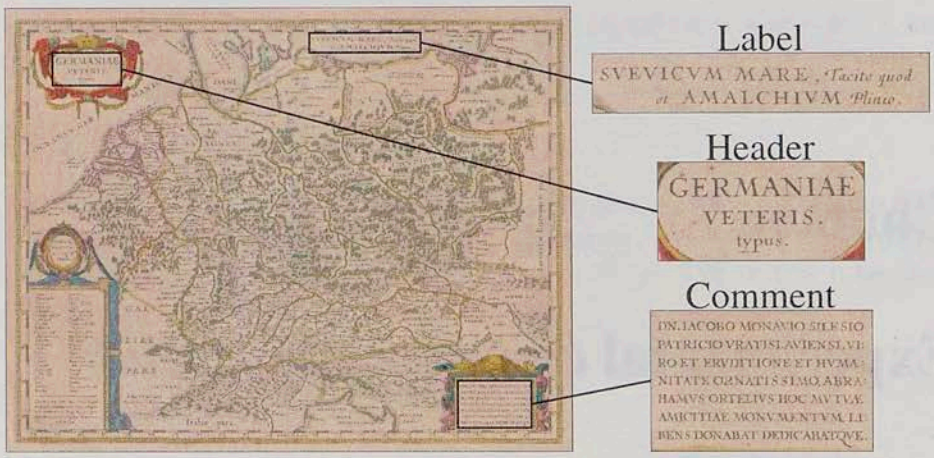


Figure 7.1 – *Example of map which contains a header, label and comment.*

7.1 Test Images

Historical documents usually present several challenges and varied forms of degradation, such as ink stains, smears, weak ink strokes and wide variations in the background. Because of this, the binarization algorithms were tested with digitalized images of the historical atlas **Theatrum orbis terrarum, sive, Atlas novus** (Blaeu Atlas)[1] at 150 **dpi** resolution.

Dots per inch (dpi) is a measure of spatial printing or video dot density, in particular, the number of individual dots that can be placed in a line within the span of 1 inch (2.54 cm).

I report the results of $n = 86$ color images randomly extracted from 61 maps. These images are mainly composed of map headers, map comments and region labels without stylized handwriting characters; see Fig. 7.1. Each color image i is transformed to a gray image I_i with the transformation defined in (2.7).

7.2 OCR measures

Accuracy of an algorithm

The **accuracy of an algorithm** is intuitively defined as how close the algorithm's output is from the desirable result. In OCRs, the desirable result is the text contained in the tested image.

[1] This images can be found in: http://www.library.ucla.edu/yrl/reference/maps/blaeu

Definition 7.1: *The **accuracy measure**(AC) of an binary image is defined as*

$$AC(\hat{\mathcal{F}}) = \frac{\#T_{match}}{\#T_{in}}, \tag{7.1}$$

*where $\hat{\mathcal{F}}$ is the estimated foreground of the evaluated image, T_{in} is the original text in the image and T_{match} is the **maximum matching string**, and the notation $\#\cdot$ refers to the number of characters in the string $\cdot$.*

Junker et al. [33] introduced several definitions of maximum matching string C of two strings A and B. In this thesis, however, I define maximum matching string as follows.

Definition 7.2: *Given two strings A and B, we say that A is substring of B ($A \prec B$) if B can be transformed to A by removing characters from it; a **maximum matching string** C of A and B is a string of maximum length such that $C \prec A$ and $C \prec B$.*

The maximum matching string can be computed with the Needleman and Wuntsh [56] algorithm. This algorithm was originally developed for finding similarities in the amino acid sequences of two proteins.

AC measure is an important measure for OCR engines, because the higher the AC measurement, the greater the possibility to extract, by further algorithms, relevant information from the recognized text.

Observe that AC measure does not penalize "*extra characters*" in the output. Then, two different images may lead to the same accuracy but with different number of "*extra characters*". In that case, I judge that an image is better than another one if its OCR output has fewer "*extra characters*". The following measure quantifies number of the "*extra characters*"

Definition 7.3: *The **precision measure** (AC) is defined as*

$$PR(\hat{\mathcal{F}}) = \frac{\#T_{match}}{\#T_{out}}, \tag{7.2}$$

where T_{out} is recognized text from the image.

Table 7.1 – *Pairwise comparison of OCR accuracy. Each cell (y-row,x-column) of the pairwise tables contains two values, n_{yx} and p_{yx}. The number n_{yx} represents the times that the algorithm y has a higher score than the algorithm x, while $p_{yx} = \frac{n_{yx}}{n_{yx}+n_{xy}}$ represents the conditional probability of y's score being higher than x's score. I ascertain that algorithm x is better than algorithm y if $0.75n_{yx} \geq n_{xy}$, which is equivalent to $p_{yx} \geq 0.57$; see Appendix C.*

	Rank	FineReader		OneNote		TopOCR		FreeOCR		MoreDataFast		SimpleOCR	
		n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
FineReader	1	-		40	0.59	59	0.78	56	0.77	63	0.82	68	0.85
OneNote	2	28	0.41	-		53	0.65	45	0.64	56	0.77	61	0.77
TopOCR	3	17	0.22	28	0.35	-		40	0.52	49	0.63	62	0.77
FreeOCR	3	17	0.23	25	0.36	37	0.48	-		47	0.69	59	0.71
MoreDataFast	4	14	0.18	17	0.23	29	0.37	21	0.31	-		47	0.6
SimpleOCR	5	12	0.15	18	0.23	19	0.23	24	0.29	31	0.4	-	

7.3 OCR comparison

I compared six OCR engines: **ABBYY FineReader** 10 Professional (FineReader), **OneNote** 2010 (OneNote)[2], **TopOCR** v3.1, **FreeOCR** 3.0[3], **MoreDataOCR** v3.0, and **SimpleOCR** v3.1.

I ranked the OCRs by the **uncertainty test**, see Appendix C, from pairwise tables of AC measurements, see Table 7.1.

With an α-uncertainty less than 0.9, FineReader is the best, followed by OneNote in second; both TopOCR and FreeOCR rank third. Unfortunately, both FineReader and OneNote are **payware software**, which is an inconvenience for academic software, and both lack **command-line interface** which is essential for my comparison studies. TopOCR is better than FreeOCR with an α-uncertainty around 0.37, which is too high to rank TopOCR over FreeOCR. Nevertheless, I elected TopOCR to carry on with the comparative studies.

A software is **payware software** if it is distributed for money

A **command-line interface** is a mechanism for interacting with a computer operating system or software by typing commands to perform specific tasks.

TopOCR was tested with four parameter sets, some of which include despeckled filters. The program tester reports the maximum AC measurement for each image.

[2]Microsoft OCR Engine Microsoft included in Office Professional Plus 2010.

[3]FreeOCR uses **Tesseract** v2.04 as OCR engine

Table 7.2 – *Each parameter is sampled according the increments of the third column between the range specified in the second column.*

Algorithm	Parameter From/To	Increment
Johannsen's, Kapur's, Kittler's and **Otsu's**	$r : 10/50$	$r : 5$
Kavallieratou's	$\alpha : 1/20$, $r : 10/50$	$\alpha : 1$, $r : 5$
Niblack's	$\alpha : 0/6$, $r : 10/50$	$\alpha : 0.1$, $r : 5$
Portes's	$\alpha : 0/5$, $r : 10/50$	$\alpha : 0.1$, $r : 5$
Sauvola's	$\alpha : 0/1$, $\beta : 32/196$, $r : 10/50$	$\alpha : 0.01$, $\beta : 32$, $r : 5$
Wolf's	$\alpha : 0/1$, $r : 10/50$, $r^* : 50$	$\alpha : 0.01$, $r : 5$

7.4 Experiment I

This section reports the results in [70] where I proposed a mechanism for systematic comparison of the efficiency of unsupervised evaluation methods for parameter selection of binarization algorithms in OCRs.

I performed an extensive comparison of unsupervised evaluation measures, binarization algorithms and OCRs, and I used it to show the strengths of the **unbiased WV measure (normal distribution)**.

7.4.1 Binarization algorithms

I compare the performance of nine binarization algorithms in OCRs: **Johannsen's**, **Kapur**, **Kavallieratou's**, **Kittler's**, **Niblack's**, **Otsu's**, **Portes's**, **Sauvola's**, and **Wolf's**. Authors like Sezgin and Sankur [82], Stathis et al. [84], and Trier and Jain [87] ranked Kittler's, Niblack's, Otsu's and Sauvola's among the best binarization algorithms.

Table 7.2 presents the range and increments of the parameter sampling for each binarization algorithm. I denote $\Omega_{j,k}$ the parameter combination k of the binarization algorithm j, which is constructed by combining the sampled parameters. Sauvola's threshold, for instance, has $5,454$ $\Omega_{j,k}$'s considering that α, β and r are sampled with 101, 6, and 9 different values, respectively.

7.4.2 Evaluation measures

I define the following values in order to evaluate the OCR performance:

Definition 7.4: *The **absolute potential AC measure** of an image I_i is defined as*

$$w_i^* = \max_{j,k} \left\{ AC(\hat{\mathcal{F}}_{i,j,k}) \right\} \tag{7.3}$$

where $\hat{\mathcal{F}}_{i,j,k}$ denotes the estimated foreground of I_i by the binarization algorithm j with parameters $\Omega_{j,k}$.

The value w_i^* approximates the maximum accuracy that the OCR (TopOCR) can compute for I_i in combination with any of the nine binarization methods. Similarly, we can compute $w_{i,j}$ which approximates the maximum accuracy with the binarization algorithm j as:

Definition 7.5: *Given an image I_i, the **relative potential AC measure** of a binarization algorithm j is defined as*

$$w_{i,j} = \max_{k} \left\{ AC(\hat{\mathcal{F}}_{i,j,k}) \right\}. \tag{7.4}$$

The absolute and relative potential AC may change if the number of sampled parameters or tested algorithms is incremented; nevertheless, I consider such values as the groundtruth.

We cannot infer from $w_{i,j}$ the "*goodness*" of the binarization method j to maximize the OCR accuracy because $w_{i,j}$ highly depends on w_i^*. For example, suppose that whichever binarization method is used, the OCR accuracy is equal or lower than 0.5 ($w_i^* \leq 0.5$). Then, if $w_{i,j} = 0.45$ for some j, this could be interpreted either as a low OCR performance, or as a low binarization method performance. However, the ratio of w_i^* to $w_{i,j}$ is 0.90, which means that the binarization method j is highly efficient to maximize the OCR accuracy despite the intrinsic low OCR performance in I_i. Hence, our observations are mainly based on pairwise tables and statistics of the following ratios.

Definition 7.6: *Given an image I_i, the **potential AC efficiency measure** of a binarization algorithm j is defined as the ratio of the relative potential AC measure to the absolute potential AC measure. That is,*

$$x_{i,j} = \frac{w_{i,j}}{w_i^*} \tag{7.5}$$

I also tested the efficiency of unsupervised evaluation measures for the parameter selection of binarization algorithms. For that, I selected the best binarized image in term of each measure and compared their accuracy. The following definition formalizes this concept.

Definition 7.7: *The AC efficiency measure is defined as*

$$y_{i,j}^{(u)} = \frac{AC(\hat{\mathcal{F}}_{i,j}^{(u)})}{w_i^*}, \tag{7.6}$$

where

$$\hat{\mathcal{F}}_{i,j}^{(u)} = \arg\min_{\hat{\mathcal{F}}_{i,j,k}} \left\{ Eval\left(M_r^{(u)}, \hat{\mathcal{F}}_{i,j,k}\right) \right\}, \tag{7.7}$$

$Eval(\cdot,\cdot)$ is defined as in (6.2), and $M_r^{(u)}$ denotes the measure u.

The ratio $x_{i,j}$ approximates the potential efficiency of the binarization algorithm j to maximize the accuracy in I_i. The ratio $y_{i,j}^{(u)}$ approximates the efficiency of measure u to tune the parameters of algorithm j in order to maximize the accuracy in I_i.

In this experiment, I tested the measures:

- **local gray-intensity uniformity measure** (GU_r),
- **local region non-uniformity measure** (NU_r),
- **unbiased uniform variance measure** with normal distribution ($\widehat{UV}_r$),
- **unbiased uniform variance measure** with lognormal distribution ($\widetilde{UV}_r$),
- **unbiased weighted variance measure** with normal distribution ($\widehat{WV}_r$), and
- **weighted variance measure** with lognormal distribution ($\widetilde{WV}_r$).

The radius of all measures was set to $r = 50$ because it is approximately the minimum radius that entirely contains any character in the tested images.

7.4.3 Results and conclusions

The absolute potential AC is greater than 0.60 in all test images; see Fig. 7.2. Indeed, 93% of them are equal or greater than 0.80, which indicates that the OCR (TopOCR) is capable of recognizing most of the characters in our test images. In the same figure, the corresponding relative potential AC measurements of **Niblack's** and **Kavallieratou's** algorithms fluctuate irregularly. A visual comparison between Niblack's and Kavallieratou's graphs is consequently difficult. Because of that, all the following graphs are in decreasing order to make the visual inspection easier.

Table 7.3 – *Pairwise comparison of absolute efficiency. Both Wolf's and Portes's methods marked with (∗) are ranked fourth because their p_{yx}'s values differ from each other slightly. See Table 7.1 for a description of values n_{yx} and p_{yx}.*

	Rank	Joh.		Kap.		Kav.		Kit.		Nib.		Otsu		Por.		Sau.		Wolf	
		n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Johannsen	9	-		2	0.03	0	0.00	13	0.18	0	0.00	2	0.03	0	0.00	0	0.00	0	0.00
Kapur	7	78	0.98	-		1	0.01	57	0.89	0	0.00	18	0.31	0	0.00	0	0.00	2	0.03
Kavallieratou	3	85	1.00	73	0.99	-		80	0.99	10	0.32	57	0.92	30	0.70	4	0.11	28	0.60
Kittler	8	59	0.82	7	0.11	1	0.01	-		0	0.00	5	0.08	0	0.00	0	0.00	0	0.00
Niblack	2	85	1.00	73	1.00	21	0.68	80	1.00	-		62	0.98	33	0.77	7	0.17	31	0.65
Otsu	6	77	0.97	40	0.69	5	0.08	61	0.92	1	0.02	-		4	0.07	1	0.01	5	0.08
Portes	4(∗)	84	1.00	71	1.00	13	0.30	79	1.00	10	0.23	52	0.93	-		4	0.08	19	0.48
Sauvola	1	85	1.00	77	1.00	34	0.89	81	1.00	34	0.83	69	0.99	46	0.92	-		40	0.83
Wolf	4(∗)	85	1.00	70	0.97	19	0.40	76	1.00	17	0.35	57	0.92	21	0.53	8	0.17	-	

Table 7.4 – *Mean (μ) and standard deviation (σ) of the AC efficiency for each binarization algorithm and unsupervised evaluation method. For each algorithm, the best values of $y_{i,j}^{(u)}$ are shown in bold.*

	Potential		$y_{i,j}^{(u)}$											
			GU_r		NU_r		$\widetilde{UV}_r$		$\widehat{UV}_r$		$\widetilde{WV}_r$		$\widehat{WV}_r$	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Johannsen	0.600	0.239	0.483	0.253	0.487	0.258	**0.496**	**0.250**	0.493	0.256	0.486	0.257	0.496	0.252
Kapur	0.845	0.168	0.750	0.201	**0.756**	**0.197**	0.756	0.199	0.751	0.198	0.751	0.200	0.750	0.200
Kavallieratou	0.963	0.048	0.601	0.220	0.517	0.227	**0.763**	**0.224**	0.728	0.222	0.715	0.195	0.763	0.227
Kittler	0.741	0.215	0.640	0.244	**0.658**	**0.243**	0.631	0.250	0.629	0.252	0.646	0.239	0.651	0.238
Niblack	0.964	0.063	0.538	0.233	0.007	0.046	0.767	0.230	0.716	0.241	0.711	0.207	**0.773**	0.227
Otsu	0.864	0.189	0.795	0.217	0.796	0.219	0.789	0.217	0.787	0.217	**0.797**	**0.217**	0.794	0.216
Portes	0.941	0.122	0.777	0.184	0.777	0.184	0.770	0.220	0.753	0.216	0.778	0.185	**0.785**	**0.209**
Sauvola	**0.989**	**0.027**	0.531	0.229	0.058	0.117	0.761	0.247	0.724	0.244	0.712	0.206	**0.798**	**0.210**
Wolf	0.936	0.141	0.801	0.204	0.804	0.191	0.769	0.235	0.740	0.249	0.806	0.193	**0.812**	**0.220**

The results of this experiment are shown in Figure 7.2 (graphs of absolute and potential efficiency), Table 7.4 (mean and variances of AC efficiency), and Table 7.3 (pairwise tables of potential AC efficiency).

Figure 7.3 shows the ranking of all six evaluation measures for each binarization method. This ranking is given by pairwise tables of AC efficiency with an α-Uncertainty lower than 0.9.

A visual inspection of the binarized images suggests that **Johannsen's**, **Kapur's**, **Kittler's**, and **Otsu's** threshold usually wrongly classify a pixel if its neighborhood is completely contained in the background. In contrast, the rest of the algorithms, which have one or two parameters more besides the radius, can successfully binarize this kind of neighborhood by tuning their parameters. My conclusions are also supported for the means and standard deviations of the relative potential AC presented in Table 7.4.

In my test images, the radius used to compute the best binarized images, in terms of relative potential AC, range randomly between 10 and 50 independent of the binarization algorithm. However, all six evaluation measures usually chose a set of parameters where $r = 50$ whichever the binarization algorithm is present (with an exception of **Wolf's algorithm** in which usually $r = 10$). This behavior led Otsu's threshold to have almost the same mean and variance whichever one of the evaluation measures adjusted the Otsu's radius; see Table 7.4 and Fig. 7.4 (right). For example, the histogram in Fig. 7.4 (left) shows in light gray bars the radius's probability of being selected by $\widetilde{WV}_r$. The probability of $r = 50$ is close to 0.90, even though the radius's probability of being optimal in terms of the relative potential AC of Otsu's method is around 0.3. This unwanted effect also appears in Johannsen's, Kapur's and Kittler's methods, pointing out that all six measures are ineffective to adjust the neighborhood radius. I conjectured that all four binarization methods estimate the foreground in such manner that, for a given binarization method, all six measures reach their minimum with the same $\hat{\mathcal{F}}$ (the same radius). Unfortunately, my mathematical analysis is unable to explain this pattern.

I observed that the OCR accuracy in an image depends mostly on how well binarized the image is. In fact, the OCR accuracy of two binarized images mainly differs due broken characters, large false positive spots, and overestimated foreground boundaries.

The ranking given in Fig. 7.3 is based on pairwise tables and not in the measurement magnitude. Therefore, the AC efficiency ranking for my dataset may be similar with other OCRs, but not so the accuracy measurements.

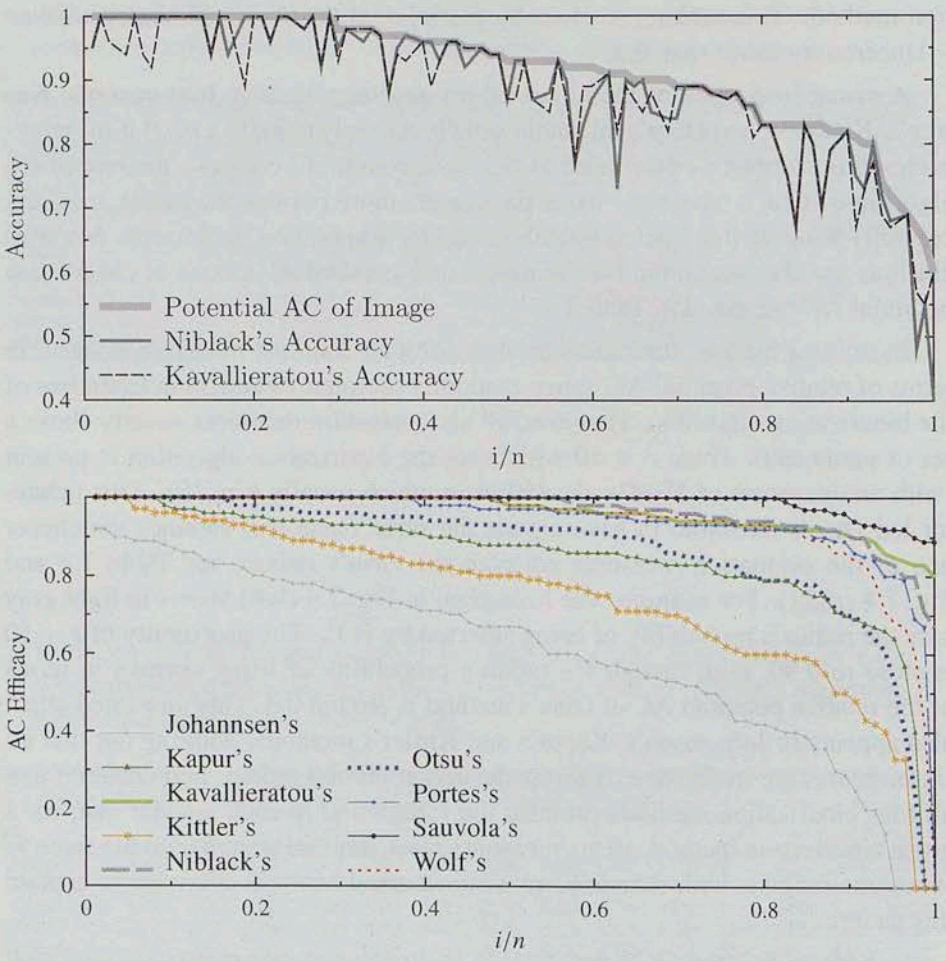


Figure 7.2 – *At the top, graph of the absolute potential AC. On the bottom, ordered graphs of the potential AC efficiency.*

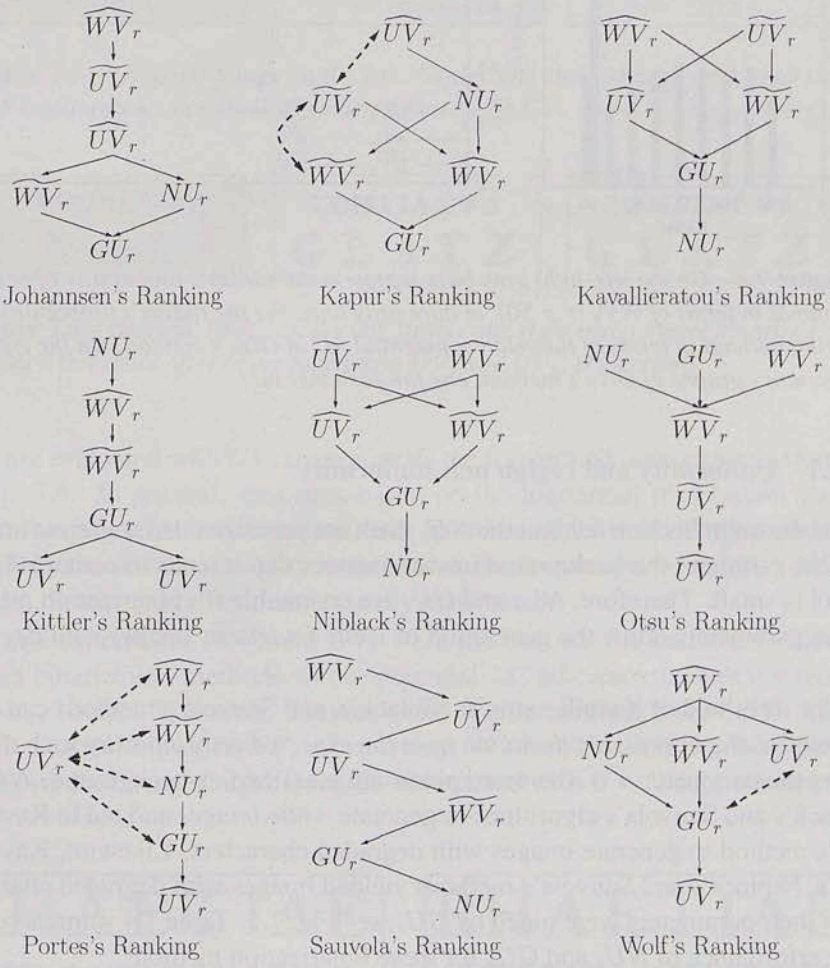


Figure 7.3 – *Measure ranking for each binarization algorithm. The ranking is in decreasing order from top to bottom. Two algorithms with the same ranking either lie on the same level, or are linked with a dash line with double arrow.*

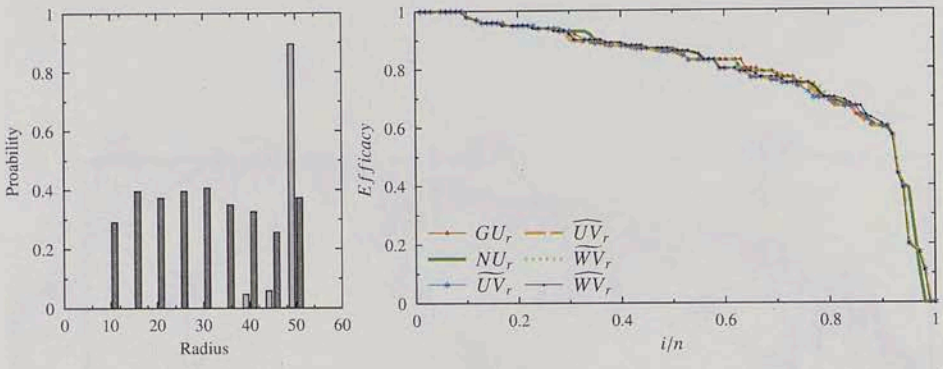


Figure 7.4 – *On the left, light gray bars represent the radius's probability of being optimal in terms of* $\widehat{WV}_r$ *(*$r = 50$*), in dark gray bars, the the radius's probability of being optimal in terms of the relative potential AC of Otsu's method. On the right, efficiency graphs of Otsu's method, one for each measure.*

7.4.3.1 Uniformity and region non-uniformity

I have shown in Section 6.2 that the NU_r does not penalize false negatives and that the GU_r estimates the background in such manner that it tends to contain $\mathcal{F}_r(\boldsymbol{p})$ if $|\mathcal{F}_r(\boldsymbol{p})|$ is small. Therefore, NU_r and GU_r are unsuitable for binarization methods whose parameters allow the generation of *white images* or *images with degraded text*.

The threshold of Kavallieratou's, Niblack's, and Sauvola's methods can be interpreted as *the acceptable deviation from the expected gray intensity* such that the higher the parameter α is, the more pixels are classified as background. NU_r led Niblack's and Sauvola's algorithms to generate *white images* and led to Kavallieratou's method to generate images with degraded characters. Likewise, Kavallieratou's, Niblack's and Sauvola's methods yielded images with degraded characters when their parameters were tuned by GU_r; see Fig. 7.5. Table 7.4 summarizes the low performance of NU_r and GU_r for these binarization methods.

7.4.3.2 Weighted and uniform variance

After inspecting the binarized images visually, I concluded that $\widetilde{UV}_r$ outperforms $\widehat{UV}_r$ in all binarization algorithms (Table 7.4) because $\widehat{UV}_r$ generates more false positive spots (connected components with four or more pixels) which are scattered all around the background. In addition to this noise, binarization algorithms

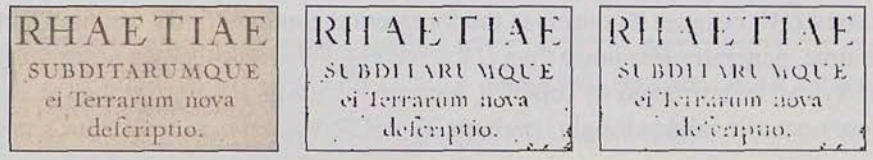


Figure 7.5 – *Original Image on the left. Center and right images were binarized by Kavallieratou's threshold after being tuned with GU_r and NU_r, respectively.*

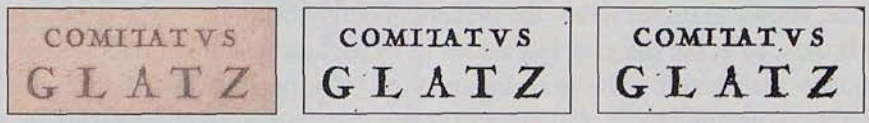


Figure 7.6 – *Original Image on the left. center and right images were binarized by Portes's threshold after being tuned with $\widetilde{UV}_r$ and $\widetilde{UV}_r$, respectively.*

which are evaluated with $\widetilde{UV}_r$ overestimate the foreground contours occasionally; see Fig. 7.6. In general, measures based on the lognormal distribution yielded sharper foreground boundaries than those based on the normal distribution in this test. This indicates that the gray intensities at the foreground boundaries are log-normally distributed rather than normally distributed.

In this experiment, $\widetilde{WV}_r$ and $\widetilde{UV}_r$ were the best for the parameter selection of those binarization methods whose potential AC efficiency is over 0.9 (Kavallieratou's, Niblack's, Porte's, Sauvola's and Wolf's methods); see Table 7.4 and Fig. 7.3. Particularly since $\widetilde{WV}_r$ is better than $\widetilde{UV}_r$ for Sauvola's and Wolf's methods despite observing sharper foreground contours with $\widetilde{UV}_r$. I suppose that $\widetilde{WV}_r$

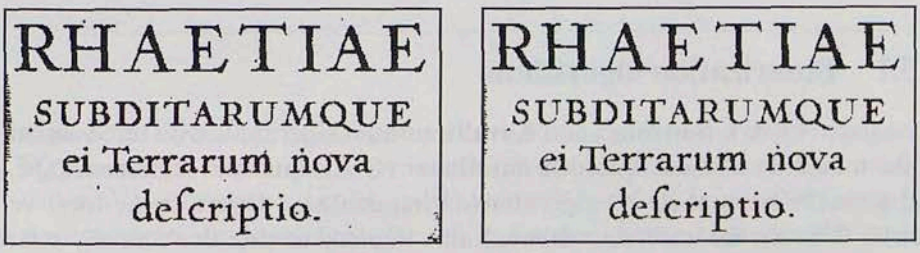


Figure 7.7 – *Left and right images were binarized by Wolf's threshold after being tuned with $\widetilde{UV}r$ and $\widetilde{WV}_r$, respectively.*

surpasses $\widetilde{UV}_r$ because it conserves the foreground contours fairly well and, at the same time, generates less noise than $\widetilde{UV}_r$; see Fig. 7.7. Another reason for this superiority can be attributed to TopOCR because it classifies a character with sharp contours occasionally wrongly; perhaps TopOCR was trained with slim characters.

In the practice, images satisfy the conditions of r-simple images partially. In an image, the performance of $\widetilde{WV}_r$ and $\widetilde{WV}_r$ is directly related with the number of neighborhoods with radius r which satisfy both Model 1 and (6.1). Figure 7.8, for instance, shows an image where the percent of neighborhoods ($r \geq 10$) that satisfy (6.1) is close to 1, but the gray intensities in its background are not approximately identically distributed. The gray intensity of false positive pixels from Wolf's binarization, denoted by $\mathcal{X}$, follows a different distribution to those pixels in $\mathcal{Y} = \mathcal{B} \backslash \mathcal{X}$. As a result, $\widetilde{WV}_r$ leads Wolf's method to generate $\hat{\mathcal{F}} = \mathcal{F} \cup \mathcal{X}$ since

$$\hat{\mu}_{\mathcal{Y}} - \hat{\mu}_{\mathcal{F}} < \sqrt{2} \cdot \max(\hat{\sigma}_{\mathcal{Y}}, \hat{\sigma}_{\mathcal{F}}), \tag{7.8}$$

and

$$\hat{\mu}_{\mathcal{X}} - \hat{\mu}_{\mathcal{Y}} > \sqrt{2} \cdot \max(\hat{\sigma}_{\mathcal{X}}, \hat{\sigma}_{\mathcal{Y}}). \tag{7.9}$$

7.5 Experiment II

Although I present a conscientious analysis for the transition method in Section 7.6, this section reports the tables from [72] for completeness.

In this experiment, I analyzed the performance and running time of the quantile transition threshold in combination with the normal and lognormal transition thresholds.

7.5.1 Binarization algorithms

I compared **Otsu's**, **Sauvola's** and **Kavallieratou's** algorithms with three variants of the transition method: **quantile autolinear** (Q-A), **quantile lognormal**(Q-L), and **quantile normal** (Q-N) algorithms. I implemented Otsu's in the local version to increase the accuracy, although this implementation dramatically raised the running-time. I implemented all the algorithms with **integral images** to compute local values except for Otsu's method, which uses histogram tracking as in Fig. 7.9.

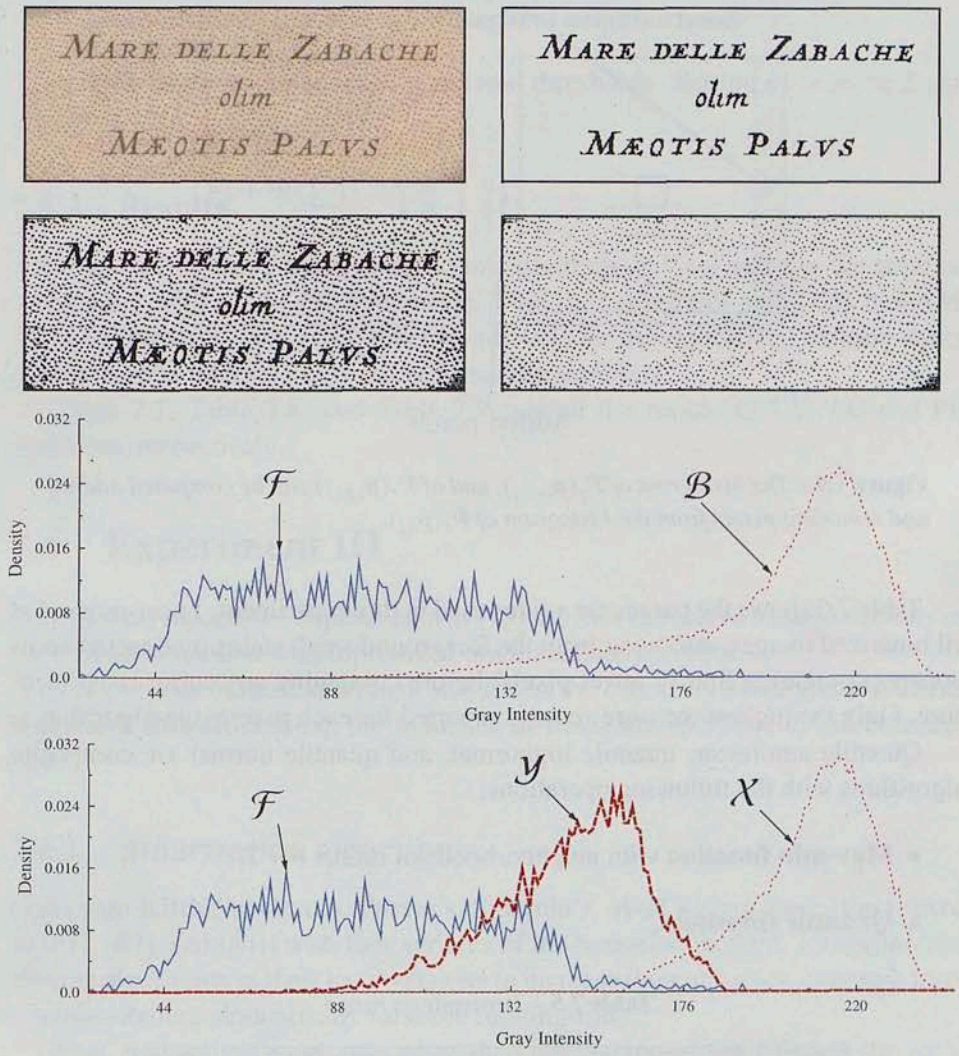


Figure 7.8 – *At the top, an example of a non-ideal image (left) and its corresponding groundtruth (right). In the second row, Wolf's binarization which was tuned with $\widetilde{WV}_r$ (left) and the set of false positives (set $\mathcal{Y}$) generated by the Wolf's binarization (right). In the third row, the density function of gray intensities of $\mathcal{F}$ and $\mathcal{B}$. On the bottom, the density function of gray intensities of $\mathcal{F}$, $\mathcal{X}$, and $\mathcal{Y}$, where $\mathcal{X} = \mathcal{F} \setminus \mathcal{Y}$.*

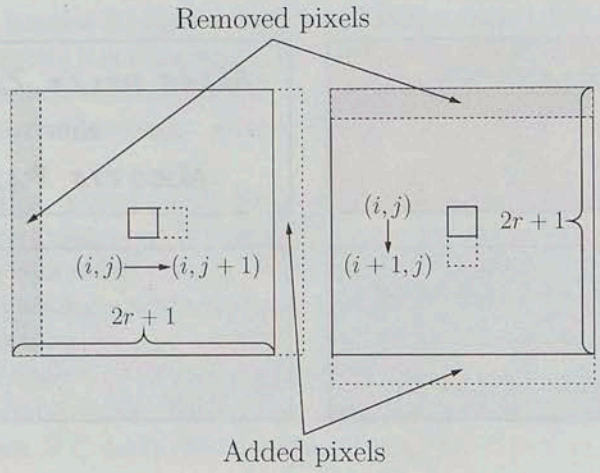


Figure 7.9 – *The histogram of $\mathcal{P}_r(\boldsymbol{p}_{i+1,j})$, and of $\mathcal{P}_r(\boldsymbol{p}_{i,j+1})$ can be computed adding and removing pixels from the histogram of $\mathcal{P}_r(\boldsymbol{p}_{i,j})$.*

Table 7.5 shows the parameter values used in this experiment. I post-processed all binarized images, removing from the foreground small stains (connected components containing four or fewer pixels) before computing any comparison measure. Only the highest measure score is reported for each pair image-algorithm.

Quantile autolinear, quantile lognormal, and quantile normal are composite algorithms with the following operations:

- **Max-min function** with neighborhoods of radius $r = 2$.
- **Quantile threshold**,

Table 7.5 – *Parameter's range*

Algorithm	From/To	Increment
Kavallieratou	0/9	1
Quantile Autolinear	α : 0.1/0.975	0.025
Quantile Lognormal	α : 0.1/0.975	0.025
Quantile Normal	α : 0.1/0.975	0.025
Sauvola	α : 0.025/0.6 β : 128	α : 0.025

- Two **isolation transition operators** ($a = b = 1$). the former using **cross neighborhood**, the later using **diagonal neighborhood**.
- **Autolinear** or **lognormal** or **normal** thresholds. Setting $n_+ = n_- = 5$ and $c = 15$.

7.5.2 Results

For this experiment, I implemented the algorithms in C++ and ran the tests on a computer with a 3.2 GHz Pentium IV Dual core processor and 2 GB in RAM. Table 7.6 presents the 95% confidence intervals for the algorithms' running-times expressing the interval limits on millisecond/megapixel.

Table 7.7, Table 7.8, and Table 7.9 present the results of UV, AC and PR measures, respectively.

7.6 Experiment III

This section reports the experiments in [73] where I compared several variants of the transition method with top-ranked binarization algorithms.

The purpose of this experiment was to test the efficiency of the **double-linear transition threshold** along the influence of transition operators in the binarization.

7.6.1 Binarization algorithms

I compare **Kittler's**, **Otsu's**, **Portes's**, **Sauvola's**, **Wolf's** algorithms (top ranked in [87], [82], and [84]) with four variants of the transition method. I implemented all nine algorithms in their local versions to increase their accuracy, although local implementations dramatically raise the running-time.

Real applications rarely use more than one parameter set. That is the main reason why I fixed Sauvola's $\alpha = 0.5$ and $\beta = 128$, Portes's $\alpha = 2$, and Wolf's $\alpha = 0.5$, which are the recommended parameters; see Section 3.3.

I set the primary neighborhood radius to $r = 50$, local windows of 101x101 pixels, and set the secondary neighborhood radius to 100 for Wolf's method; see Fig. 3.2.

The transition algorithms, denoted by the prefix T, are composite methods with the following combination of operators:

Table 7.6 – *95% confidence intervals for binarization-running time. The intervals are normalized with respect to Sauvola's running-time (millisecond/megapixel) which is the fastest.*

	Raw	Normalized
Kavallieratou	(1718,1727)	(4.0,4.0)
Otsu	(265757,266908)	(630.9,630.9)
Quantile Autolinear	(1802,1813)	(4.2,4.2)
Quantile Lognormal	(2568,2580)	(6.1,6.1)
Quantile Normal	(2039,2051)	(4.8,4.8)
Sauvola	(421,423)	(1.0,1.0)

Table 7.7 – *Pairwise comparison of UV measure. See Table 7.1 for a description of values n_{yx} and p_{yx}.*

	Kav.		Otsu		Q-A		Q-L		Q-N		Sau.	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kavallieratou	-	-	12	0.14	1	0.01	1	0.01	1	0.01	0	0.00
Otsu	71	0.86	-	-	6	0.07	3	0.04	3	0.04	0	0.00
Quantile Autolinear	82	0.99	77	0.93	-	-	2	0.02	19	0.23	3	0.04
Quantile Lognormal	82	0.99	80	0.96	81	0.98	-	-	56	0.67	8	0.10
Quantile Normal	82	0.99	80	0.96	64	0.77	27	0.33	-	-	4	0.05
Sauvola	83	1.00	83	1.00	80	0.96	75	0.90	79	0.95	-	-

Table 7.8 – *Pairwise comparison of AC measure. See Table 7.1 for a description of values n_{yx} and p_{yx}.*

	Kav.		Otsu		Q-A		Q-L		Q-N		Sau.	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kavallieratou	-	-	55	0.75	8	0.12	10	0.15	4	0.06	10	0.16
Otsu	18	0.25	-	-	1	0.01	2	0.03	1	0.01	1	0.01
Quantile Autolinear	58	0.88	70	0.99	-	-	25	0.46	13	0.33	33	0.61
Quantile Lognormal	56	0.85	71	0.97	29	0.54	-	-	17	0.40	31	0.58
Quantile Normal	61	0.94	71	0.99	27	0.68	25	0.60	-	-	33	0.67
Sauvola	54	0.84	69	0.99	21	0.39	22	0.42	16	0.33	-	-

Table 7.9 – *Pairwise comparison of PR measure. See Table 7.1 for a description of values n_{yx} and p_{yx}.*

	Kav.		Otsu		Q-A		Q-L		Q-N		Sau.	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kavallieratou	-	-	36	0.45	7	0.09	3	0.04	6	0.08	5	0.06
Otsu	44	0.55	-	-	4	0.05	6	0.08	3	0.04	1	0.01
Quantile Autolinear	74	0.91	74	0.95	-	-	23	0.36	22	0.44	26	0.39
Quantile Lognormal	76	0.96	71	0.92	41	0.64	-	-	39	0.61	31	0.44
Quantile Normal	74	0.93	74	0.96	28	0.56	25	0.39	-	-	21	0.35
Sauvola	75	0.94	76	0.99	41	0.61	39	0.56	39	0.65	-	-

- **Max-min function** with neighborhoods of radius r.
- **Double-linear** threshold for transition values using either the empirical scaled density function denoted by DF or the empirical complementary cumulative distribution function denoted by CCD.
- **Isolate transition operators** in the following order:
  1. **cross transition operator**,
  2. **diagonal transition operator**, and
  3. **frame transition operator** ($x = y = 2$).
- **Incidence transition operator** ($k = 4$, $a = b = 3$).
- **Dilation transition operator** ($a = b = 3$).
- Gray-intensity threshold. Setting $n_+ = n_- = 25$, $c = 15$, and using either the **normal threshold** (simple form) denoted by N or the **lognormal threshold** (simple form) denoted by L.

I named these four variants T-DF-N, T-DF-L, T-CCD-N, and T-CCD-L, depending on how the algorithm computes the transition and gray-intensity thresholds.

I also tested three variants of T-CCD-L in order to analyze the influence of transition operators on the transition method: T-CCD-L-A does not include any transition operator, T-CCD-L-B includes only the isolate transition operators, and T-CCD-L-C includes both isolate transition operators and incidence transition operators.

All binarized images were post-processed removing from the foreground small stains (connected components containing four or fewer pixels) before computing any evaluation measure. The following operators were applied in this order:

1. **cross isolate operator**,
2. **diagonal isolate operator**, and
3. **frame isolate operator** ($x = y = 2$).

7.6.2 Evaluation measures

In this experiment, I used the **unbiased uniform variance measure** with normal distribution ($\widehat{UV}_r$) to assess the segmentation quality.

As I stated in Section 7.4, the mean and variance of AC measures are unsuitable to assess the performance of a binarization algorithm in OCRs. Hence, for the purpose of this experiment, I redefined the **AC efficiency measure** as:

Definition 7.8: *Given an image I_i and an binarization algorithm j,*

$$y_{i,j} = \frac{AC(\hat{\mathcal{F}}_{i,j})}{w_i^*} \tag{7.10}$$

*where w_i^* is the **absolute potential AC measure** (Definition 7.4), and $\hat{\mathcal{F}}_{i,j}$ is the estimated foreground of I_i by the binarization algorithm j.*

Remark 7.1: The values w_i^* in this experiment are the same as those values w_i^* in experiment I (Section 7.4).

7.6.3 Results and conclusions

I arranged the test images such that the graph of AC accuracy is decreasing for an easier visual comparison; see Section 7.4.3 for details.

$\widehat{UV}_r$ measure penalizes eroded and overestimated foreground boundaries, but it also penalizes stains (ink stains and dark background spots) that are classified as background so that algorithms that compute foreground boundaries correctly and classify stains as foreground are highly scored, like **Wolf's algorithm** which is the best in terms $\widehat{UV}_r$ measure; see Table 7.10. However, scattered stains and a slight overestimation of the foreground contour lead Wolf's algorithm to a low OCR performance; see Table 7.11, Table 7.12, and Table 7.13.

Kitller's algorithm also classifies stains as foreground but, contrary to Wolf's algorithm, it overestimates the foreground boundaries greatly. In consequence, Kittler's algorithm is a medium rank in terms of UV measure and reports the lowest AC efficiency because of the overestimated foreground boundary; see Fig. 7.10.

Sauvola's algorithm computes low thresholds, which discard stains from the foreground, but low thresholds also produce eroded foreground boundaries that are strongly penalized by UV measure; see Table 7.10. As a result, Sauvola's algorithm was the worst in terms of UV. What is more, this also affects the OCR performance badly; see Table 7.10.

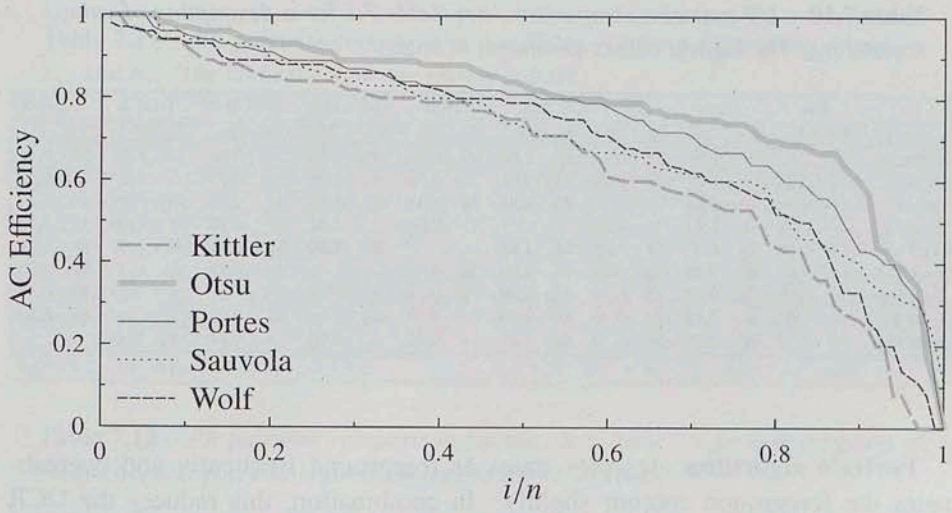


Figure 7.10 – *Ordered AC efficiency of no-transition algorithms.*

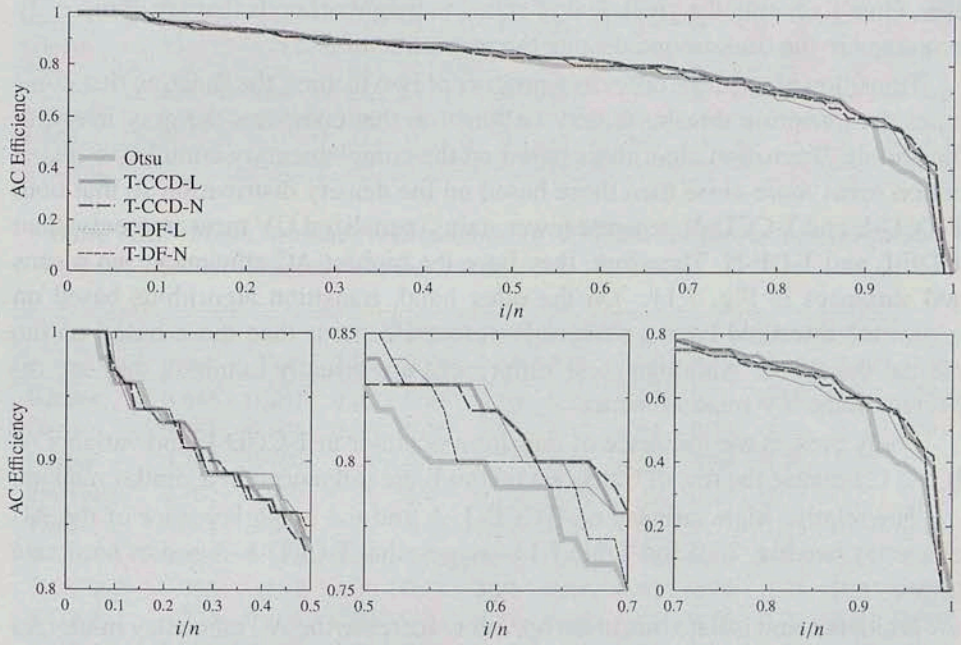


Figure 7.11 – *At the top, ordered AC efficiency graphs of transition algorithms; details on the bottom. Otsu's graph is plotted as a graph of reference.*

Table 7.10 – *UV pairwise comparison. See Table 7.1 for a description of values n_{yx} and p_{yx}. The highest values are shown in bold.*

	Kit.		Otsu		Por.		Sau.		Wolf		T-CCD-L		T-CCD-N		T-DF-L		T-DF-N	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kittler	-	-	72	0.84	35	0.41	**86**	**1.00**	29	0.34	30	0.35	56	0.65	32	0.37	58	0.67
Otsu	14	0.16	-	-	12	0.14	85	0.99	7	0.08	4	0.05	8	0.09	4	0.05	6	0.07
Portes	51	0.59	74	0.86	-	-	**86**	**1.00**	**36**	**0.42**	46	0.53	61	0.71	44	0.51	60	0.70
Sauvola	0	0.00	1	0.01	0	0.00	-	-	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
Wolf	57	0.66	79	0.92	**50**	**0.58**	**86**	**1.00**	-	-	**50**	**0.58**	66	0.77	**52**	**0.60**	68	0.79
T-CCD-L	**56**	**0.65**	**82**	**0.95**	40	0.47	**86**	**1.00**	**36**	**0.42**	-	-	**83**	**0.97**	44	0.52	79	0.92
T-CCD-N	30	0.35	78	0.91	25	0.29	**86**	**1.00**	20	0.23	3	0.03	-	-	4	0.05	47	0.56
T-DF-L	54	0.63	**82**	**0.95**	42	0.49	**86**	**1.00**	34	0.40	40	0.48	82	0.95	-	-	**81**	**0.94**
T-DF-N	28	0.33	80	0.93	26	0.30	**86**	**1.00**	18	0.21	7	0.08	37	0.44	5	0.06	-	-

Portes's algorithm classifies stains as foreground frequently and overestimates the foreground contour slightly. In combination, this reduces the OCR performance but increases the UV measurements.

Otsu's and transition algorithms determine sharp foreground contours. However, Otsu's generated a great deal of stains in neighborhoods that are completely contained in the background despite the restriction of (3.5).

Transition algorithms differ as a product of two factors: the function that computes the transition thresholds and the function that computes the gray-intensity thresholds. Transition algorithms based on the complementary cumulative distribution resist more noise than those based on the density distribution so that both T-CCD-L and T-CCD-N generate fewer stains (penalized UV measurements) than T-DF-L and T-DF-N. Therefore, they have the highest AC efficiency; see means and variances in Fig. 7.11. On the other hand, transition algorithms based on lognormal threshold have a sharper foreground contour than those based on the normal threshold. Although these differences are visually minimal, they are reflected on the UV measurements.

I only present the influence of transition operator in T-CCD-L and variants A, B and C because the rest of transition methods are influenced in a similar manner.

The relative high variance of T-CCD-L-A and the graph behavior of the AC efficiency, see Fig. 7.12 and Table 7.14, suggest that T-CCD-L-A resists moderate noise.

Incidence and isolate transition operators increase the AC efficiency in images with high noise level at the cost of dropping the cardinality of transition set and, in consequence, the AC efficiency decreases in images whose foreground contains small connected components like punctuation marks and small characters. Note

Table 7.11 – *AC pairwise comparison. See Table 7.1 for a description of values n_{yx} and p_{yx}. The highest values are shown in bold.*

	Kit.		Otsu		Por.		Sau.		Wolf		T-CCD-L		T-CCD-N		T-DF-L		T-DF-N	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kittler	-	-	15	0.21	18	0.25	26	0.35	20	0.33	14	0.18	13	0.18	17	0.22	11	0.14
Otsu	58	0.79	-	-	44	0.62	**55**	**0.76**	52	0.78	20	0.34	23	0.41	29	0.45	22	0.39
Portes	53	0.75	27	0.38	-	-	37	0.51	40	0.61	21	0.30	27	0.38	25	0.36	24	0.35
Sauvola	48	0.65	17	0.24	35	0.49	-	-	38	0.54	19	0.26	20	0.27	25	0.32	17	0.25
Wolf	40	0.67	15	0.22	26	0.39	33	0.46	-	-	14	0.20	17	0.23	21	0.28	13	0.20
T-CCD-L	64	0.82	**38**	**0.66**	**49**	**0.70**	54	0.74	**56**	**0.80**	-	-	**29**	**0.62**	24	**0.60**	**27**	**0.54**
T-CCD-N	61	0.82	33	0.59	44	0.62	53	0.73	**56**	0.77	18	0.38	-	-	28	0.52	23	0.47
T-DF-L	61	0.78	35	0.55	44	0.64	52	0.68	53	0.72	16	0.40	26	0.48	-	-	25	0.42
T-DF-N	**66**	**0.86**	34	0.61	45	0.65	52	0.75	53	**0.80**	**23**	**0.46**	26	0.53	**34**	0.58	-	-

Table 7.12 – *PR pairwise comparison for text. See Table 7.1 for a description of values n_{yx} and p_{yx}. The highest values are shown in bold.*

	Kit.		Otsu		Por.		Sau.		Wolf		T-CCD-L		T-CCD-N		T-DF-L		T-DF-N	
	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Kittler	-	-	20	0.24	32	0.39	44	0.52	32	0.41	20	0.24	21	0.25	21	0.25	21	0.25
Otsu	**65**	**0.76**	-	-	52	0.63	59	0.73	51	0.64	**32**	**0.42**	31	0.42	**37**	0.49	30	0.43
Portes	51	0.61	31	0.37	-	-	53	0.64	45	0.56	28	0.34	31	0.38	26	0.31	33	0.40
Sauvola	41	0.48	22	0.27	30	0.36	-	-	34	0.41	19	0.23	20	0.25	24	0.29	20	0.24
Wolf	46	0.59	29	0.36	35	0.44	49	0.59	-	-	25	0.31	27	0.33	27	0.33	26	0.33
T-CCD-L	62	**0.76**	**44**	**0.58**	55	0.66	**64**	**0.77**	**55**	**0.69**	-	-	**40**	**0.59**	31	**0.58**	**38**	**0.58**
T-CCD-N	62	0.75	42	**0.58**	51	0.62	61	0.75	**55**	0.67	28	0.41	-	-	33	0.49	32	0.52
T-DF-L	62	0.75	39	0.51	**57**	**0.69**	59	0.71	54	0.67	22	**0.42**	35	0.51	-	-	34	0.50
T-DF-N	63	0.75	39	0.57	50	0.60	62	0.76	53	0.67	28	**0.42**	30	0.48	34	0.50	-	-

Table 7.13 – *Mean, variance and quantiles of AC efficiency for each binarization method. The best values are shown in bold.*

			Values i/n such that $y_{z_i,j}$ equal or greater than								
	mean	**Var**	**1.00**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.60**	**0.50**
Kittler	0.646	0.261	0.02	0.05	0.10	0.21	0.37	0.50	0.55	0.65	0.78
Otsu	0.787	0.196	0.06	**0.16**	0.27	0.47	0.59	0.73	0.80	0.90	0.91
Portes	0.748	0.203	0.05	0.10	0.21	0.34	0.56	0.64	0.70	0.83	0.88
Sauvola	0.702	0.212	0.06	0.09	0.19	0.28	0.42	0.47	0.55	0.77	0.81
Wolf	0.691	0.246	0.00	0.06	0.14	0.34	0.47	0.57	0.60	0.74	0.84
T-CCD-L	**0.805**	**0.175**	0.08	0.13	0.30	0.48	**0.67**	**0.76**	**0.84**	0.91	**0.95**
T-CCD-N	0.798	0.182	0.08	0.13	**0.31**	0.45	**0.67**	**0.76**	0.79	0.90	**0.95**
T-DF-L	0.795	0.196	**0.09**	0.12	**0.31**	**0.51**	0.65	**0.76**	0.79	0.87	**0.95**
T-DF-N	0.796	0.189	0.08	0.15	0.28	**0.51**	0.63	0.72	0.81	**0.92**	0.94

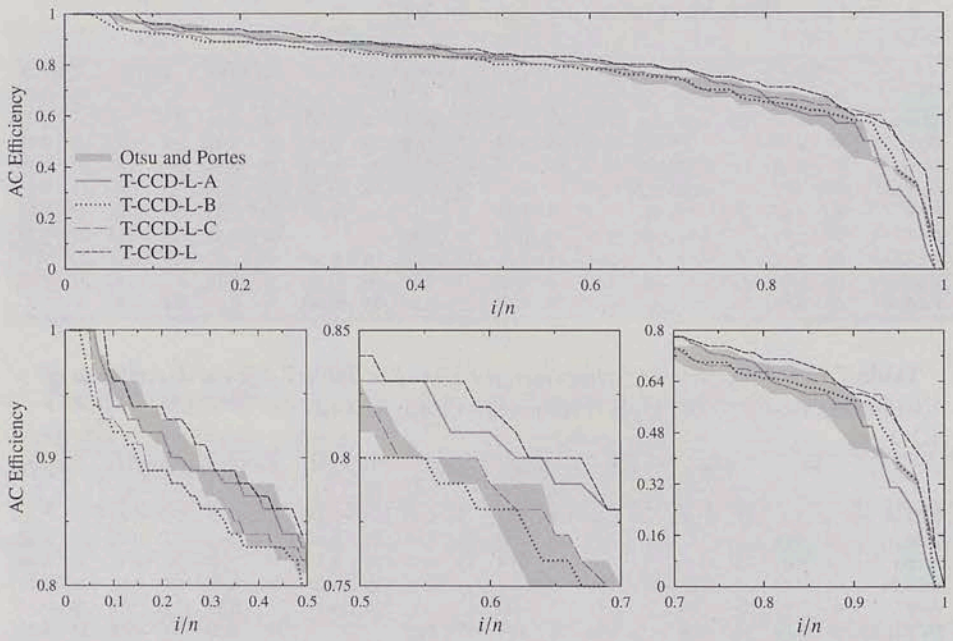


Figure 7.12 – *Ordered AC efficiency graphs of T-CCD-L and variants. The area between Otsu's and Portes's graphs is plotted in light gray as reference.*

Table 7.14 – *Mean, variance and quantiles of AC efficiency for T-CCD-L and variants. The best values are shown in bold.*

			Values i/n such that $y_{z_i,j}$ equal or greater than								
	mean	**Var**	**1.00**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**	**0.70**	**0.60**	**0.50**
T-CCD-L-A	0.771	0.213	0.06	0.09	0.24	0.44	0.64	0.73	0.79	0.87	0.92
T-CCD-L-B	0.758	0.186	0.3	0.05	0.15	0.33	0.55	0.69	0.76	0.88	0.93
T-CCD-L-C	0.771	0.175	0.5	0.06	0.16	0.36	0.59	0.70	0.73	**0.93**	**0.95**
T-CCD-L	**0.805**	**0.175**	**0.08**	**0.13**	**0.30**	**0.48**	**0.67**	**0.76**	**0.85**	0.91	**0.95**

that the incidence operator does not remove dense salt and pepper noise. Thus, it has to be applied after isolate transition operators.

The dilation transition operator counterbalances the unwanted effect of the incidence and isolate transition operators by increasing the cardinality of diminished transition sets. I should remark that this operator has to be applied in images with low noise level, or after isolate and incidence transition operators. Otherwise, the noise is magnified.

7.7 Summary

In this chapter, I presented the four main contributions of my thesis: an analysis of the performance of binarization algorithms and unsupervised measures. In concrete, I proposed two mechanisms for systematic comparison of the efficacy of algorithms using OCR's and historical documents (Blaeu maps).

The data set used in all tests is described in Section 7.1. Later on, OCR's measures based on the maximum matching string (Definition 7.2) are discussed in Section 7.2.

Six commercial OCRs are evaluated in Section 7.3. TopOCR is chosen to carry on with all comparative studies since it has performed the best among freeware software and has command-line mode (essential tool for massive evaluations).

In Section 7.4, I proposed a mechanism for systematic comparison of the efficacy of unsupervised evaluation methods for parameter selection of binarization algorithms in optical character recognition (OCR). The comparison process is streamlined in several steps. Given an unsupervised measure and a binarization algorithm, I:
(i) find the best parameter combination for the algorithm in terms of the measure,
(ii) use the best binarization of an image on an OCR, and
(iii) evaluate the accuracy of the characters detected.

The performance of the transition method is evaluated in Section 7.5 and Section 7.6. The running-time of three variants of the transition method is determined under a normalization by the running-time of Sauvola's algorithm. It turns out that the transition method is between 4.2 and 6.1 times slower than Sauvola's methods, which is one of the fastest algorithms. However, it is between 100 and 150 times faster than Otsu's methods, which is considered as one of the best binarization algorithms.

Results presented in Table 7.8 and Table 7.13 indicate that the transition method outperforms top-ranked binarization algorithms, namely Otsu's, Wolf's, Sauvola's,

and Kittler's methods. Table 7.13 also indicates that:
(i) the transition method resists highest levels of noise,
(ii) the complementary cumulative distribution function decreases the impact of outliers on the double-linear threshold, and
(iii) the lognormal threshold generates sharper foreground contours than both normal and autolinear threshold.

Since all variants are influenced by transition operators in a similar manner, Table 7.14 presents the influence of transition operator in a particular variant of the transition method. The incidence transition operator can remove noise that isolated operators cannot, and the dilation transition operator can improve the performance of normal and lognormal thresholds.

Chapter 8

Slope estimators (chapter n+1)

Always strive to win, because in so doing even when you lose, you still win!

Salomé Angulo Romero
Mexican professor of mathematics (1949-2010)

I wrote "*chapter n+1*" in the title of this chapter because I introduce the **differences-rate estimator** for the slope in a **linear regression model**, which is apparently unrelated to binarization and the general topics of my thesis. However, I developed this novel estimator for the **double-linear threshold** in which the slope of two lines from a histogram are estimated.

In this chapter, I prove that this novel estimator is an unbiased estimator with low computational cost. Although the **breakdown point** of differences-rate estimator is zero, it can accurately estimate the slope on histograms of empirical complementary cumulative distribution functions where the effect of outliers is faded. Moreover, the alternative form of this estimator is linearly computed in the number of samples and, in consequence, it is suitable for estimating the slope of lines in large histograms with extreme values, and for time-consuming algorithms.

I describe a potential application of this estimator to estimate the exponent parameters in overestimated measurements drawn from a power-law distribution.

8.1 Simple Linear regression model

Consider the **simple linear regression model**

$$y_i = x_{i,1} \cdot \beta + \alpha + \epsilon_i, \qquad \text{for } 1 \leq i \leq n \tag{8.1}$$

where $\beta \in \mathbb{R}$ is the **slope parameter**, $\alpha \in \mathbb{R}$ is the **intercept parameter**, the observations are of the form $z_i = (x_i, y_i) \in \mathbb{R}^2$, and ϵ_i is a random variable depicting the error from the observed data.

estimator

An **estimator** is a measure calculated from a sample of data that is used to infer the value of an unknown parameter in a statistical model. In the simple linear regression model, β and α are the parameters to be estimated. Four concepts are usually employed to evaluate an estimator: **unbiasedness**, **asymptotic efficiency**, **breakdown point**, and **run-time complexity**.

Definition 8.1: *Assume that the parameter of a model is defined in (a,b). For an estimator $\hat{\theta}$ to be **unbiased**, we mean that on the average the estimator will yield the true value for all $\theta \in (a, b)$. That is, the estimator is unbiased if*

$$E\left(\hat{\theta}\right) = \theta \quad \textit{for all } \theta. \tag{8.2}$$

The theorem of the **Cramer-Rao lower bound – (simple linear model)** states that if $\hat{\theta}$ is an unbiased estimator of θ, then there exists $LB(\theta)$ such that

$$Var(\hat{\theta}) \geq LB(\theta), \tag{8.3}$$

which is known as Cramer-Rho lower bound. The calculation of the Cramer-Rho lower bound depends on the distribution of ϵ_i's, and it is derived from the inverse of a **Fisher information matrix**.[1]

efficiency

Given an estimator $\hat{\theta}$ of an unknown parameter θ, the **efficiency** is a measure of how close $Var(\hat{\theta})$ is to Cramer-Rho lower bound. It is defined by

$$\text{Efficiency}(\hat{\theta}) = \frac{LB(\theta)}{Var(\hat{\theta})}, \tag{8.4}$$

[1] Readers interested in further pursuing the Cramer-Rao lower bound and related topics are encouraged to consult the book by Kay [38].

where $LB(\theta)$ is the Cramer-Rho lower bound for $Var(\hat{\theta})$. An estimator with efficiency 1.0 is said to be an **efficient estimator**.

The efficiency is usually computed for independent observations such that $\epsilon_i \sim N(0, \sigma^2)$. Hence, for the simple linear model, the Cramer-Rao lower bound is given by

$$Var(\hat{\beta}) \geq \frac{n \cdot \sigma^2}{n \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} = LB(\beta) \quad \text{for all } \beta. \tag{8.5}$$

For histograms, (8.5) is simplified to

$$Var(\hat{\beta}) \geq \frac{12\sigma^2}{x^2 \cdot [n-1] \cdot n \cdot [n+1]} = LB(\beta) \quad \text{for all } \beta. \tag{8.6}$$

where the observations have the form $(i \cdot x, y_i)$ for $i = 1, 2, \ldots, n$.

asymptotic efficiency

The **asymptotic efficiency** of an estimator is then defined as the estimator efficiency for $n \to \infty$. For example, in the presence of Gaussian noise, the mean estimator has an asymptotic (large sample) efficiency of 1.0 (achieving the lower bound) while the median estimator's efficiency is only $\frac{2}{\pi} \approx 0.64$.

breakdown point

The notion of **breakdown point** was coined, defined, and discussed by Hampel [28]. The breakdown point of an estimator is informally defined as the smallest percentage of contaminated data that may cause an estimator to take misleading values. For example, the breakdown point of the sample mean is $\frac{1}{n}$ since a single large outlier can corrupt the result. The median remains reliable if less than half of the data are contaminated. Indeed, 50% is the best that can be expected; for larger amounts of contamination, it becomes impossible to distinguish between the "good" and the "bad" parts of the sample.

For the run-time complexity, we use the conventional **Big – O** notation $O(\cdot)$.

8.2 Estimators

The following estimators of both slope and intercept parameters are defined in terms of the simple linear regression model. In the same manner, their complexity analysis is simplified. Their generalization for higher dimensions can be found in their references.

The **least square estimator**[2] proposed around 1795 (LSS) is defined as

$$(\hat{\beta}_{LSS}, \hat{\alpha}_{LSS}) = \arg\min_{\beta',\alpha'} \sum_{i=1}^{n} r_i^2(\beta', \alpha'), \tag{8.7}$$

where

$$r_i(\beta', \alpha') = y_i - x_i \cdot \beta' - \alpha'. \tag{8.8}$$

The values $r_i(\beta', \alpha')$ are known as **residuals**. A more complete name for this estimator would be **least sum of squares estimator**, which I adopt for the rest of this chapter.

If ϵ_i's are **independent and identically distributed**, such that $E(\epsilon_i)$ is finite, then

$$\hat{\beta}_{LSS} = \frac{n \cdot \sum_{i=1}^{n} y_i \cdot x_i - \left[\sum_{i=1}^{n} y_i\right] \cdot \left[\sum_{i=1}^{n} x_i\right]}{n \cdot \sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \tag{8.9}$$

is unbiased. The efficiency of $\hat{\beta}_{LSS}$ at Gaussian noise is 1.0. Moreover, (8.9) has the mathematical beauty of being an **arithmetic form** and the computational beauty of being linear on the number of observations. However, a single outlier can lead $\hat{\beta}_{LSS}$ to misleading values. Therefore, its breakdown point is zero.

I refer as **arithmetic form** to those equations with close form such that only arithmetic operations (addition, subtraction, multiplication, and division) are involved.

In 1887, Edgeworth [21] [22] proposed the **least sum of absolute errors estimator** (LSAE), improving a proposal by Boscovich:

$$(\hat{\beta}_{LSAE}, \hat{\alpha}_{LSAE}) = \arg\min_{\beta',\alpha'} \sum_{i=1}^{n} \|r_i(\beta', \alpha')\|, \tag{8.10}$$

where $\|\cdot\|$ denotes the absolute value. This estimator is less sensitive to outliers than the least sum of squares estimator, but even so, its breakdown point is zero. Another drawback is that $\hat{\beta}_{LSAE}$ depends on $\hat{\alpha}_{LSAE}$. The estimator $\hat{\beta}_{LSAE}$ is unbiased only if $E(\epsilon_i) = 0$ where ϵ_i are independent and identically distributed. Narula and Wellington [55] presented a survey of algorithms to calculate this estimator.

Huber [30] introduced the **M-estimators** in 1973, which is defined as

$$(\hat{\beta}_M, \hat{\alpha}_M) = \arg\min_{\beta',\alpha'} \sum_{i=1}^{n} \rho\left(r_i(\beta', \alpha')\right), \tag{8.11}$$

[2] This estimator is attributed to Carl Friedrich Gauss. Adrien-Marie was the first to publish the method, however. See Stigler [85] for historical discussion.

where $\rho(x)$ is not monotone with one minimum in zero such that $\varphi(x) = (\frac{d}{dx})\rho(x)$ is continuous and bounded. If $\rho(x)$ is convex, Huber proved that (8.11) is equivalent to solve the system

$$\begin{aligned} \sum_{i=1}^{n} \varphi\left(r_i(\beta', \alpha')\right) \cdot \beta' &= 0 \\ \sum_{i=1}^{n} \varphi\left(r_i(\beta', \alpha')\right) \cdot \alpha' &= 0. \end{aligned} \tag{8.12}$$

$\hat{\beta}_M$ is unbiased only if $E(\phi(\epsilon_i)) = 0$ and the ϵ_i's are independent and approximately identically distributed. Choosing an adequate $\psi(x)$, M-estimators are statistically more efficient than the least sum of absolute errors estimator at central model and Gaussian error; M-estimators, however, cannot cope with grossly aberrant values in x_i's, namely **leverage points**, which have a large influence in (8.12). Furthermore, solving (8.12) may need numerical optimization algorithms. Subsequent variants of M-estimators achieved around 30% of the breakdown point; see [77] for more references of these estimators.

A point (x_i, y_i) whose x_i is outlying is called a **leverage point**.

The **repeated medians estimator**, proposed by Siegel [83] in 1982, can resist the effects of outliers having the best breakdown point (50%). For the simple linear regression model, this estimator is defined as

$$\hat{\beta}_{RM} = \underset{i}{median} \left\{ \underset{j \neq i}{median} \left\{ \frac{y_i - y_j}{x_i - x_j} \right\} \right\}. \tag{8.13}$$

Although $\hat{\beta}_{RM}$ has no close form, it can be calculated in a deterministic manner with a running-time of $O(n^2 \ln(n))$; see [68], Chapter 8.5. This estimator is unbiased assuming that ϵ_i's are independent and approximately identically distributed and $E(\epsilon_i)$ exists. It is robust against a high percentage of outliers. The Gaussian efficiency of the repeated median method was found experimentally as being around 0.60.

Two years later, in 1984, Rousseeuw [77] proposed the **least median of squares estimator** defined as

$$(\hat{\beta}_{LMS}, \hat{\alpha}_{LMS}) = \min_{\beta', \alpha'} \left\{ \underset{i}{median} \left\{ r_i^2(\beta', \alpha') \right\} \right\}. \tag{8.14}$$

$\hat{\beta}_{LMS}$ depends on $\hat{\alpha}_{LMS}$, and its Gaussian asymptotic efficiency is 0%. Another drawback is that $\hat{\beta}_{LMS}$ is unbiased only if the ϵ_i's are independent and identically distributed such that $E(\epsilon_i) = 0$ and finite $E(\epsilon_i^2)$. In addition, the best algorithm known to compute $\hat{\beta}_{LMS}$, by Edelsbrunner and Souvaine [20], has a run-time

$O(n^2)$. In the same publication, Rousseeuw also proposed the **least trimmed of squares estimator**; Rousseeuw said that he has "In press" a publication about the least trimmed of squares estimator, however, I was unsuccessful in tracking such a publication. (LTSS) which minimizes the sum of squares of the smallest k residuals. That is

$$(\hat{\beta}_{LTSS}, \hat{\alpha}_{LTSS}) = \min_{\beta',\alpha'} \left\{ \sum_{i=1}^{k} r^2(\beta', \alpha')_{(i)} \right\}, \tag{8.15}$$

where $r^2(\beta', \alpha')_{(i)}$ denotes the smallest i value from the set

$$\left\{ r_i^2(\beta', \alpha') \mid \text{for } i = 1, \ldots, n \right\}. \tag{8.16}$$

$\hat{\beta}_{LTSS}$ is unbiased with the same assumptions of $\hat{\beta}_{LMS}$, and it has a Gaussian asymptotic efficiency of 8%. Moreover, it also reaches a 50% breakdown point for $k = \frac{n+3}{2}$. However, known algorithms for its calculation have a run-time of $O(n^2 ln(n))$ or higher; see Li [43].

Rousseeuw et al. [78] proposed, in 1993, the **least quartile difference estimator** (LQAD), which is defined as

$$\hat{\beta}_{LQAD} = \min_{\beta'} \{ \| r(\beta') \|_{(k)} \} \tag{8.17}$$

where $\| r(\beta') \|_{(k)}$ denotes the k-smallest element from $\binom{n}{2}$ elements of the set

$$\begin{aligned} &\left\{ \| r_i(\beta', \alpha') - r_j(\beta', \alpha') \| \; ; \; 0 \le j < i \text{ for } i = 1, \ldots, n \right\} \\ &= \left\{ \| y_i - \beta' \cdot x_i - y_j + \beta' \cdot x_j \| \; ; \; 0 \le j < i \text{ for } i = 1, \ldots, n \right\}. \end{aligned} \tag{8.18}$$

This estimator has a breakdown point of 50% if

$$k = \binom{\frac{n+3}{2}}{2}. \tag{8.19}$$

Furthermore, $\hat{\beta}_{LQAD}$ does not depend on $\hat{\alpha}_{LQAD}$, and it is unbiased if the ϵ_i's are independent and approximately identically distributed such that $E(\epsilon_i)$ exists. Its asymptotic efficiency at Gaussian noise is 0.67. However, known algorithms for the exact solution of $\hat{\beta}_{LQAD}$ have a run-time of $O(n^2 \ln^2 n)$ or higher, see [2].

Croux et al. [17] proposed the generalization of this estimator, namely **generalized S-estimator** (GS-Estimator). Berrendero [4] studied the GS-estimators

robustness and Roelant et al. [75] introduced the GS-estimators for the multivariate regression model.

The **least trimmed differences** (LTSSD), proposed by Stromberg et al. [86] in 2000, also exploits the pairwise differences minimizing the sum of the smallest quartile of the squared differences of the residual pairs.

$$\hat{\beta}_{LTSSD} = \sum_{i=1}^{k} r^2(\beta')_{(i)}, \tag{8.20}$$

where $r^2(\beta')_{(k)}$ denotes the k-smallest element from $\binom{n}{2}$ elements of the set

$$\begin{aligned} &\left\{[r_i(\beta', \alpha') - r_j(\beta', \alpha')]^2 \; ; \; 0 \leq j < i \leq n\right\} \\ =&\left\{[y_i - \beta' \cdot x_i - y_j + \beta' \cdot x_j]^2 \; ; \; 0 \leq j < i \leq n\right\}. \end{aligned} \tag{8.21}$$

It is unbiased if the ϵ_i's are independent and approximately identically distributed such that $E(\epsilon_i)$ and $E(\epsilon_i^2)$ exist. The breakdown of this point is 50% if k is defined as (8.19) with asymptotic efficiency of 0.66 at Gaussian noise. However, it is computationally expensive. $\hat{\beta}_{LTSSD}$ has a run-time complexity $O(n^4 \ln^2(n))$ by adapting algorithms for $\hat{\beta}_{LTSD}$. Nevertheless, for $k = n$, the (8.20) (no trimmed) is equivalent to

$$\hat{\beta}_{LSSD} = \frac{\sum\limits_{1 \leq i < j \leq n} [y_j - y_i] \cdot [x_j - x_i]}{\sum\limits_{1 \leq i < j \leq n} [x_j - x_i]^2}. \tag{8.22}$$

which is an arithmetic form and quadratic in the number of observations, but then its breakdown point is 0%.

8.3 Differences-rate estimator

Let me introduce a definition used in the assumptions of the differences-rate estimator.

Definition 8.2: *A set of values $x_1 \leq x_1 \leq \ldots \leq x_n$ are in **n-general position** if there exists a pair of values x_i and x_j in the set such that $x_i \neq x_j$ for some indexes $1 \leq i < j \leq n$.*

Definition 8.3: *Assume the simple linear regression model with $n > 2$ observations such that $x_1 \leq x_2 \leq \ldots \leq x_n$ are in n-general position. Define the* ***differences-rate estimator*** *as*

$$\hat{\beta}_{DR} = \frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[y_j - y_i]}{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]} \tag{8.23}$$

Proposition 8.1. *The differences-rate estimator for simple linear regression model is equivalent to*

$$\hat{\beta}_{DR} = \frac{\sum_{i=1}^{n}[2\cdot i - n - 1]\cdot y_i}{\sum_{i=1}^{n}[2\cdot i - n - 1]\cdot x_i} \tag{8.24}$$

Proof. I prove (8.1) by induction on the number n. Since the numerator and denominator of (8.24) are dual, I will prove the identity for the denomiator. That is,

$$\sum_{j=2}^{n+1}\sum_{i=1}^{j-1}[x_j - x_i] = \sum_{i=1}^{n+1}[2\cdot i - [n+1] - 1]\cdot x_i. \tag{8.25}$$

Trivially, (8.24) holds for $n = 2$. Suppose that it holds for

$$\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i] = \sum_{i=1}^{n}[2\cdot i - n - 1]\cdot x_i. \tag{8.26}$$

Now consider

$$\begin{aligned}\sum_{j=2}^{n+1}\sum_{i=1}^{j-1}[x_j - x_i] &= \sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i] + \sum_{i=1}^{n}[x_{n+1} - x_i] \\ &= \left[\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]\right] + n\cdot x_{n+1} - \sum_{i=1}^{n} x_i\end{aligned} \tag{8.27}$$

By grouping the first term with the third term, we obtain

$$\sum_{j=2}^{n+1}\sum_{i=1}^{j-1}[x_j - x_i] = \left[\sum_{i=1}^{n}[x_i\cdot[2\cdot i - n - 1] - x_i]\right] + n\cdot x_{n+1} \tag{8.28}$$

The proof is concluded by rewriting

$$n \cdot x_{n+1} = [2 \cdot [n+1] - [n+1] - 1] \cdot x_{n+1} \tag{8.29}$$

which is the $n+1$ term of (8.25). □

Theorem 8.1. *Assume the simple linear regression model such that $x_1 \leq x_2 \leq \ldots \leq x_n$ are in n-general position, and $\epsilon_1, \ldots, \epsilon_n$ are random variables independent and identically distributed with finite $E(\epsilon_i)$. Then,*

$$E(\hat{\beta}_{DR}) = \beta. \tag{8.30}$$

Therefore, $\hat{\beta}_{DR}$ is an unbiased estimator of β.

Proof. The expected value of (8.23) is given by

$$E\left(\hat{\beta}_{DR}\right) = E\left(\frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[y_j - y_i]}{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]}\right) = \frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}E\left(y_j - y_i\right)}{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]}. \tag{8.31}$$

Observe that

$$\begin{aligned} E\left(y_j - y_i\right) &= E\left(\beta \cdot x_j + \alpha + \epsilon_j - \beta \cdot x_i - \alpha - \epsilon_i\right) \\ &= E\left(\beta \cdot x_j + \epsilon_j - \beta \cdot x_i - \epsilon_i\right) \end{aligned} \tag{8.32}$$

for all pair i and j. Since ϵ_i and ϵ_j are independent observations, (8.32) is equivalent to

$$E\left(y_j - y_i\right) = \beta \cdot [x_j - x_i] + E(\epsilon_j) - E(\epsilon_i). \tag{8.33}$$

For identically distributed observations, $E(\epsilon_i) = E(\epsilon_j)$. Therefore,

$$E\left(y_j - y_i\right) = \beta \cdot [x_j - x_i]. \tag{8.34}$$

Thus, we conclude that

$$E\left(\hat{\beta}_{DR}\right) = \frac{\sum_{j=2}^{n}\sum_{i=1}^{j-1}E\left(y_j - y_i\right)}{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]} = \frac{\beta\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]}{\sum_{j=2}^{n}\sum_{i=1}^{j-1}[x_j - x_i]} = \beta. \tag{8.35}$$

□

Theorem 8.2. *Assume the simple linear regression model such that $x_1 \leq x_2 \leq \ldots \leq x_n$ are in n-general position, and $\epsilon_1, \ldots, \epsilon_n$ are random variables independent and identically distributed with finite $Var(\epsilon_i) = \sigma^2$. Then,*

$$Var(\hat{\beta}_{DR}) = \frac{[n-1] \cdot n \cdot [n+1] \cdot \sigma^2}{3\left[\sum_{i=1}^{n} [2i - n - 1] \cdot x_i\right]^2}. \tag{8.36}$$

Proof. Observe that

$$Var(y_i) = Var(\beta \cdot x_i + \alpha + \epsilon_i) = Var(\epsilon_i) = \sigma^2, \tag{8.37}$$

and

$$Var(\hat{\beta}_{DR}) = Var\left(\frac{\sum_{i=1}^{n} [2i - n - 1] \cdot y_i}{\sum_{i=1}^{n} [2i - n - 1] \cdot x_i}\right) = \frac{\sum_{i=1}^{n} [2i - n - 1]^2 \cdot Var(y_i)}{\left[\sum_{i=1}^{n} [2i - n - 1] \cdot x_i\right]^2}. \tag{8.38}$$

Simplifying the summatory in the numerator

$$\begin{aligned} \sum_{i=1}^{n} [2i - n - 1]^2 &= 4\left[\sum_{i=1}^{n} i^2\right] - 4[n+1]\left[\sum_{i=1}^{n} i\right] + n \cdot [n+1]^2 \\ &= \frac{[n-1] \cdot n \cdot [n+1]}{3}, \end{aligned} \tag{8.39}$$

where we have used the identities

$$\sum_{i=1}^{n} i = \frac{n \cdot [n+1]}{2}, \tag{8.40}$$

and

$$\sum_{i=1}^{n} i^2 = \frac{n \cdot [n+1] \cdot [2n+1]}{6} \tag{8.41}$$

Therefore, we conclude the proof using (8.39) in (8.38). □

Theorem 8.3. *The efficiency of $\hat{\beta}_{DR}$ at Gaussian noise is given by*

$$Efficiency(\hat{\beta}_{DR}) = \frac{3\left[\sum_{i=1}^{n}[2i-n-1]\cdot x_i\right]^2}{[n-1]\cdot[n+1]\cdot\left[n\sum_{i=1}^{n}x_i^2-\left[\sum_{i=1}^{n}x_i\right]^2\right]}. \tag{8.42}$$

Proof. The efficiency of $\hat{\beta}_{DR}$ derives directly from the ratio (8.5) to (8.42) □

Corollary 8.1. *Assume a simple linear regression model where the observations have the form $(x, y_1), (2\cdot x, y_2), \ldots (n\cdot x, y_n)$, such that the ϵ_i's are independent and identically distributed with finite $E(\epsilon_i)$. Then, β is unbiased estimated by*

$$\hat{\beta}_{DR} = \frac{6\cdot\sum_{i=1}^{n}[2\cdot i-n-1]\cdot y_i}{x\cdot[n-1]\cdot n\cdot[n+1]}. \tag{8.43}$$

Proof. Corollary 8.1 is derived from (8.24) and the identity

$$\begin{aligned}\sum_{i=1}^{n}[2i-n-1]\cdot i\cdot x &= x\cdot\left[2\left[\sum_{i=1}^{n}i^2\right]-[n+1]\left[\sum_{i=1}^{n}i\right]\right] \\ &= \frac{x\cdot[n-1]\cdot n\cdot[n+1]}{6}\end{aligned} \tag{8.44}$$

□

Corollary 8.2. *Assume a simple linear regression model where the observations have the form $(x, y_1), (2\cdot x, y_2), \ldots (n\cdot x, y_n)$. such that the ϵ_i's are independent and identically distributed with finite $E(\epsilon_i)$ and $Var(\epsilon_i)$. Then,*

$$Var\left(\hat{\beta}_{DR}\right) = \frac{12\sigma^2}{x^2\cdot[n-1]\cdot n\cdot[n+1]}, \tag{8.45}$$

which is identical to the Cramer-Rho lower bound at Gaussian noise. Therefore, $\hat{\beta}_{DR}$ is an efficient estimator of β.

Proof. Direct calculus yields

$$Var\left(\hat{\beta}_{DR}\right) = Var\left(\frac{6 \cdot \sum_{i=1}^{n} [2 \cdot i - n - 1] \cdot y_i}{x \cdot [n-1] \cdot n \cdot [n+1]}\right) = \frac{6^2 \cdot \sum_{i=1}^{n} [2 \cdot i - n - 1]^2 \cdot Var(y_i)}{[x \cdot [n-1] \cdot n \cdot [n+1]]^2} \tag{8.46}$$

To simplify (8.46) I used the identity (8.39). □

Unfortunately, the breakdown point of Differences-Rate estimator is zero because $\tilde{y} \to \infty$ if any $y_i \to \infty$.

8.4 Complexity and computational stored cost

The complexity of $\hat{\beta}_{DR}$ computed with (8.24) is $O(n)$, where n is the number of observations.

Since a single variable overflow in running-time could crash the whole system, the computational stored cost of variables is an important matter for applications where values are computed from a large amount of data. For instance, in the standard programming language of C++, if two variables x and y are integers of 32 bits, then the sum $x + y$ may result in an integer higher than 32 bits, in which case, a variable overflow will happen if $x + y$ is assigned to a variable of 32 bits or less and, as a result, the calculation of any variable which depends on this sum will fail.

The following definitions formalize the **stored cost of variable**.

Definition 8.4: *The **precision** of $\Lambda(x)$ is defined as the number of bits used to store the value x.*

Proposition 8.2. *$\Lambda(x)$ fulfills the following properties for x, y integers:*

1. $\Lambda(x + y) \leq \max\{\Lambda(x), \Lambda(y)\} + 1$.
2. $\Lambda(-x) \leq \Lambda(x) + 1$.

3. $\Lambda(x \cdot y) \leq \Lambda(x) + \Lambda(y)$.

Proof. Without losing generality, assume $x \geq y > 0$ such that $\Lambda(x) = n$ and $\Lambda(y) = m$.

1) Let n = m. Then,

$$x \leq 2^n - 1 \quad \text{and} \quad y \leq 2^n - 1 \Rightarrow x + y \leq 2^{n+1} - 2 \tag{8.47}$$

2) Computationally, a variable needs an extra bit to store the number sign when it can be either plus or minus.

3) Assume $\Lambda(x) = n$ and $\Lambda(y) = m$. Then,

$$x \leq 2^n - 1 \quad \text{and} \quad y \leq 2^m - 1 \Rightarrow x \cdot y \leq 2^{n+m} - 2^n - 2^m + 1. \tag{8.48}$$

□

Definition 8.5: *Define* $\tilde{\Lambda}(\frac{x}{y}) = \max\{\Lambda(x), \Lambda(y)\}$ *as the maximum number of bits stored in x and y in order to compute* $\frac{x}{y}$.

A computational advantage of $\hat{\beta}_{DR}$ over $\hat{\beta}_{LSS}$ is that $\tilde{\Lambda}(\hat{\beta}_{DR}) < \tilde{\Lambda}(\hat{\beta}_{LSS})$.

Proposition 8.3. *Suppose that* $\Lambda(x_i), \Lambda(y_i) \leq a$ *for* $i = 1, \ldots n$, $\Lambda(n) \leq b$, *and* $\hat{\beta}_{DR}$ *is computed by (8.24). Then,* $\tilde{\Lambda}(\hat{\beta}_{DR}) \leq a + 2b + 1$.

Proof. Note that

$$-n < 2 \cdot i - n - 1 < n \quad \text{for } i = 1, \ldots, n \tag{8.49}$$

then

$$\Lambda(2 \cdot i - n - 1) \leq \Lambda(-n) \leq b + 1. \tag{8.50}$$

Without losing generality assume $\Lambda(\tilde{y}) > \Lambda(\tilde{x})$. Let

$$z = \underset{y_i,\, i=1,\ldots,n}{\arg\max} \{\Lambda(y_i)\} \tag{8.51}$$

be the variable with the maximum stored cost from the sample. Thus

$$\tilde{\Lambda}(\hat{\beta}_{DR}) = \max\{\Lambda(\tilde{y}), \Lambda(\tilde{x})\} = \Lambda(\tilde{y})$$

$$\leq \Lambda\left(\sum_{1}^{n}[-n] \cdot z\right) = \Lambda\left(n \cdot [-n] \cdot z\right) \Lambda(n) + \Lambda(-n) + \Lambda(z) = b + (b + 1) + a. \tag{8.52}$$

□

Similarly, I calculated $\tilde{\Lambda}(\hat{\beta}_{LSS}) = \tilde{\Lambda}(\hat{\beta}_{LSSD}) \leq 2a + 2b + 1$ according (8.9) and (8.22).

8.5 Application in power-law distributions

The populations of cities, the intensities of earthquakes, and the sizes of power outages, for example, are all thought to have **power-law distributions** [57]. Quantities such as these are not well characterized by their typical or average values. For instance, according to the Mexican Census (1995) [31], the average population of a city, town, or village in Mexico is around 453. But this statement is not a useful one for most purposes because a significant fraction of the total population lives in cities whose population is larger by several orders of magnitude, like Mexico City, in which more than 8.84 million people live.

power law distribution

Power-law distributions is a family of statistical distributions, such as **Pareto** and **Zipf distribution**, where values with extreme deviation of the median have a significant probability of being observed. Such distributions lead to much heavier tails than other common models, such as **exponential distributions**. Mathematically, a quantity $x \geq x_{min} > 0$ obeys a **power law** if it is drawn from a probability distribution function

$$\Pr(X = x) = \psi(x) = c \cdot x^{-\alpha}, \tag{8.53}$$

where $\alpha > 0$ is a constant parameter of the distribution known as the **exponent** or **scaling parameter** and $c > 0$ is the normalization constant.

Power-law distributions occur in diverse models of pattern recognition and computer vision [7], [25], [74]. However, the estimation of the parameters of a power law distribution from observed data is a serious challenge if the measured quantities are noisy. Hence the importance of using robust estimators.

8.5.1 Estimators for the exponenent of a power-law distribution

The **maximum likelihood estimator** of α gives an accurate parameter estimate in the limit of large sample size. For the continuous case, this estimator is given by

$$\hat{\alpha} = 1 + n \cdot \left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{min}} \right) \right]^{-1}, \tag{8.54}$$

where $x_1, \ldots, x_n$ are drawn from a power law distribution such that $x_i \geq x_{min}$. An estimate of the standard error $\hat{\sigma}$ on $\hat{\alpha}$ is

$$\hat{\sigma} = \frac{\hat{\alpha} - 1}{\sqrt{n}}. \tag{8.55}$$

I strongly recommend the publications by Clauset et al. [15], and Newman [58] for a useful discussion of these and related points.

An alternative method to estimate α is based on the linearity of the **complementary cumulative distribution function** on logarithmic scales. That is,

$$\Pr(X \geq x) = \Psi^c(x) = \int_x^\infty c \cdot y^{-\alpha}\, dy = \frac{c}{(\alpha - 1)} \cdot x^{-(\alpha-1)} \tag{8.56}$$

$$ln\,(\Psi^c(x)) = -(\alpha - 1) \cdot \ln(x) + constant. \tag{8.57}$$

Thus, the parameter α can be estimated from the absolute slope of the **empirical complementary cumulative distribution function** on a doubly logarithmic plot, which is an estimate of (8.57). However, such a graph should be truncated in order to avoid the noise introduced by fluctuations in its right tail. Then, the truncated empirical complementary cumulative distribution function is given by

$$\widehat{\Psi}^c(x) = |\{x_i \mid x_i \leq x\}|, \quad \text{for } \widehat{\Psi}^c(x) > y_{min} \tag{8.58}$$

where $x_1 \leq x_2 \leq \ldots \leq x_n$ are the observed measurements, and y_{min} is a parameter. Experimentally, I computed good results with $y_{min} = 0.01$.

The simple linear regression model for (8.56) can potentially lead us to estimates with large bias. The noise, for instance, could obey a distribution whose expected value is different to zero or does not exist, like the **Cauchy distribution** or for some parameters of the **Pareto distribution**.[3]

Clauset et al. [15] pointed out that the assumptions to calculate the **standard error** on the slope of a regression line, which include independent and Gaussian noise in the dependent variable at each value of the independent variable, do not hold for (8.56). In fact, assuming a Gaussian noise in the observed samples $x_1, \leq \ldots, \leq x_n$; $\hat{\Psi}^c(x_i)$ will have a Gaussian noise but the noise in the logarithm is not Gaussian. Furthermore, the assumption of independence fails because $\hat{\Psi}^c(x_i) = \hat{\Psi}^c(x_{i+1}) + \hat{\psi}(x_i)$ for $x_i < x_{i+1}$, where $\hat{\psi}(x_i)$ is the empirical probability density function, and hence adjacent values of the empirical complementary cumulative distribution function are strongly correlated. However, I will show that estimators based on linear regression are more robust than the maximum likelihood estimator when the data has been contaminated. For that, I tested the ability of the least sum of squares, the least sum of squared differences, and the differences-rate estimator to extract a known exponent parameter from noisy synthetic power-law distributed data. I also addressed the maximum likelihood estimator as a reference point.

The **standard error** is an alternative name of the muestral standard deviation

[3]The probability density function of the Pareto distribution is $f(x; k, \alpha) = \alpha \cdot k^\alpha \cdot x^{-\alpha-1}$, where $\alpha, k > 0$ are parameters. The Pareto distribution has infinite variance if $0 < \alpha \leq 2$. If $\alpha \leq 1$, also has infinite mean.

8.5.2 Noisy measurements from power-law distributed data

Noisy measurements of the true quantities of the phenomenon come in three basic flavors: all are underestimated values (**noise type left**), all are overestimated values (**noise type right**), and there are both underestimated and overestimated values (**noise type left-right**). In this context, "*underestimated value*" means that at most the value equals the true value; "*overestimated value*" means that at least the value equals the true value

I model underestimated and overestimated measurements as follows:

$$x_i = [1 + u_i] \cdot v_i \tag{8.59}$$

where x_i is the i-observed measurement, u_i is a random variable uniformly distributed in $[a, b]$, and v_i is the true i-quantity of the phenomenon which obeys a power-law distribution.

The factor $[1 + u]$ in (8.59) models a complete ignorance of the noise distribution in the measurements. However, this model restricts x_i between the lower bound $[1 + a] \cdot v_i$ and the upper bound $[1 + b] \cdot v_i$. Therefore, a and b determine which kind of noisy measurements there are in the observations. For example, all our measurements are overestimated if $0 \leq a, b$.

In the following paragraphs I state some real and hypothetical examples where the measurements are underestimated and overestimated.

The **net wealth** of an individual is the total of his or her assets minus the total his or her debts.

The **net wealth** in US dollars of the richest individuals in the wold is an example where noisy measurements are both underestimated and overestimated values. **Forbes Magazine**, for instance, publishes a ranking of the world's billionaires annually. This ranking is based on the net wealth of each individual. The complementary cumulative histogram of the 24th edition of this ranking[4] appears to obey a power-law distribution; see Fig. 8.1. However, these data are biased. Forbes Magazine says that there are billionaires that may not be in the list since some billionaires were not detected by their reporters. Another reason for this bias is that some billionaires cooperate to assess their fortune, but others do not.[5] Therefore, for those billionaires who cooperated, their fortune is underestimated since they may not (be able to) report all that they own. However, the fortune of those billionaires that did not cooperate may be either underestimated, or overestimated.

[4] This ranking can be found in: http://www.forbes.com/2010/03/10/worlds-richest-people-slim-gates-buffett-billionaires-2010_land.html Readers interested in this data can mail me.

[5] Forbes's methodology can be found in: http://www.forbes.com/forbes/2010/0329/billionaires-2010-wealth-estates-stocks-yachts-fortunes-methodology.html

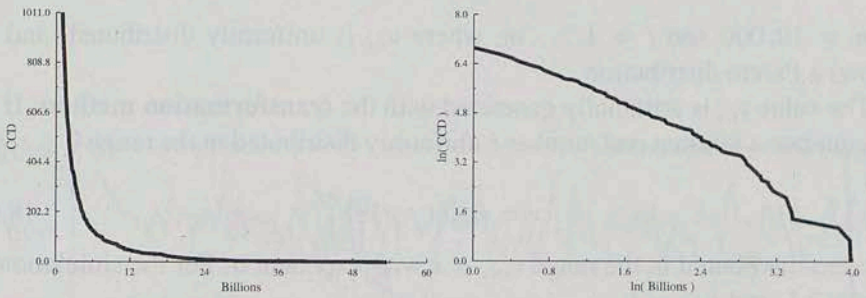


Figure 8.1 – *On the left, histogram of the number of billionaires that have a net worth equal or higher than x according to the list of the world's billionaires (February 2010, Forbe's 24th ranking). On the right, histogram of the same data, but plotted on logarithmic scales*

Underestimated measurements may occur in phenomena where the quantity is the number of existing "*things*" of "*something*". For example, as first observed by de Solla [18], the numbers of citations received by scientific papers appear to have a power-law distribution. These data are strictly underestimated because, in the practice, a significant number of publications is not indexed and not all citations can be extracted from a manuscript because of inconsistent references and digitalization problems (low OCR accuracy, wrong text extraction, to mention some).

Overestimated measurements may occur in phenomena where their quantities are measurable only after they take place, for example, hypothetically speaking, the time for detecting the presence of a disease. Overestimated measurements may also occur in measuring techniques which intrinsically overestimate the phenomenon quantities. For example, the area of a bacterium may be approximated for a convex hull that contains the whole bacterium.

8.5.3 Simulations of noisy measurements

For each estimator method, I computed several simulations with different types and levels of noise. A simulation consists of $m = 1,000$ data sets. Each data set has the form

$$\begin{aligned} X_i &= \{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\} \\ &= \{[1 + u_{i,1}] \cdot v_{i,1}, [1 + u_{i,2}] \cdot v_{i,2}, \ldots, [1 + u_{i,n}] \cdot v_{i,n}\} \end{aligned} \tag{8.60}$$

for $n = 10,000$ and $i = 1, \ldots, m$, where $u_{i,j}$ is uniformly distributed, and $v_{i,j}$ follows a Pareto distribution.

The value $v_{i,j}$ is artificially generated with the **transformation method**: If we can generate a random real number r uniformly distributed in the range $0 \leq r < 1$, then

$$v = v_{min}(1 - r)^{\frac{-1}{\alpha - 1}} \tag{8.61}$$

is Pareto-distributed in the range $v_{min} \leq x$ with exponent α. For the simulations, I set $\alpha = 2.5$, and $x_{min} = 1$.

The noise $u_{i,j}$ is uniformly distributed between $[a, b]$ for all data set in a simulation, where a and b depend on the type of noise that the simulation performs:

- Noise type left: $u_{i,j} \sim U(-\delta, 0)$
- Noise type right: $u_{i,j} \sim U(0, \delta)$
- Noise type left-right: $u_{i,j} \sim U(-\delta, \delta)$

where δ is the noise level in the simulation.

In each simulation, the absolute error $e_i = \|\hat{\alpha}_i - \alpha\|$ for $i = 1, 2, \ldots m$, was computed, where $\hat{\alpha}_i$ is the estimated value of the exponent parameter from X_i.

The mean and standard deviation of e_i's of each simulation are show in Fig. 8.2, Fig. 8.3, and Fig. 8.4.

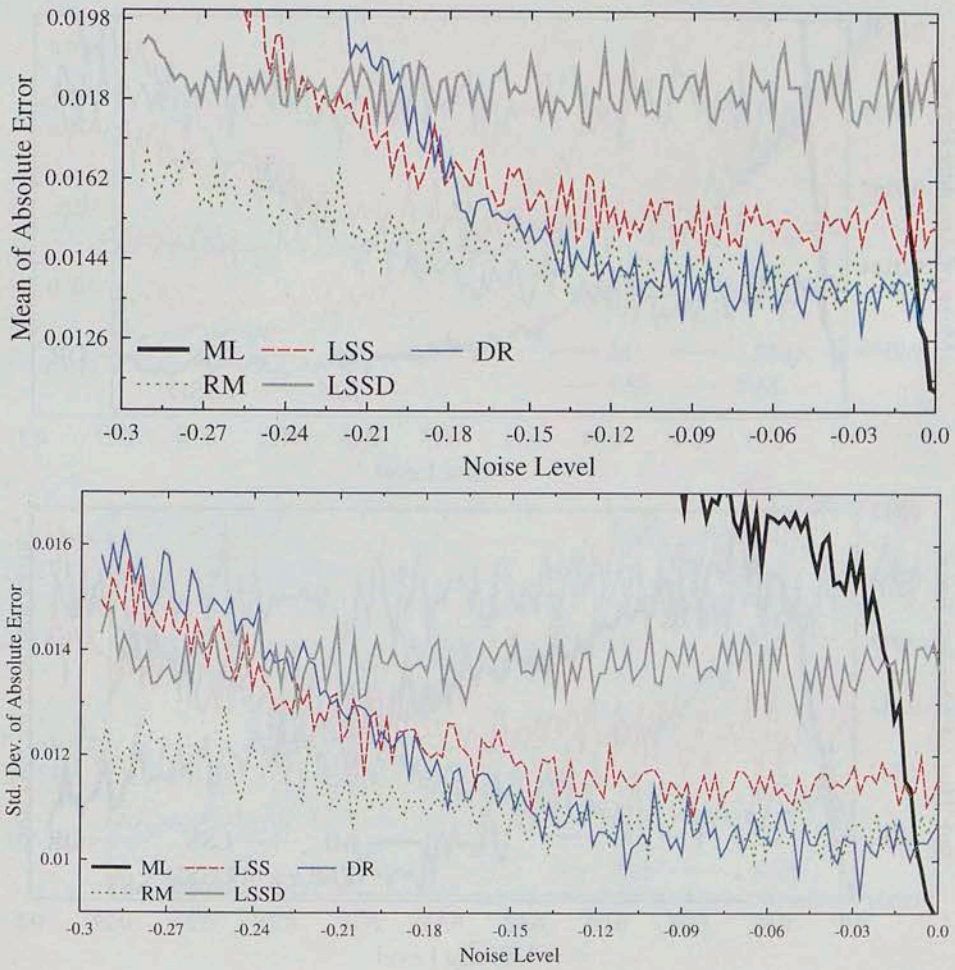


Figure 8.2 – *At the top, mean of absolute error from samples with noise type left. On the bottom, standard deviation of absolute error from samples with noise type left.*

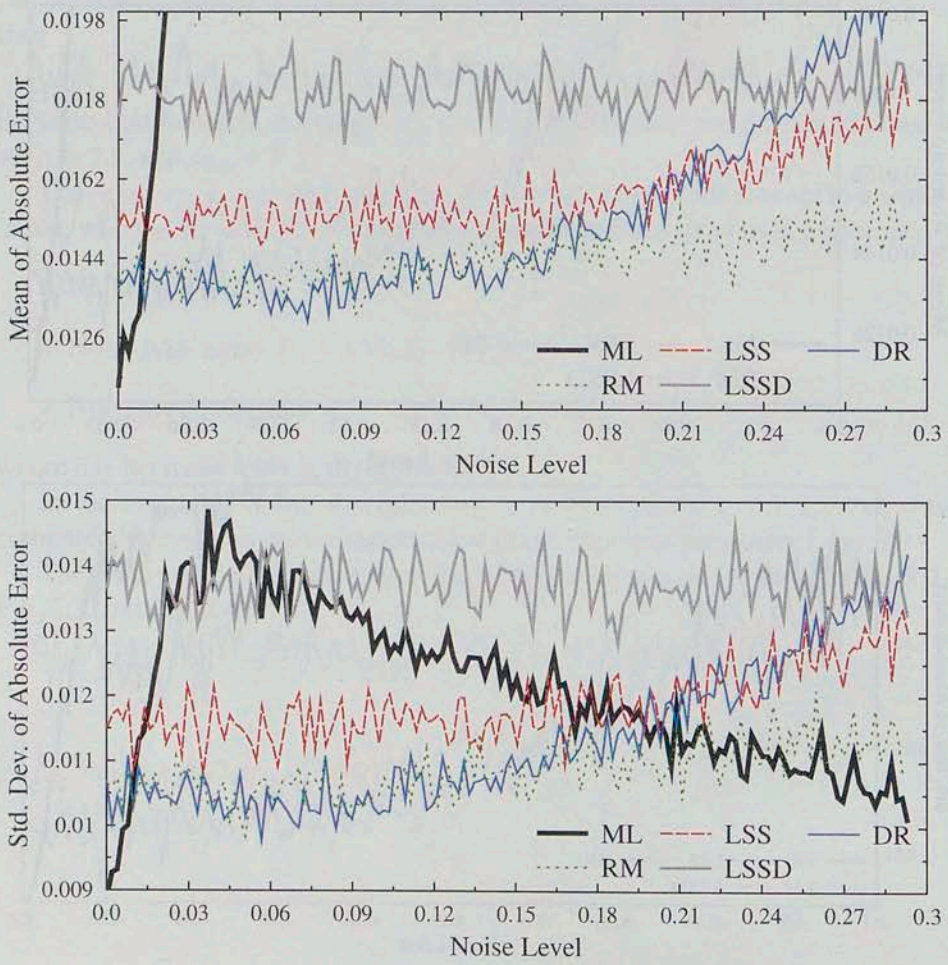


Figure 8.3 – *At the top, mean of absolute error from samples with noise type right. On the bottom, standard deviation of absolute error from samples with noise type right.*

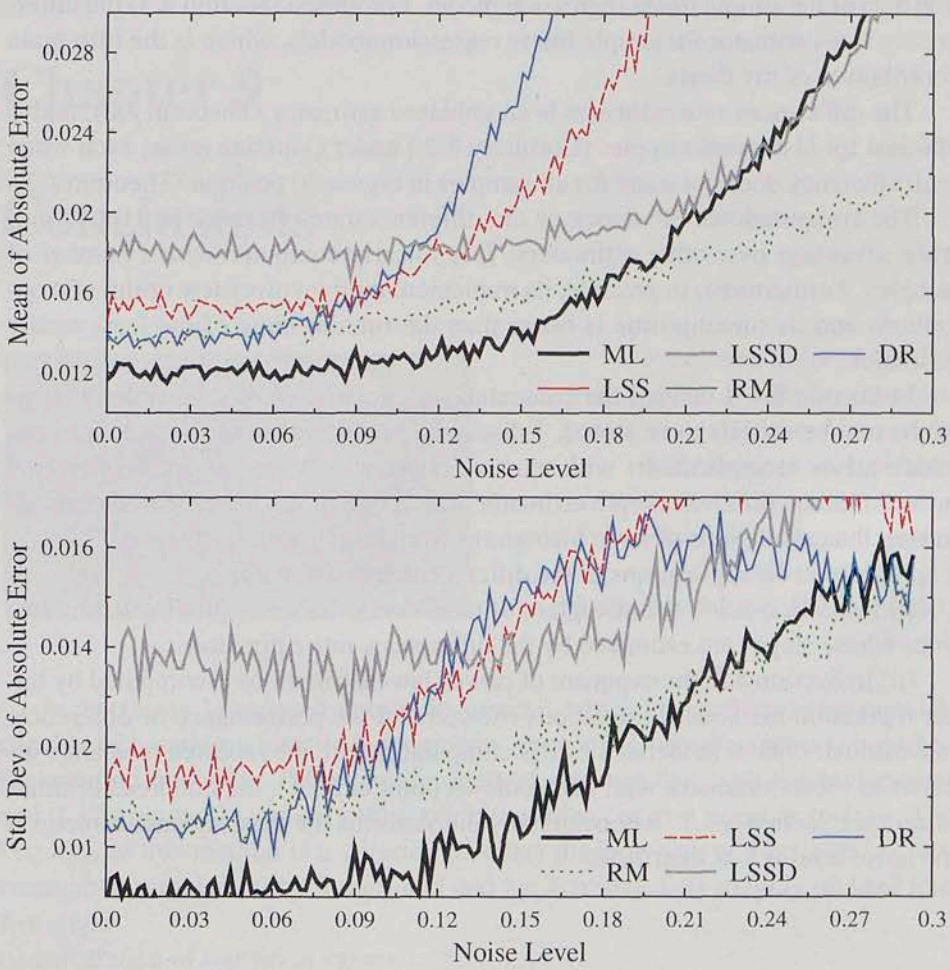


Figure 8.4 – *At the top, mean of absolute error from samples with noise type left-right. On the bottom, standard deviation of absolute error from samples with noise type left-right.*

8.6 Summary

After I reviewed preliminary concepts (Section 8.1) and the state of the art (Section 8.2) of the simple linear regression model, I proposed (Section 8.3) the differences - rate estimator for simple linear regression models, which is the fifth main contribution of my thesis.

The differences-rate estimator is an unbiased estimator (Theorem 8.1), and is efficient for histogram samples (Corollary 8.2) under Gaussian noise, even when such efficiency does not stand for all samples in n-general position (Theorem 8.3).

The computational efficiency of the difference-rate estimator is its most notable advantage over other estimators. It is linearly computed with a number of samples. Furthermore, in practice, its implementation involves few arithmetic operations and its running-time is better than the running-time of the least square estimator.

In Section 8.4, I defined the computational stored cost of a variable in terms of the number of bits to be stored. I discussed the relevance of these definitions, which arises in applications with massive numerical elements. Subsequently, I proved that the differences-rate estimator stands one of the lower stored costs allowing the calculations of large histograms from large numeric elements.

I showed two applications of the differences-rate estimator:

(i) In Section 5.2.3, the histogram of transition values is approximated by two lines whose slopes are estimated by the differences-rate estimator.

(ii) In Section 8.5, the exponent of power-law distributions is computed by linear regression methods. Simulations showed that the performance of difference-rate estimator under moderate noise is comparable with the repeated medians estimator (a robust estimator with a breakdown point of 50%), but hundreds of times faster; see Section 8.5.3. It considerably outperforms the least square estimator if the noise level is less than 20%.

Chapter 9

Conclusions

Pretty and beloved Mexico
If I die far from you
Say that I am asleep
And bring me back to here.

By Jesús Monge Ramírez
(Chucho Monge)
Mexican composer (1910 – 1964)

In this thesis, I proposed a novel framework, which I named transition method, capable of binarizing historical documents more efficiently than other top-ranked binarization methods. The transition method assumes that both the background and the foreground vary smoothly, exhibiting high contrast at the boundary. The key idea of this method is a criterion to select pixels which will be taken as representative samples of the foreground and background. It is roughly divided into five steps:
(i) calculation of transition values,
(ii) calculation of transition thresholds,
(iii) restoration of transition sets,
(iv) detection of regions of interest, and
(v) calculation of thresholds of gray intensities.

In Chapter 4, I mathematically modeled the distribution of gray intensities. As first suggested by Chow and Kaneko [14], gray intensities appear to obey a normal distribution. Indeed, experimental observations in historical documents

confirm such a conjecture but are restricted to small neighborhoods. Following this line, contrast and smoothness play the most important role in my approach rather than spatial relationships between pixels. Even though spatial relationships are ignored, my proposed model is capable of determining certain bounds and properties which led me to propose the transition method.

The strength of the transition method stems not only in its images modeling, but also in its capacity of "*plugging*" different models in each method's stages. For example, to compute transition thresholds, I first proposed the quantile transition method, which has a crucial parameter. However, in further publications, I proposed the double linear and Rosin's transition thresholds, both of which lack parameters and, as a consequence, are suitable methods for unsupervised applications.

I concluded that the restoration of transition sets is critical in images with high levels of noise. In particular, isolated, incidence, and dilation transition operators may be applied (in that order) to enhance the transition sets. This combination tends to improve our transition set approximation. Moreover, it is robust for different levels of noise.

I derived the dilation transion operator from the concept of transition balance (Definition 5.3). However, such a concept was not fully justified, and further techniques may be developed from it, like a direct binarization and edge detection method.

Comparative studies in Chapter 7 strongly indicate that the transition method performs better with the lognormal threshold than with the normal threshold, even when the transition thresholds were computed with different methods. Hence, I conjectured in [72], [73], and [71] that the gray intensities of transition pixels are lognormally distributed rather than normally distributed. However, this conjecture is contrary to empirical observations (gray intensities are normally distributed). I suspect that this pattern could be due to maxmin function, and/or due to sampling process in the very boundary between the foreground and background. The gray intensities of pixels along boundaries may obey a distribution that is not Gaussian.

Although the transition method has promising results in historical documents, it cannot cope with sudden illumination changes, and with large isolated bleed-through artifacts. But in fact, none of the binarization methods described in this thesis can cope with such problems. The transition method has the potential of overcoming such problems by developing extended techniques in the transition set restoration, region of interest detection, and gray intensities thresholds.

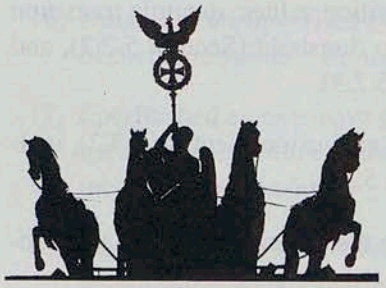
In this thesis, I also studied unsupervised measures for segmentation quality based on variances of gray intensities. Technical conclusions are widely discussed

in Section 6.2 and Section 7.4.3. Nevertheless, I would like to remark that values of unsupervised measures may be used to compare the performance of two different parameterizations of a single algorithm rather than comparing the performance of two different algorithms. An unsupervised measure may not "*share*" the same assumptions as the evaluated binarization method. As I showed in Section 7.4, certain unsupervised measures are unsuitable to evaluate the performance of certain binarization methods.

In Chapter 8, I proposed the differences-rate estimator, which is an unbiased estimator for the slope in simple linear regression models. It can accurately estimate the slope on histograms of empirical complementary cumulative distribution functions where the effect of outliers had faded. Moreover, its alternative form is linearly computed in the number of samples and, hence, it is suitable for estimating the slope of lines in large histograms with extreme values, and for time-consuming algorithms.

Chapter 10

Summary of contributions



When you want something, all the universe conspires in helping you to achieve it.

The Alchemist by Paulo Coelho
Brazilian lyricist and novelist (1947 –)

In this thesis, I proposed a novel approach for binarization, edge detection, and the detection of region of interest. Additionally, I proposed novel unsupervised measures to evaluate the binarization performance, a novel slope estimator, and a novel statistical test for pairwise comparisons. In concrete terms:

1. I proposed the ***t*-transition pixels**, a generalization of **edge pixels**; see Definition 4.4 and Definition 4.5.

2. I proposed the term **ideal image** based on smooth surfaces and contrast; see Definition 4.2 and Definition 4.3.

3. I defined the **transition functions** and characterized the transition pixels with extreme values for those functions; see Section 4.3 and Definition 4.10.

4. I proved that the function maxmin is a transition function in ideal images; see Section 4.4 and Theorem 4.1.

5. I pointed out how the statistical distribution of gray intensities of transition sets approximate the statistical distribution of gray intensities of the foreground and background; see Section 5.1.

6. I proposed and described the transition method with five steps:

   (a) calculus of transition values,

   (b) selection of transition thresholds,

   (c) restoration of transition set,

   (d) detection of region of interest, and

   (e) binarization, or edge detection.

7. I proposed three novel thresholding for transition values: quantile transition threshold (Section 5.2.1), Rosin's transition threshold (Section 5.2.2), and double-linear transition threshold (Section 5.2.3).

8. I proposed three novel transition operators: expansion (Section 5.3.2), incidence (Section 5.3.3), and dilation (Section 5.3.4).

9. I proposed two simple criteria for detecting the region of interest; see Section 5.4.

10. I proposed a simple algorithm for edge detection; see Section 5.6.

11. I proposed several algorithms for binarization. Particularly, I described:

    (a) linear mean-variance threshold (Section 5.5.1),

    (b) autoliear threshold (Section 5.5.2),

    (c) minimum-error-rate threshold (Section 5.5.4),

    (d) normal threshold (Section 5.5.5), and

    (e) lognormal threshold (Section 5.5.6).

12. I introduced in Section 6.1 the concept of simple images (Definition 6.1).

13. I statistically analized local implementations of three well-known unsupervised measures:

    (a) uniformity measure (Section 6.2.1),

(b) region non-uniformity measure (Section 6.2.2), and

(c) weighted variance measure (Section 6.2.3).

14. I proposed four novel evaluation measures for binarization:

    (a) normal uniform variance (Section 6.2.4),

    (b) unbiased weighted variance (Section 6.2.5),

    (c) lognormal uniform variance (Section 6.2.6), and

    (d) lognormal weighted variance (Section 6.2.6).

15. I proved in Theorem 6.1 that the expected value of the unbiased weighted variance measure is minimum in a perfect binarization.

16. I analyzed unsupervised evaluation measures by describing statistically which of them are suitable for nine binarization methods; see Section 7.4.

17. I performed an extensive comparison of several unsupervised measures, binarization algorithms , and OCRs. I used it to show the strength of the WV measure; Section 7.4.

18. I performed an extensive comparison between the transition method and several top-ranked binarization algorithms; see Section 7.5 and Section 7.6.

19. I proposed and described a novel estimator (differences-rate estimator) for the slope of the simple linear regression (Section 8.3).

20. I proved the computational goodness of differences-rate estimator; Section 8.4.

21. I showed a suitable application of the differences-rate estimator in power-law distributions; see Section 8.5.

22. I proposed a statistical test to compare measures based on an intuitive triad of possible results: better, worse, or comparable performance; Appendix C.

Appendix A

Integral Images

Ramírez-Ortegón et al. [72] extended the **integral image** [89] to compute efficiently any statistical moment in subsets of pixels in neighborhoods with radius r of an image F. This is particularly useful for the transition method, and for statistical binarization methods.

Definition A.1: ***The integral image*** $\tilde{F}_S$ *of a subset* $S \subset \mathcal{P}$ *in an image* F *is an image defined as*

$$\tilde{F}_S(\boldsymbol{p}_{i,j}) = \sum_{0 \le h \le i} \sum_{0 \le k \le j} F(\boldsymbol{p}_{i,j}) \cdot \mathbf{1}_S(\boldsymbol{p}_{i,j}), \tag{A.1}$$

where $\mathbf{1}_S(\boldsymbol{p}_{i,j})$ *denotes the indicator function*

$$\mathbf{1}_S(\boldsymbol{p}_{i,j}) = \begin{cases} 1 & \text{if } \boldsymbol{p}_{i,j} \in S \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

The efficiency of integral images emerges from the linearity of its calculation, see Fig. A.1, given by

$$\tilde{F}_S(\boldsymbol{p}_{i,j}) = F(\boldsymbol{p}_{i,j}) \cdot \mathbf{1}_S(\boldsymbol{p}_{i,j}) + \begin{cases} \tilde{F}_S(\boldsymbol{p}_{i,j-1}) & \text{if } j > 0 \\ \tilde{F}_S(\boldsymbol{p}_{i-1,j}) & \text{if } i > 0 \\ -\tilde{F}_S(\boldsymbol{p}_{i-1,j-1}) & \text{if } i > 0 \text{ and } j > 0 \end{cases} \tag{A.3}$$

As an immediate result of (A.3), the sum of $F(\boldsymbol{q})$ for $\boldsymbol{q} \in S_r(\boldsymbol{p}_{i,j})$ is computed

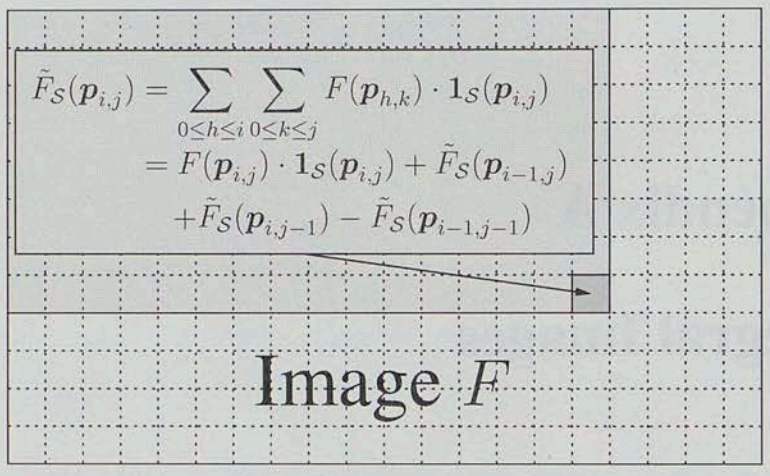


Figure A.1 – *Calculation of the integral image of a pixel in general position.*

as

$$\sum_{q\in\mathcal{S}_r(\boldsymbol{p}_{i,j})} F(\boldsymbol{q}) = \biguplus_{\mathcal{S}_r(\boldsymbol{p}_{i,j})} F = \tilde{F}_{\mathcal{S}}(\boldsymbol{p}_{h,w}) + \begin{cases} -\tilde{F}_{\mathcal{S}}(\boldsymbol{p}_{i,j-r-1}) & \text{if } j-r>0 \\ -\tilde{F}_{\mathcal{S}}(\boldsymbol{p}_{i-r-1,j}) & \text{if } i-r>0 \\ +\tilde{F}_{\mathcal{S}}(\boldsymbol{p}_{i-r-1,j-r-1}) & \text{if } {}^{i-r>0}_{j-r>0} \end{cases} \tag{A.4}$$

with $h = \max\left\{i+r, n_y\right\}$, and $w = \max\left\{j+r, n_x\right\}$, where n_y is the number of rows in F, and n_x is the number of columns. Figure A.2 shows the calculation (A.4) of a pixel in general position (Definition 2.10).

Remark 2.5 states that the frame isolate operator is quickly computed since the cardinality of any subset $\mathcal{S}$ of a rectangular partition $\mathcal{P}$ is computed by integral images in constant time. For this,

$$|\mathcal{S}_r(\boldsymbol{p})| = \sum_{q\in\mathcal{S}_r(\boldsymbol{p})} \mathbf{1}_{\mathcal{S}}(\boldsymbol{q}) = \biguplus_{\mathcal{S}_r(\boldsymbol{p})} \mathbf{1}_{\mathcal{S}} \tag{A.5}$$

Moreover, given any image F,

$$\mu_{F,\mathcal{S}_r(\boldsymbol{p})} = \frac{1}{|\mathcal{S}_r(\boldsymbol{p})|}\sum_{q\in\mathcal{S}_r(\boldsymbol{p})} F(\boldsymbol{q}) = \frac{\biguplus_{\mathcal{S}_r(\boldsymbol{p})} F}{\biguplus_{\mathcal{S}_r(\boldsymbol{p})} \mathbf{1}_{\mathcal{S}}} \tag{A.6}$$

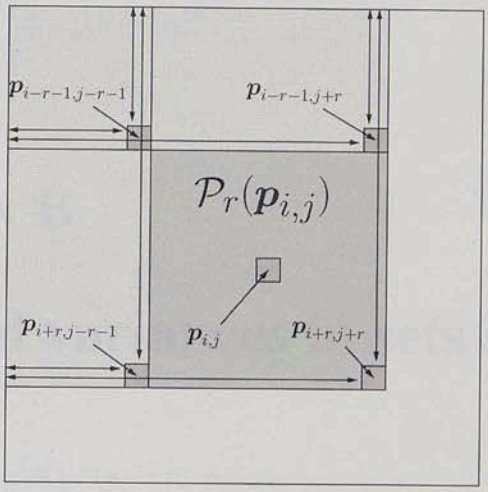


Figure A.2 – *Calculation of (A.4) for a pixel in general position.*

and

$$\sum_{q \in \mathcal{S}_r(p)} [F(q)]^2 = \biguplus_{\mathcal{S}_r(p)} F^2, \tag{A.7}$$

from which $\hat{\sigma}^2_{F,\mathcal{S}_r(p)}$ can be computed in constant time according (B.2) as

$$\begin{aligned} \hat{\sigma}^2_{F,\mathcal{S}_r(p)} &= \left[\frac{1}{|\mathcal{S}_r(p)| - 1} \sum_{q \in \mathcal{S}_r(p)} [F(q)]^2 \right] - \hat{\mu}^2_{F,\mathcal{S}_r(p)} \\ &= \frac{\biguplus_{\mathcal{S}_r(p)} F^2}{\biguplus_{\mathcal{S}_r(p)} \mathbf{1}_{\mathcal{S}} - 1} - \left[\frac{\biguplus_{\mathcal{S}_r(p)} F}{\biguplus_{\mathcal{S}_r(p)} \mathbf{1}_{\mathcal{S}}} \right]^2 \end{aligned} \tag{A.8}$$

Appendix B

Mean and variances in sets

Given $H_{F,\mathcal{A}}(i) = |\{\boldsymbol{p} \in \mathcal{A} \mid F(\boldsymbol{p}) = i\}|$, denote

- The **mean of F** in $\mathcal{A}$ as

$$\hat{\mu}_{F,\mathcal{A}} = \begin{cases} \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{p} \in |\mathcal{A}|} F(\boldsymbol{p}) & \text{if } \mathcal{A} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \tag{B.1}$$

- The **unbiased sample variance of F** in $\mathcal{A}$ as

$$\hat{\sigma}^2_{F,\mathcal{A}} = \begin{cases} \frac{1}{|\mathcal{A}|-1} \sum_{\boldsymbol{p} \in \mathcal{A}} F^2(\boldsymbol{p}) - \hat{\mu}^2_{\mathcal{A}} & \text{if } |\mathcal{A}| > 1 \\ 0 & \text{otherwise,} \end{cases} \tag{B.2}$$

  where $\hat{\mu}^2_{F,\mathcal{A}} = [\hat{\mu}_{F,\mathcal{A}}]^2$ and $F^2(\boldsymbol{p}) = [F(\boldsymbol{p})]^2$.

- The **biased sample variance of values F** in $\mathcal{A}$ as

$$S^2_{F,\mathcal{A}} = \begin{cases} \frac{1}{|\mathcal{A}|} \sum_{\boldsymbol{p} \in \mathcal{A}} F^2(\boldsymbol{p}) - \hat{\mu}^2_{F,\mathcal{A}} & \text{if } |\mathcal{A}| > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

- The **unbiased sample variance of logarithms** in $\mathcal{A}$

$$\tilde{\sigma}^2_{F,\mathcal{A}} = \ln\left(1 + \frac{\hat{\sigma}^2_{F,\mathcal{A}}}{\hat{\mu}^2_{F,\mathcal{A}}}\right). \tag{B.4}$$

Appendix C

Uncertainty test

In [70], I developed the **uncertainty test** to compare measures based on an intuitive triad of possible results: better, worse, or comparable performance.

Given an image, suppose that we are able to compare the performance of two methods x and y based on some criterion. In this context, performance means how well the method performs its task. Also suppose that there are only three possible outcomes of the method's comparison in a single image: *Method x better than method y* (E_1), *method y better than method x* (E_2), and *method x as good as method y* (E_3). Therefore, we ascertain that method x is better than method y in an image population if E_1 occurs more frequently than E_2. More formally, let $p_i = \Pr(E_i)$ for $i = 1, 2, 3$ be the probability of occurrence of E_i in an image which was randomly drawn from an image population. Then, our assessment is based on the numerical relation between p_1 and p_2.

Let the random variable N_i indicate the number of occurrences of E_i in a sample of n images which were independently and randomly drawn from a large population of images. Then, the triad (N_1, N_2, N_3) follows a trinomial distribution[1].

Assume that (n_1, n_2, n_3) is an observed vector of (N_1, N_2, N_3); the probability of observing (n_1, n_2, n_3) is given by

$$\begin{aligned}\psi(n_1, n_2, n_3; n, p_1, p_2, p_3) &= \Pr(N_1 = n_1, N_2 = n_2, N_3 = n_3) \\ &= \frac{n!}{n_1! \cdot n_2! \cdot n_3!} p_1^{n_1} \cdot p_2^{n_2} \cdot p_3^{n_3}\end{aligned} \tag{C.1}$$

where $\cdot!$ denotes the **factorial function**. Therefore, p_i can be estimated by $\hat{p}_i = \frac{n_i}{n}$.

Given a postive integer n, the factorial of n is defined as $n! = n \cdot [n-1] \cdot \ldots 2 \cdot 1$. Factorial of zero is defined as $0! = 1$.

[1] Technically speaking, this is sampling without replacement, so the correct distribution is the multivariate hypergeometric distribution, but the distributions converge as the population grows large.

Unfortunately, large samples to ensure convergence may be unavailable, and the probability of observing $\hat{p}_1 < \hat{p}_2$ may be significant if $p_1 - p_2 > 0$ is small.

The problem is then to measure how unlike $\alpha \cdot \hat{p}_1 \geq \hat{p}_2$ for $\alpha < 1$ is, given that $p_1 \leq p_2$. Therefore, the upper bound of $\Pr(\alpha \cdot \hat{p}_1 \geq \hat{p}_2 \mid p_1 \leq p_2)$ for all possible pairs $p_1 \leq p_2$ is the maximum probability of observing $\alpha \cdot \hat{p}_1 \geq \hat{p}_2$ while the true probabilities p_1 and p_2 are such that $p_1 \leq p_2$.

I named this probability as α-**uncertainty**, which can be estimated by

$$UN(n, \alpha) = \max_{(y_1, y_2) \in \mathcal{Y}} \left\{ \sum_{(x_1, x_2, x_3) \in \mathcal{X}} \psi(x_1, x_2, x_3; n, y_1, y_2, y_3) \right\} \tag{C.2}$$

where

$$\mathcal{Y} = \left\{ (y_1, y_2) \in \mathbb{R}^2 \mid 0 \leq y_1 \leq y_2 \leq 1 \text{ and } y_1 + y_2 \leq 1 \right\}, \tag{C.3}$$

$$y_3 = 1 - y_1 - y_2, \tag{C.4}$$

and

$$\mathcal{X} = \left\{ (x_1, x_2, x_3) \in \mathbb{N}^3 \mid \alpha \cdot x_1 \geq x_2 \text{ and } x_1 + x_2 + x_3 = n \right\}. \tag{C.5}$$

Table C.1 presents values of α-uncertainty for different values of n and α.

Table C.1 – *α-uncertainty values.*

α	n								
	25	**50**	**75**	**86**	**100**	**150**	**200**	**300**	**400**
0.95	0.500	0.444	0.409	0.373	0.382	0.403	0.362	0.343	0.291
0.90	0.345	0.336	0.322	0.295	0.309	0.284	0.218	0.193	0.147
0.85	0.345	0.240	0.244	0.225	0.184	0.144	0.115	0.074	0.049
0.80	0.345	0.240	0.178	0.166	0.136	0.082	0.052	0.028	0.012
0.75	0.212	0.161	0.124	0.080	0.067	0.043	0.020	0.006	0.002
0.70	0.212	0.101	0.053	0.053	0.044	0.014	0.007	0.001	0.000
0.65	0.115	0.059	0.032	0.020	0.018	0.006	0.001	0.000	-
0.60	0.115	0.032	0.018	0.011	0.006	0.001	0.000	-	-
0.55	0.054	0.016	0.005	0.003	0.002	0.000	-	-	-
0.50	0.054	0.008	0.003	0.001	0.000	-	-	-	-
0.45	0.022	0.003	0.001	0.000	-	-	-	-	-
0.40	0.022	0.001	0.000	-	-	-	-	-	-
0.35	0.007	0.000	-	-	-	-	-	-	-
0.30	0.002	-	-	-	-	-	-	-	-
0.25	0.002	-	-	-	-	-	-	-	-

Index

Bibliography

[1] Jose Agulló, Christophe Croux, and Stefan Van Aelst. The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99(3):311–338, 2008. Cited in page(s) [55]

[2] Thorsten Bernholt, Robin Nunkesser, and Karen Schettlinger. Computing the least quartile difference estimator in the plane. *Computational Statistics & Data Analysis*, 52(2):763 – 772, 2007. Cited in page(s) [128]

[3] J. Bernsen. Dynamic thresholding of grey-level images. In *Proceedingsof the Eighth International Conference on Pattern Recognition (ICPR)*, pages 1251–1255, Paris, France, October 1986. Cited in page(s) [6]

[4] José R. Berrendero. On the global robustness of generalized s-estimators. *Journal of Statistical Planning and Inference*, 102:287–302, 2002. Cited in page(s) [128]

[5] Ilya Blayvas, Alfred Bruckstein, and Ron Kimmel. Efficient computation of adaptive threshold surfaces for image binarization. *Pattern Recognition*, 39:89–101, 2006. Cited in page(s) [6]

[6] John Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence*, 8(1):679–698, 1986. Cited in page(s) [6, 81]

[7] Yves Caron, Pascal Makris, and Nicole Vincent. Use of power law models in detecting region of interest. *Pattern Recognition*, 40(1):2521–2529, January 2007. Cited in page(s) [4, 63, 136]

[8] Sebastien Chabrier, Bruno Emile, Christophe Rosenberger, and Helene Laurent. Unsupervised performance evaluation of image segmentation. *Journal on Applied Signal Processing*, 2006:1–12, 2006. Cited in page(s) [86]

[9] Raymond H. Chan, Chung-Wa Ho, and Mila Nikolova. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on Image Processing*, 14(10):1479–1485, October 2005. Cited in page(s) [61]

[10] Arijit Chaudhuri and Horst Stenger. *Survey Sampling Theory and Methods*. M. Dekker, New York, 2nd edition, 1992. Cited in page(s) [35]

[11] Qiang Chen, Quan-sen Sun, Pheng Ann Heng, and De-shen Xia. A double-threshold image binarization method based on edge detector. *Pattern Recognition*, 41(4):1254–1267, 2008. Cited in page(s) [4, 6]

[12] Sungzoon Cho, Robert Haralick, and Seungku Yi. Improvement of Kittler and Illingworth's minimum error thresholding. *Pattern Recognition*, 22(5):609 – 617, 1989. Cited in page(s) [25]

[13] Chien-Hsing Chou, Wen-Hsiung Lin, and Fu Chang. A binarization method with learning-built rules for document images produced by cameras. *Pattern Recognition*, 43(4):1518–1530, April 2010. Cited in page(s) [4, 21, 23]

[14] C.K. Chow and T. Kaneko. Boundary detection and volume determination of the left ventricle from a cineangiogram. *Computers in Biology and Medicine*, 3(1):13 – 16, IN1–IN2, 17–26, 1973. Cardiology and Blood. Cited in page(s) [31, 43, 145]

[15] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *arXiv*, 0:0, February 2009. Cited in page(s) [137]

[16] Corinna Cortes, Mehryar Mohri, , Michael Riley, and Afshin Rostamizadeh. *Algorithmic Learning Theory*, volume 5254/2010, chapter Sample selection bias correction theory, pages 38–53. Springer Berlin / Heidelberg, 2010. Cited in page(s) [35]

[17] Christophe Croux, Peter J. Rousseeuw, and Ola Hossjer. Generalized s-estimators. *Journal of the American Statistical Association*, 89(428):1271–1281, December 1994. Cited in page(s) [128]

[18] Dereck J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510 – 515., July 1965. Cited in page(s) [139]

[19] W. Doyle. Operations useful for similarity-invariant pattern recognition. *Journal of the ACM*, 9(2):259–267, 1962. Cited in page(s) [51]

[20] Herbert Edelsbrunner and Diane L. Souvaine. Computing least median of squares regression lines and guided topological sweep. *Journal of the American Statistical Association*, 85(409):115–119, 1990. Cited in page(s) [127]

[21] Francis Ysidro Edgeworth. On observations relating to several quantities. *Philosophical Magazine Series 5*, 24(147):222–223, 1887. Cited in page(s) [126]

[22] Francis Ysidro Edgeworth. On a new method of reducing observations relating to several quantities. *Philosophical Magazine Series 5*, 25(154):184–191, 1888. Cited in page(s) [126]

[23] David A. Forsyth and Jean Ponce. *Computer Vision a Modern Approach*. Prentice Hall, August 2002. Cited in page(s) [61]

[24] Basilios Gatos, K. Ntirogiannis, and Ioannis Pratikakis. ICDAR 2009 document image binarization contest (DIBCO 2009). In *Tenth International Conference on Document Analysis and Recognition*, pages 1375–1382, October 2009. Cited in page(s) [73]

[25] Jan-Mark Geusebroek and Arnold W. M. Smeulders. Fragmentation in the vision of scenes. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, volume 1, pages 130 –135, 2003. Cited in page(s) [136]

[26] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, August 2007. Cited in page(s) [10, 12, 14, 22]

[27] Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia. OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition*, 40:389 – 397, 2007. Cited in page(s) [21]

[28] Frank R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971. Cited in page(s) [125]

[29] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schlkopf. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 601–608. MIT Press, Cambridge, MA, 2007. Cited in page(s) [35]

[30] Peter J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821, 1973. Cited in page(s) [126]

[31] Geografa e Informtica Instituto Nacional de Estadstica, editor. *Indicadores Sociodemogrficos de Mxico (1930-2000)*. INEGI, 2001. Cited in page(s) [136]

[32] G. Johannsen and J. Bille. A threshold selection method using information measures. In *Proceedings of the Sixth International Conference on Pattern Recognition*, pages 140 – 143, 1982. Cited in page(s) [24]

[33] Markus Junker, Andreas Dengel, and Rainer Hoch. On the evaluation of document analysis components by recall, precision, and accuracy. In *Proceedings of the Fifth International Conference (ICDAR '99)*, volume 0, page 713, Los Alamitos, CA, USA, 1999. IEEE Computer Society. Cited in page(s) [99]

[34] Mohamed Kamel and Aiguo Zhao. Extraction of binary character/graphics images from grayscale document images. *CVGIP: Graphical Models and Image Processing*, 55(3):203 – 217, 1993. Cited in page(s) [6]

[35] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics and Image Processing*, 29:273–285, 1985. Cited in page(s) [25]

[36] Ergina Kavallieratou. A binarization algorithm specialized on document images and photos. In *ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 463–467, Washington, DC, USA, 2005. IEEE Computer Society. Cited in page(s) [5, 27]

[37] Ergina Kavallieratou and Stamatatos Stathis. Adaptive binarization of historical document images. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 742–745, Washington, DC, USA, 2006. IEEE Computer Society. Cited in page(s) [5, 27]

[38] Steven M. Kay. *Fundamentals of statistical signal processing: Estimation theory*. Prentice-Hall Signal Processing Series, 1993. Cited in page(s) [124]

[39] Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, Wiley Classics Library 1995 edition, 1965. Cited in page(s) [35]

[40] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, July 1985. Cited in page(s) [5, 24, 31, 43]

[41] Hanako Kohmura and Toru Wakahara. Determining optimal filters for binarization of degraded characters in color using genetic algorithms. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 661–664, Washington, DC, USA, 2006. IEEE Computer Society. Cited in page(s) [3]

[42] Martin D. Levine and Ahmed M. Nazif. Dynamic measurement of computer generated image segmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):155–164, 1985. Cited in page(s) [30, 88]

[43] Lei M. Li. An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints. *Computational Statistics & Data Analysis*, 48(4):717 – 734, 2005. Cited in page(s) [128]

[44] Yun Li, Ching Y. Suen, and Mohamed Cheriet. A threshlod selection method based on multiscale and graylevel co-occurrence matrix analysis. In *ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 575–579, Washington, DC, USA, 2005. IEEE Computer Society. Cited in page(s) [6]

[45] Ping-Sung Liao, Tse-Sheng Chen, and Pau-Choo Chung. A fast algorithm for multilevel thresholding. *Journal of Information Science and Engineering*, 17:713–727, 2001. Cited in page(s) [23]

[46] S. J. Lu and C. L. Tan. Binarization of badly illuminated document images through shading estimation and compensation. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 312–316, Washington, DC, USA, 2007. IEEE Computer Society. Cited in page(s) [6]

[47] Raman Maini and J.S. Sohal. Performance evaluation of Prewitt edge detector for noisy images. *International Journal on Graphics, Vision and Image Processing*, 6(3):39–46, December 2006. Cited in page(s) [6, 81]

[48] Stéphane Marchand-Maillet and Yazid M. Sharaiha. *Binary Digital Image Processing. A Discrete Approach*. Academic Press, San Diego, 2000. Cited in page(s) [14, 57]

[49] The MathWorks. *MatLab*. 2009. Cited in page(s) [81]

[50] Carlos Mello, Ángel Sanchez, Adriano Oliveira, and Alberto Lopes. An efficient gray-level thresholding algorithm for historic document images. *Journal of Cultural Heritage*, 9(2):109–116, 2008. Cited in page(s) [3, 4, 26]

[51] Carlos A.B. Mello and Luciana A.Schuler. Thresholding images of historical documents using a Tsallis-entropy based algorithm. *Journal of Software*, 3(6):39–36, 2008. Cited in page(s) [26]

[52] Estados Unidos Mexicanos. *Archivo General de la Nación*. http: www.agn.gob.mx, 2009. Cited in page(s) [1]

[53] Robert Milewski and Venu Govindaraju. Binarization and cleanup of handwritten text from carbon copy medical form images. *Pattern Recognition*, 41(4):1308–1315, 2008. Cited in page(s) [3, 4]

[54] Reza Farrahi Moghaddam and Mohamed Cheriet. A multi-scale framework for adaptive binarization of degraded document images. *Pattern Recognition*, 43(6):2186 – 2198, June 2010. Cited in page(s) [21, 23]

[55] Subhash C. Narula and John F. Wellington. The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review*, 50(3):317–322, 1982. Cited in page(s) [126]

[56] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970. Cited in page(s) [99]

[57] M. E. J. Newman. Power laws, Pareto distributions and Zipfs law. *Contemporary Physics*, 46(46):323–351, September-October 2005. Cited in page(s) [53, 136]

[58] M. E. J. Newman. Power laws, Pareto distributions and Zipfs law. *arXiv*, 0(0):0, May 2006. Cited in page(s) [137]

[59] Hui-Fuang Ng. Automatic thresholding for defect detection. *Pattern Recognition Letters*, 27:1644–1649, 2006. Cited in page(s) [23]

[60] W.S. Ng and C.K. Lee. Comment on using the uniformity measure for performance measure in image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:933 – 934, 1996. Cited in page(s) [89]

[61] Wayne Niblack. *An Introduction to Digital Image Processing*. Prentice Hall, Birkeroed, Denmark, Denmark, 1985. Cited in page(s) [5, 14, 26]

[62] Alfonso Nieto-Castanona, Satrajit S. Ghosh, Jason A. Tourvillea, and Frank H. Guenthera. Region of interestnext term based analysis of functional imaging data. *NeuroImage*, 19:1303–1316, 2003. Cited in page(s) [63]

[63] United States of America. *Library of Congress*. , 2009. Cited in page(s) [1]

[64] Arab Republic of Egypt. *The National Archives of Egypt*. , 2009. Cited in page(s) [1]

[65] Il-Seok Oh. Document image binarization preserving stroke connectivity. *Pattern Recognition Letters*, 16:743–748, 1995. Cited in page(s) [6]

[66] N. Otsu. A threshold selection method from grey-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979. Cited in page(s) [5, 22, 89]

[67] M. Portes de Albuquerque, I.A. Esquef, A.R. Gesualdi Mello, and M. Portes de Albuquerque. Image thresholding using Tsallis entropy. *Pattern Recognition Letters*, 25:1059–1065, 2004. Cited in page(s) [5, 25]

[68] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C: The art of scientific computing*. the press syndicate of the University of Cambridge, second edition edition, 1992. Cited in page(s) [127]

[69] Marte Ramírez-Ortegón, Ernesto Tapia, Marco Block, and Ral Rojas. Quantile linear algorithm for robust binarization of digitalized letters. In *Ninth International Conference on Document Analysis and Recognition*, volume 2, pages 1158 –1162, September 2007. Cited in page(s) [29, 51, 65, 66, 67]

[70] Marte A. Ramírez-Ortegón, Edgar A. Duéñez-Guzmán, Raúl Rojas, and Erik Cuevas. Unsupervised measures for parameter selection of binarization algorithms. *Pattern Recognition*, 44(3):491 – 502, March 2011. Cited in page(s) [7, 90, 101, 159]

[71] Marte A. Ramírez-Ortegón and Raúl Rojas. Transition thresholds for binarization of historical documents. In *20th International Conference on Pattern Recognition*, pages 2362 – 2365. IEEE Computer Society, August 2010. Cited in page(s) [52, 57, 83, 146]

[72] Marte A. Ramírez-Ortegón, Ernesto Tapia, Lilia L. Ramírez-Ramírez, Raúl Rojas, and Erik Cuevas. Transition pixel: A concept for binarization based on edge detection and gray-intensity histograms. *Pattern Recognition*, 43:1233 – 1243, 2010. Cited in page(s) [7, 22, 26, 29, 31, 34, 39, 51, 57, 64, 65, 67, 75, 78, 86, 89, 110, 146, 153]

[73] Marte Alejandro Ramírez-Ortegón, Ernesto Tapia, Raúl Rojas, and Erik Cuevas. Transition thresholds and transition operators for binarization and edge detection. *Pattern Recognition*, 43(10):3243–3254, October 2010. Cited in page(s) [7, 31, 57, 60, 64, 65, 68, 80, 86, 113, 146]

[74] Xiaofeng Ren, Charless C. Fowlkes, and Jitendra Malik. Learning probabilistic models for contour completion in natural images. *International Journal of Computer Vision*, 77(1):47 – 63, May 2008. Cited in page(s) [136]

[75] E. Roelant, S. Van Aelst, and C. Croux. Multivariate generalized s-estimators. *Journal of Multivariate Analysis*, 100(5):876–887, 2009. Cited in page(s) [55, 129]

[76] Paul L. Rosin. Unimodal thresholding. *Pattern Recognition*, 34(11):2083–2096, 2001. Cited in page(s) [52]

[77] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984. Cited in page(s) [127]

[78] Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, December 1993. Cited in page(s) [128]

[79] P. K. Sahoo, S. Soltani, A. K.C. Wong, and Y. C. Chen. A survey of thresholding techniques. *Computer Vision, Graphics. and Image Processing*, 41(2):233–260, 1988. Cited in page(s) [4, 30, 86, 88]

[80] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000. Cited in page(s) [4, 5, 27]

[81] Jean Serra. *Image Analysis and Mathematical Morphology*. Academic Press, Inc., Orlando, FL, USA, 1983. Cited in page(s) [14]

[82] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, January 2004. Cited in page(s) [5, 30, 33, 86, 88, 101, 113]

[83] Andrew F. Siegel. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982. Cited in page(s) [127]

[84] Pavlos Stathis, Ergina Kavallieratou, and Nikos Papamarkos. An evaluation technique for binarization algorithms. *Journal of Universal Computer Science*, 14(18):3011–3030, October 2008. Cited in page(s) [30, 33, 101, 113]

[85] Stephen M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474, 1981. Cited in page(s) [126]

[86] Arnold J. Stromberg, Ola Hossjer, and Douglas M. Hawkins. The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association*, 95(451):853–864, September 2000. Cited in page(s) [129]

[87] Øivind Due Trier and Anil K. Jain. Goal-directed evaluation of binarization methods. *Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1191–1201, 1995. Cited in page(s) [4, 22, 26, 30, 33, 101, 113]

[88] Constantino Tsallis. *Nonextensive Statistical Mechanics and Thermodynamics*. Group of Statistical Physics, 2009. Cited in page(s) [5, 25]

[89] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, pages 1–25, 2001. Cited in page(s) [153]

[90] Christian Wolf and Jean-Michel Jolion. Extraction and recognition of artificial text in multimedia documents. *Pattern Analysis and Applications.*, 3:309–326, 2003. Cited in page(s) [5, 27]

[91] Hui Zhang, Jason E. Fritts, and Sally A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110:260 –280, September 2008. Cited in page(s) [85, 86]

[92] Yu Jin Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996. Cited in page(s) [85]


colorchecker CLASSIC
x-rite
mm


Freie Universität Berlin