Transition Thresholds for Binarization of Historical Documents

Marte A. Ramírez-Ortegón* and Raúl Rojas
Institut für Informatik, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany
* mars.sasha@gmail.com, rojas@inf.fu-berlin.de

Abstract—This paper extends the transition method for binarization based on transition pixels, a generalization of edge pixels. This method originally computes transition thresholds using the quantile thresholding algorithm, that has a critical parameter. We achieved an automatic version of the transition method by computing the transition thresholds with the Rosin’s algorithm. We experimentally tested four variants of the transition method combining the density and cumulative distribution functions of transition values, with gray-intensity thresholds based on the normal and lognormal density functions. The results of our experiments show that these unsupervised methods yields superior binarization compared with top-ranked algorithms.

Index Terms—binarization; transition; method; documents, historical

I. INTRODUCTION

Binarization classifies each pixel in an image either as foreground or background. The foreground is a subset of pixels $\mathcal{F}$ containing objects used for further analysis and recognition, while the background $\mathcal{B}$ is the complement of $\mathcal{F}$.

The transition method, proposed by Ramírez-Ortegón et al. [1], is a composite binarization algorithm based on the transition pixel concept, a generalization of edge pixels.

Empirical evidence lead us to propose the Rosin’s threshold [2] in order to avoid the parameter $\alpha$ that was originally used in the transition method. In this manner, we upgrade the transition method to the category of unsupervised method.

The rest of this paper is organized as follows. Section II introduces the related concepts. The transition method is developed in section III and III-A. In section IV, the experiments results are shown. Conclusions are presented in section V.

II. PRELIMINARY CONCEPTS

This paper only considers one input image. An image function $F$ can be defined as the mapping $F : \mathcal{P} \rightarrow \mathcal{A}$ where $\mathcal{P} = \{p \mid p \in \mathbb{N} \times \mathbb{N}\}$ and $\mathcal{A} \subset \mathbb{Z}$. $I : \mathcal{P} \rightarrow \mathcal{G} = \{0, 1, \ldots, l\}$ is the gray image function. $B : \mathcal{P} \rightarrow \{0, 1\}$ represents the binarization of $I$, where one is considered as foreground. Local binarization algorithms compute a threshold surface $T : \mathcal{P} \rightarrow \mathcal{G}$ over the whole image: $B(p) = 1$ if $I(p)$ is lower than the threshold $T(p)$. The information to compute $T(p)$ is gathered from the pixels within a square $\mathcal{P}_t(p)$ centered at the pixel $p$ of sides with length $2t + 1$.

Several kind of histograms $H_{F,A}$ will be used to keep track of the pixels in relation to their transition values or gray intensities. We have to specify the function $F$ that is used on the pixel set $A$. Then, $H_{F,A}(x)$ is the cardinality of $\{p \in A \mid F(p) = x\}$.

A pixel $p$ is a t-transition pixel if the neighborhood $\mathcal{P}_t(p)$ contains foreground and background pixels. The set of these pixels is named $\mathcal{P}_t$. If $t = 1$, then the t-transition pixel is an edge pixel.

Transition pixels can be detected by selecting those pixels that have extreme transition values. These values can be computed with max-min function [1], [3]:

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

(1)

Fig. 1. $\mathcal{N}_t(p)$ has four cases considering the pixels contained and its central pixel: $\mathcal{B} \setminus \mathcal{B}_t$, $\mathcal{B}_t$, $\mathcal{F}_t$ and $\mathcal{F} \setminus \mathcal{F}_t$.

We abbreviate $\mathcal{F}_t(p) = \mathcal{F} \cap \mathcal{P}_t(p)$, $\mathcal{F}_{t,r}(p) = \mathcal{F}_t \cap \mathcal{P}_r(p)$ and $\mathcal{F}_{t,r}(p) = \mathcal{F}_t \cap \mathcal{P}_r(p)$. The sets $\mathcal{P}_{t,r}(p)$, $\mathcal{B}_t(p)$, $\mathcal{B}_{t,r}(p)$, and $\mathcal{B}_{t,r}(p)$ are defined in a similar way.

III. OVERVIEW OF THE TRANSITION METHOD

Since $\mathcal{F}_t$ and $\mathcal{B}_t$ are dual sets, we will only explain models for $\mathcal{F}_t$, leaving out details for $\mathcal{B}_t$.

Consider that Fig. 2 (a) is a neighborhood $\mathcal{P}_r(p)$; the left peak of $H_{I,P_r(p)}$ is mainly formed by foreground pixels, while the right peak is formed by background pixels.
If we knew the **class-conditional density** $P(I(q) \mid q \in \mathcal{F}_r(p))$, we could consider the **maximum likelihood estimation** or **Bayesian estimation** approach to solve the binarization problem. Unfortunately, the class-conditional densities are rarely known, nevertheless we can reasonably assume that the gray-intensity density is approximately normally distributed in small neighborhoods $\mathcal{F}_r(p)$:

$$H_{I,\mathcal{F}_r}(p)(i) \approx c_{\mathcal{F}_r(p)} \phi(i; \mu_{\mathcal{F}_r(p)}, \sigma^2_{\mathcal{F}_r(p)}),$$

where $\phi(x; \mu, \sigma)$ is the normal probability density function with mean $\mu$ and variance $\sigma^2$. $T(p)$ is quickly computed if there is an analytic intersection between (2) and the corresponding background function, see Fig. 2 (b). Thus, we can approximate $P(I(q) \mid q \in \mathcal{F}_r(p))$ by drawing a representative sample of $\mathcal{F}_r(p)$.

The positive transition set $\mathcal{F}_+$ satisfies

$$P(I(q) \mid q \in \mathcal{F}_+(p)) \approx P(I(q) \mid q \in \hat{\mathcal{F}}_{r,+}(p)).$$

Therefore, $\mathcal{F}_{r,+}(p)$ is a representative sample of $\mathcal{F}_r(p)$, see Fig. 2 (c). Although the transition sets are also unknown, our method provides $\hat{\mathcal{F}}_{r,+}(p)$ (Fig. 2(d)) which is an accurate estimate of $\mathcal{F}_r(p)$. Then, (3) change to

$$P(I(q) \mid q \in \mathcal{F}_+(p)) \approx P(I(q) \mid q \in \hat{\mathcal{F}}_{r,+}(p)).$$

We are now able to compute the gray-threshold with usual classification procedures. The complete method consists of the following steps:

1) Compute the transition values for each pixel with a transition function. We suggest the max-min function using neighborhoods of radius 2.

2) Calculate the thresholds $t_+$ and $t_-$. Take $\hat{\mathcal{F}}_+ = \{p \mid V(p) \geq t_+\}$ and $\hat{\mathcal{F}}_- = \{p \mid V(p) \leq t_-\}$.

3) Restore $\hat{\mathcal{F}}_+$ and $\hat{\mathcal{F}}_-$. 

4) Compute the threshold image $T$ and generate the binary image $B$.

5) Remove noise from $B$ by standard algorithms.

For better understanding of the transition method, the reader is referred to [1].

### A. Transition threshold

We found suitable transition thresholds by analyzing functions of transition values: **Empirical scaled density function**

$$u_i = \frac{1}{k} H_{V,T}(i), \quad \text{with } k = \max_{i \in [1,l]} H_{V,T}(i)$$

(5)

and **empirical complementary cumulative distribution function**

$$v_i = \frac{1}{t} \sum_{j=1}^{r} H_{V,T}(j), \quad \text{with } t = \sum_{j=1}^{r} H_{V,T}(j).$$

(6)

The behavior of (5) and (6) is ideal for Rosin’s threshold [2], which proposes a threshold for unimodal histograms.

Rosin’s method assumes that one of the two classes produces one dominant peak located at one of the sides of the histogram of gray intensities. 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. A straight line $L$ is drawn from the peak to the lowest non-zero value of the histogram, the threshold point is selected as the histogram index $i$ that maximizes the perpendicular distance between $L$ and the point $(i, H_{F,A}(i))$ (See Fig. 3).
C. Statistical thresholds

Only pixels in \( \hat{P}_{i,r}(p) = \hat{F}_{i,r}(p) \cup \hat{B}_{i,r}(p) \) are considered to compute \( T(p) \). At the same time, outliers are discarded by labeling as background those pixels \( p \) that satisfy either \( |\hat{F}_{i,r}(p)| < n_+ \) or \( |\hat{B}_{i,r}(p)| < n_- \), where \( n_+ \) and \( n_- \) depend on \( r \) and objects of interest; the higher \( n_+ \), the larger the objects that can be removed from the foreground. We suggest \( n_+ = n_- = 5 \) for detecting small foreground objects.

A second criterion to discard outliers uses the difference between the gray-intensity means of the transition set. The pixel \( p \) is labeled as background if

\[
\mu_I,\hat{B}_{i,r}(p) - \mu_I,\hat{F}_{i,r}(p) < c, \tag{12}
\]

where \( c \) is an integer, which depicts the minimum contrast expected between the foreground and background. We suggest \( c = 15 \).

Normal threshold

Given \( H_{I,\hat{F}_{i,r}(p)}(i) \approx c_+ \phi(i; \mu_+, \sigma_+^2) \) where \( \phi(x; \mu, \sigma) \) is the normal probability density function with mean \( \mu \) and variance \( \sigma^2 \). Then, \( H_{I,\hat{F}_{i,r}(p)} \) is approximated when

\[
c_+ = \|\hat{F}_{i,r}(p)\| \quad \text{(complete)} \quad \text{or} \quad c_+ = 1 \quad \text{(simple)}
\]

\[
\mu_+ = \mu_I,\hat{F}_{i,r}(p), \quad \sigma_+^2 = \max \left( \sigma_I^2, \sigma_{\hat{F}_{i,r}(p)}^2 \right).
\]

The intersection of those curves is the root \( \mu_+ < x_0 < \mu_- \) of the quadratic equation with coefficients \( a, b, c \) given by

\[
a = \frac{1}{\sigma_+^2} - \frac{1}{\sigma_-^2}, \quad b = \frac{2\mu_+ - 2\mu_-}{\sigma_+^2}, \quad c = \frac{\mu_+^2}{\sigma_+^2} - \frac{\mu_-^2}{\sigma_-^2} - 2 \ln \left( \frac{\sigma_+ - c}{\sigma_+ - c} \right).
\]

Lognormal threshold

Given \( H_{I,\hat{F}_{i,r}(p)}(i) \approx c_+ \lambda(i; \mu_+, \sigma_+^2) \) where \( \lambda(\mu, \sigma^2) \) denotes the lognormal probability density function.

The intersection of these curves is \( \exp(x_0) \), where \( x_0 \) is the root of the quadratic equation with coefficients given by (14), but \( \mu_+ \) and \( \sigma_+^2 \) are estimated based on the estimated mean and variance of the lognormal distribution using the relations:

\[
\mu_+ = \ln \left( \frac{\mu_I,\hat{F}_{i,r}(p)}{2\sigma_+^2} \right) - \frac{1}{2} \sigma_+^2 \quad \text{and}
\]

\[
\sigma_+^2 = \ln \left( 1 + \frac{\sigma_+^2}{\mu_I,\hat{F}_{i,r}(p)^2} \right).
\]

IV. DESCRIPTION OF EXPERIMENTS

We compare Otsu’s, Kittler’s and Sauvola’s methods (top-ranked algorithms on [4], [5], [6]) with four variants of the transition method: T-DF-L, T-CCD-L, T-DF-N and T-CCD-N.

We implemented both Kittler’s and Otsu’s methods in their local versions [1] to increase their accuracy. Real applications rarely use more than one parameter’s set; this is the main reason we fixed Sauvola’s \( \alpha = 0.5 \) and \( \beta = 128 \), which are
the recommended parameters [7]. We set the neighborhood radius to $r = 50$.

Transition methods, denoted by the prefix T, are composite algorithms with the following operations: A) Max-min function using $N_2$. B) Rosin’s threshold for transition values using as input: (5) denoted by DF and (6) denoted by CCD. C) Three isolation operators, in order: Cross, diagonal, and rectangular isolate transition-operator ($x = y = 2$). D) Gray-intensity thresholds: Lognormal threshold denoted by L and normal threshold denoted by N. Setting $n_+ = n_- = 25$ and $c = 15$.

The binarized images were post-processed using the following operators, in order: Cross, diagonal and rectangular ($x = y = 2$) isolate operator which are defined as transition operators but over binary images.

Historical documents often are degraded with ink stains and weak ink strokes for mention some kind of degradation. Hence, we tested the binarization algorithms with the historical atlas “Theatrum orbis terrarum, sive, Atlas novus” (Blaeu Atlas) [8]. This paper reports the results of 86 text-images, see Fig. 4, extracted from 61 maps.

![Image](image_url)

**Fig. 4.** Sample from the map “Theatrum orbis terrarum, sive, Atlas novus”. On the top-right, T-CCD-N; On the bottom-left, T-DF-L; on the bottom-right, Otsu’s method.

We measure the segmentation quality by evaluating the performance of an optical character recognition (OCR) software. We used TopOCR [9] to recognize the text from the binarized images. Our evaluation measures are accuracy (AC) and precision (PR) computed as

$$AC = \frac{\#T_{\text{match}}}{\#T_{\text{in}}}, \quad \text{and} \quad PR = \frac{\#T_{\text{match}}}{\#T_{\text{out}}},$$  

where $S$ denotes the length of the string $S$, $T_{\text{in}}$ is the original text in the image, $T_{\text{out}}$ is the output text from the OCR, and $T_{\text{match}}$ is the maximum matching text between $T_{\text{in}}$ and $T_{\text{out}}$ computed by Needleman-Wuntsh algorithm [10]. AC measure is an important measure for OCR applications, because a high AC value increases the possibility to extract, by further algorithms, relevant information.

TopOCR was tested with four parameter sets. The program tester reports the AC and PR measures from the parameter set that scores higher in terms of AC measure. If there is any draw, PR measure is used.

**V. RESULTS AND CONCLUSIONS**

We based our observations in the values $p_{yx}$ and $p'_{yx}$ from Table I. We ascertain that algorithm $x$ is better than algorithm $y$ if $n_{yx} \geq 1.33n_{xy}$.

T-DF-L and T-CCD-L based on lognormal threshold are ranked first and second among the rest of the methods, respectively. T-DF-N and T-CCD-N performed similarly to Otsu’s threshold, which is the highest scored among non-transition algorithms, see Fig. 4. Our results also suggest that transition methods based on density function of transition values perform better than those based on complementary cumulative distribution of transition values.

**ACKNOWLEDGMENT**

This research was partially supported by The National Council on Science and Technology (CONACYT) of Mexico (Grant number:218253).

**REFERENCES**


http://www.library.ucla.edu/ylrl/reference/maps/blaeu

http://www.topocr.com/

the search for similarities in the amino acid sequence of two proteins,” 