Binarization of Historical Watermarks

A Review of Thresholding Techniques Applied to Historical Watermark Images

Anna N. Lantink[1]

Supervisor(s): Dr. Martin Skrodzki [1], **Dr. Jorge Martinez Castaneda**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Anna Lantink
Final project course: CSE3000 Research Project
Thesis committee: Dr. Martin Skrodzki, Dr. Jorge Martinez Castaneda, Dr. Christoph Lofi

Abstract

A watermark image is a scan of a historical paper document that contains a watermark, which is a motif embedded in the paper that provides valuable information on the origins of a document. Developing tools to automatically identify watermarks can make this information more accessible to researchers. This paper focuses on one specific binarization technique, thresholding. Thresholding selects a threshold value, which is used to turn an image binary such that one color represents foreground and the other represents background. Ideally, binarization isolates the watermark's shape by representing it as foreground, and removes unwanted information. This research compares the effectiveness of different thresholding techniques when applied to watermark images. Eight algorithms are selected from the literature, and a novel algorithm is proposed that seeks to improve on the other algorithms when applied to watermarks. The nine total algorithms are evaluated quantitatively on synthetic data, and qualitatively through a survey where participants select which algorithm appears best and rate it. The results show that there is no clear algorithm which works best for all images, however a logical adaptive approach may work marginally better than other approaches. Additionally, the presented algorithms do not adequately remove non-watermark information from the images. Further research should be conducted to analyze different binarization techniques in this context.

1 Introduction

Historical documents present value to historians and researchers, beyond just their text. Watermarks are motifs that appear in historical paper documents, which identify the manufacturers of the paper [1]. To capture a watermark, the raw watermarked paper can be scanned against a light source to make the watermark visible (Fig. 1, left), or it can be manually traced on a separate paper (Fig. 1, right). In ideal situations, researchers can use watermarks to date and locate historical documents, providing valuable context to their research [2]. Currently there are no widely accessible automated options to analyze watermarks based on appearance. This limits the accessibility of the watermarks' information, since domain knowledge from experts is required to access the desired information. An automatic watermark analysis tool focused on watermark syntax, rather than semantics, has the potential to speed this process up considerably, increasing the accessibility of watermark information.

The poor quality of raw watermark images poses a significant challenge in accurately identifying watermarks automatically. Low contrast between the watermark foreground and the non-watermark background, as well as the presence of document artifacts such as paper staining, makes it particularly difficult to identify where a watermark may be. However, raw watermark images are easier to collect than tracings, since they can be captured directly without intermediary steps. Thus, creating techniques to effectively identify watermarks in these images is a crucial step in automatically analyzing watermarks.

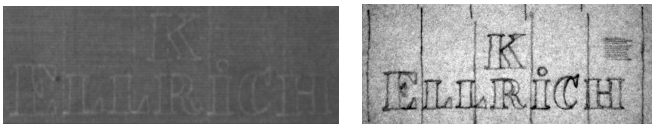


Figure 1: Examples of a raw, untraced, watermark image on the left, and a manually traced watermark image on the right.

Harmonization is a key step in the process of automatically identifying watermarks. Harmonization seeks to enhance the shape of the watermark to make analyzing its components easier. This paper focuses on one aspect of harmonization: binarization, which isolates foreground by making an image binary. Binarization techniques are useful for watermark images because they make the shape of the watermark more distinct. Yet, they are rarely researched in this context, which presents a research gap. To address this gap, the question this research seeks to answer is,

> 'To what extent can thresholding techniques be effective in binarizing watermark images with degraded quality, and how do different algorithms compare to each other?'

To answer this, watermark images are binarized using several existing thresholding algorithms. Watermark images were provided by the German Museum of Books and Writing[1], from their extensive, private collection. In addition to the selected algorithms, a new thresholding algorithm is proposed specifically for watermark binarization. The thresholding algorithms will be evaluated, both qualitatively with a human research study and quantitatively using synthetic data.

2 Background

Binarization is the process of segmenting an image into foreground and background [3]. This results in a binary image, where one color represents the foreground and the other represents the background. For watermark images, the watermark would be considered foreground, and everything else, background. Thresholding is a common form of binarization where an intensity value is determined that categorizes whether a pixel is foreground or background [3]. Thresholding techniques typically fall into two broad groups: global, where one value categorizes all pixels, and local, where different threshold values are used across different regions of the image [4]. Binarization is fundamental to watermark identification, since, ideally, it isolates the shape of the watermark and removes all other paper artifacts.

Binarization algorithms are occasionally designed for specific types of images. For this research, two types of images are particularly relevant: degraded document images, and historical document images. Degraded document images refer to images of physical documents that are degraded in some way.

[1] https://www.dnb.de/EN/Ueber-uns/DBSM/dbsm_node.html#sprg315370

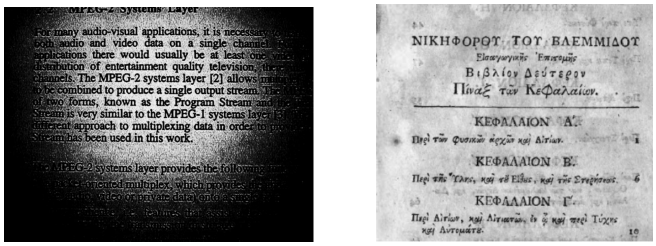


Figure 2: Examples of degraded documents analyzed in related binarization research. On the left, a document image with degraded lighting [5] and, on the right, a historical document image [6].

Degradation may refer to aspects of the image such as lighting (Fig. 2, left) or it may refer to degradation of the document itself, like staining (Fig. 2, right). Historical document images (Fig. 2, right) are a subclass of degraded document images, since their advanced age naturally leads to degradation.

3 Related Work

This section highlights previous work related to the research in this paper. Section 3.1 discusses techniques that have been developed to isolate and match historical watermarks. Section 3.2 explores both specialized and generic binarization algorithms.

3.1 Watermark Isolation Techniques

Much research has been conducted on identifying historical watermarks. Identifying specific watermarks has been performed using decision trees, for example with watermarks found in Rembrandt's etchings [7]. These algorithms are limited in that they cannot be applied to watermarks in general, since they rely on specific icons and shapes found in the watermarks in the etchings. A generalized tool for identifying and matching historical watermarks automatically has been prototyped in an earlier iteration of this research [8]. The tool performed less effectively on raw watermark images than on traced versions of these images because the raw watermark image are low contrast and degraded. The results of the prototype show room for improvement in accurately identifying watermarks in historical paper.

Al Faleh Al Hiary [9] has conducted research on isolating watermarks in degraded documents. His work focuses largely on how to remove noise from such documents and lacks detail regarding other steps of identification, such as binarization. Thus, research on improving watermark identification accuracy through the usage of specific binarization techniques has largely been unexplored.

3.2 Binarization Algorithms

Binarization of historical documents has been extensively researched. The International Document Image Binarization Contest (DIBCO) was created to track progress in document image binarization over the past decade [6, 10]. Algorithms that use thresholding are common amongst binarization algorithms. It is thus valuable to categorize such algorithms according to the techniques they use to identify thresholds. These categories have been identified for the purposes of this paper through an analysis of thresholding algorithms, with a focus on those specializing in degraded or historical documents. The following categories will inform which algorithms this paper analyzes when thresholding watermarks.

The first category contains algorithms that use color histograms. One such algorithm is proposed by Otsu [11], which uses the color histogram of an image to determine a probability distribution that is used to derive an estimate for the mean and variance. These statistics are used to determine the global threshold. An algorithm proposed by Niblack [12] also uses such statistics and applies them to regions of the images to find local thresholds. These two algorithms have inspired derivatives designed for special classes of images. Examples would be Sauvola and Pietikäinen [13] who developed an algorithm designed for degraded documents, and Khurshid et al. [14] who developed an algorithm for binarizing historical documents. Other approaches to using color histograms have been developed as well. An algorithm proposed by Kavallieratou [15] iteratively alters the global color histogram to reduce lighting variance and other irregularities. Rao [16] altered Kavallieratou's algorithm to work more effectively for historical documents by combining the global threshold in [15] with local adaptive fine-tuning.

The second category includes contrast-based and edge-based algorithms. Algorithms in this category utilize an image's contrast or its edges to determine foreground. Algorithms proposed by Su et al. [17, 18] use local minimum and maximum intensities to mimic the gradient within the image, which is used to find the image's contrast. Other edge-based algorithms have been developed for thresholding. One example would be Chen et al. [19], who developed an algorithm that adapts Canny edge detection to binarize an image.

The third category is background estimation algorithms. These algorithms calculate the grayscale background in the original grayscale image first, and then binarize the image using this information. Thus, the background that these algorithms estimate is not the single-color background of a binary image, but rather a complex grayscale image. Gatos et al. [20] propose an algorithm that calculates a degraded document's background through neighborhood pixel interpolation. Lu et al. [21] propose an algorithm that iteratively applies polynomial smoothing to isolate the background of the image.

The fourth category is entropic algorithms. This category contains algorithms which utilize entropy, a measure of information content [22], to find a threshold. Pun [23] proposed an algorithm that uses entropy to find light and dark tones in an image, which are used to calculate a threshold. Mello and Schuler [24], and Mello and Costa [22], have adapted the concept of entropic thresholding to work for historical document images by altering the threshold calculation process.

The fifth category is logic-based algorithms. These algorithms are based on Kamel and Zhao's algorithm [25], which seeks to binarize degraded documents by determining a pixel's threshold through a logical combination of its 'derivatives'. Pixel derivatives are calculated by comparing a pixel's intensity with local averages in the neighborhoods of the pixel's neighbors [25]. This process ensures that unwanted artifacts such as stains do not appear in the binarized image. Yang and Yan [5] modified Kamel's algorithm to work

for images with extreme lighting variation by automatically and adaptively calculating hyper-parameters. Ntirogiannis et al. [26] further improved Kamel's algorithm by adding contrast information to improve text enhancement in historical documents.

Beyond thresholding, there are several other methods of binarization that have been applied to historical and degraded documents. These include: Laplacian energy segmentation algorithms [27, 28], Markov Random Field algorithms [29], clustering algorithms [30, 31], and deep-learning algorithms [32]. Instead of binarizing images by computing thresholds, these algorithms rely on different ways of modeling an image to categorize the pixels. Non-thresholding algorithms tend to be more complex, both in terms of implementation and computation, and thus are not considered in this research.

Algorithms that binarize historical documents, such as those presented above, focus on isolating text as the foreground. However, watermarks are fundamentally different than text, particularly because they are in paper rather than on it, and thus have less contrast between the watermark foreground and the non-watermark background. The effectiveness of applying specialized binarization algorithms on watermarks remains unexplored, presenting a research gap that this paper seeks to address.

4 Methodology

The research process can be broken down into four main stages. First, the dataset of watermark images was gathered and split into training, validation and test sets. The training set was used to implement the thresholding algorithms, the validation set was used to fine-tune hyper-parameters, and the test set was used to evaluate. The dataset is discussed further in Section 4.1. Second, the thresholding algorithms were selected based on the categories defined in Section 3.2. The algorithm selection process is outlined in Section 4.2. Third, a specialized watermark thresholding algorithm is proposed, which is explained in Section 4.3. Finally, all of the thresholding algorithms are evaluated quantitatively, using synthetic data, and qualitatively, with a survey, which is discussed further in Section 4.4.

4.1 Dataset

To determine the effectiveness of binarization techniques on watermarks, the watermark images to test must first be chosen. A dataset of watermark images is kindly provided by the German Museum of Books and Writing[2]. Of the dataset, several types of watermark images were excluded from the final selection. Traced watermark images (Fig. 1, right) were excluded. These documents are not degraded and have high contrast, and thus are trivial to binarize. Watermarks with text overlap (Fig. 3) were excluded because this research focuses on separating watermarks from the surrounding degraded paper. Text removal is thus outside of the scope of this research. Additionally, the original dataset included watermark images that were already pre-processed, which were excluded to ensure uniformity in the data. Of the remaining watermark images, 206 were randomly selected (Fig. 4).

[2]https://www.dnb.de/EN/Ueber-uns/DBSM/dbsm_node.html#sprg315370

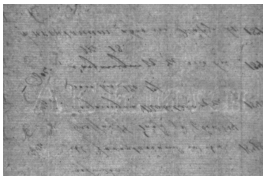


Figure 3: Example of a watermark image that is excluded from data selection because of the overlapping text present on the watermark. Text removal is outside of the scope of this research.

The selected watermark images were cropped around the watermarks to improve the accuracy of binarization results. Watermarks were cropped by the bounding box containing the watermark, padded with 20 pixels on each side. If the 20 pixel padding extended beyond the historical paper, then the padding ended at the paper's edge. If several watermarks appeared in a single raw watermark image, then each were cropped and split into different images. This process resulted in 235 total cropped watermark images. These were split into 66% training, 17% validation, and 17% testing. This split was chosen based on the validation and testing set size. These sets were intentionally kept small to make qualitative evaluation feasible, which is explained further in Section 4.4.

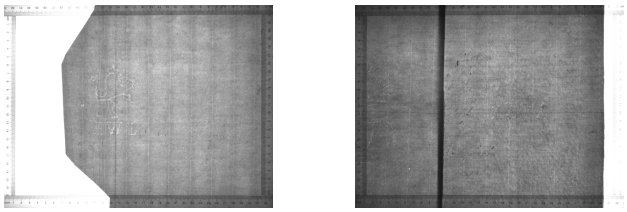


Figure 4: Examples of watermark images that were selected to be included in the watermark dataset, since they are not traced, and do not contain overlapping text.

4.2 Algorithm Selection

To compare binarization techniques, the algorithms to compare must be selected. First, five algorithms that specialize in degraded and historical documents were selected, each belonging to a category outlined in Section 3.2. This ensures that a wide range of techniques are compared. Within a category, the algorithm was chosen if it was reproducible, if it was not specialized for text, and if it did not make assumptions that exclude watermark images. For example, algorithms that assume the image is high contrast were excluded, since this assumption does not apply to the watermark images. The following specialized algorithms were selected: Rao et al. [16], which is a color histogram algorithm that uses a hybrid of global and local thresholding, Su et al. [17], which is a local contrast-based algorithm, Gatos et al. [20], which is a local background estimation algorithm, Mello and Costa [22], which is a global entropic algorithm, and Kamel and Zhao [25], which is a local logic-based algorithm.

(a) (b) (c)

Figure 5: The steps of the proposed algorithm's binarization when applied to the raw watermark image in Figure 1: (a) shows the low detail binarized watermark, (b) the high detail binarized watermark, and (c) the output of the proposed algorithm which combines (a) and (b).

In addition to these specialized algorithms, the binarization algorithm used in the previous watermark matching prototype [8] is chosen as a baseline. It first thresholds locally with Sauvola [13] and then applies a morphological closing operation.

Finally, two more baseline algorithms were selected. These were chosen to be the global algorithm proposed by Otsu [11], and the local algorithm proposed by Niblack [12]. These algorithms are generic and commonly used for thresholding [3]. Thus, they can serve as a general case to compare the specialized algorithms against. This is also why they are commonly used as a baseline when evaluating other binarization experiments [17, 18, 20, 26, 27].

All algorithms were implemented according to their descriptions[3]. Hyper-parameters were tuned using the training set to improve performance on the watermark set (Appendix A). Additionally, the algorithm proposed in [17] contained a vague explanation for single-pixel artifact filtering. This step was replaced with a morphological closing operation since it was a close parallel to the procedure alluded to by the authors.

4.3 Proposed Algorithm

This paper proposes a thresholding algorithm for historical watermark documents. First, two binarized images are generated using the watermark image (Fig. 1, left): one with low detail (Fig. 5a), and one with high detail (Fig. 5b). The low detail image should have very little misclassified foreground, but also less of the watermark present in the foreground. The high detail result should have most of the watermark present in the foreground, but also more of the non-watermark paper misclassified as foreground. Note that this algorithm assumes foreground to be black. The low and high detailed images were generated using Kamel's algorithm [25], using different hyper-parameters. This algorithm was chosen by observing which of the selected algorithms visually performed the best on the training set.

The proposed algorithm takes the low detail image and clarifies the watermark using the high detail image. This is done by iterating through each foreground pixel in the low detail image. Within a defined window in the high detailed image, centered at the pixel of the current iteration, all of the foreground pixels from the high detailed image are added to the corresponding region of the final image. The final image begins as blank, and is thus filled with foreground pixels. Since the low detail image should contain sparse but mostly correctly classified foreground pixels, windowing around foreground pixels should clarify the watermark outline without adding many non-watermark pixels to the foreground.

Algorithm 1 Proposed Algorithm

Input: The grayscale input image I
Output: The binary image $result$

$T \leftarrow -0.03$
$stroke\_width \leftarrow 5$

$initial\_binarized \leftarrow$ Kamel$(I, T, stroke_width)$
Apply morphological opening to $initial\_binarized$.

Based on the fraction of foreground pixels in $initial\_binarized$, determine T_{high} and T_{low}.

$high\_detail \leftarrow$ Kamel$(I, T_{\text{high}}, stroke_width)$
$low\_detail \leftarrow$ Kamel$(I, T_{\text{low}}, stroke_width)$
Apply morphological opening to $high\_detail$ and closing to low_detail.

Based on the fraction of foreground pixels in low_detail, determine $window\_radius$.

$result \leftarrow$ Create a white image, same shape as input
for all foreground pixels in low_detail **do**
  $window_{\text{high}} \leftarrow$ a window in $high\_detail$ centered at pixel with radius $window\_radius$.

  Add foreground pixels in $window_{\text{high}}$ to corresponding region in $result$.
end for

Apply morphological operations to $result$.

The output of the algorithm is further improved by applying morphological closing and opening operations to various stages of binarization. In addition, the fraction of foreground pixels in the intermediary binary images is used to alter hyper-parameters. For example, a lower fraction of foreground pixels should lead to a larger window radius. These adjustments are made to accommodate the diversity of watermark images being analyzed. The algorithm is summarized in Algorithm 1, and given in more detail in Appendix B. The result of applying the proposed algorithm can be seen in Figure 5c. The proposed algorithm should improve upon other algorithms because it combines information from existing bi-

[3] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Skrodzki_Castaneda/alantink-Automated-processing-of-scanned-historic-watermar

narized results to determine where the foreground is likely to be. This information is then used to reduce misclassifications.

4.4 Evaluation

To determine the effectiveness of the nine binarization algorithms, they must be evaluated. The evaluation procedure for the algorithms contains both a quantitative and a qualitative component. The quantitative component is important because it allows the algorithms to be objectively compared to each other. The qualitative component contextualizes the quantitative results by factoring in human perception, which is relevant because objective metrics may not always reflect reality in practice.

Quantitative Evaluation

A quantitative evaluation was conducted to get data on the specific performance of the binarization algorithms. The watermark images in the test set do not have ground truths. This makes it difficult to quantitatively evaluate the watermark images directly. Thus, synthetic data was created that mimicked the watermark images.

To create the synthetic data, first the ground truth for the data was generated. The ground truth was produced by randomly sampling from an existing dataset of human drawings [33]. The drawings in this dataset are simple, geometric, and resemble a watermark outline (Fig. 6, left). These drawings were binarized using Otsu [11] to produce the ground truth. The ground truth was cropped as described in Section 4.1. The synthetic watermark images were created by noising the ground truths. First, the background color was randomly chosen from an interval of 50 to 200. The watermark color was randomly selected within $+40$ to $+60$ of the background color, which ensures that there is low contrast between foreground and background. These numbers were chosen based on the watermark training set. Additionally, horizontal and vertical lines, which represent the chain and laid lines [1], are added. The color of the lines, number of the lines, and distance between the lines are all generated randomly. Finally, the image is noised and blurred randomly. An example of a synthetic image before and after noising can be seen in Figure 6. For the validation set 50 synthetic images were generated, and 100 for the test set.

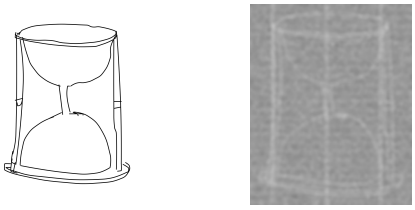


Figure 6: An example of synthetic watermark generation, used for the quantitative evaluation. Generation begins with the original drawing on the left [33], and results in the synthetic watermark image on the right.

The synthetic binarized results are generated by applying the nine selected algorithms to the raw synthetic image. The similarity between the ground truth and the binarized result was evaluated using several metrics. The metrics chosen are based on those used in DIBCO [10], and are commonly used to evaluate specialized binarization algorithms. These metrics are: the F1 score, the peak signal-to-noise ratio (PSNR), the negative rate metric (NRM), and the misclassification penalty metric (MPM) [10]. The F1 score and the NRM both use pixel-wise comparisons. Specifically, they are calculated using the distribution of true positives, false positives, true negatives, and false negatives with respect to the image pixels. For example, a false positive would be incorrectly classified foreground. The F1 and NRM combine these values in different ways. The PSNR is a weighted pixel-wise difference, placing more importance on the number of misclassificatied pixels than the F1 or NRM. The MPM, on the other hand, analyzes a pixels distance from the ground truth contour. The F1 score is best when equal to 1, the PSNR is better the higher it is, and both the NRM and MPM are best at 0.

Qualitative Evaluation

For the qualitative evaluation a human research study is conducted to determine the effectiveness of different algorithms according to human perceptions. Participants are given a survey, where they are shown a raw watermark image and must choose which of the nine binarized results isolates the watermark the best. The results are produced by applying the nine algorithms to the cropped, raw, watermark images (Fig. Figure 7), and are presented in a random order. Participants also rate how well the result they chose isolates the watermark. This rating is determined by a five-point Likert scale, ranging from strongly disagree to strongly agree. The first statement to rate is: 'The image I selected contains the complete watermark', and the second is: 'The image I selected contains only the watermark and none of the non-watermark background'. These statements are useful for determining overall effectiveness, rather than comparative effectiveness. This procedure is repeated for all 40 watermark images in the test set, which are shown to the participant in a random order.

A subset of participants in this study are individuals that work in fields that analyze historical paper. These individuals are important to include because they would be the user group for a watermark matching tool. Thus their perceptions of watermark isolation are particularly relevant. To broaden the participant pool, non-users are included in the study, which are staff and students at the Delft University of Technology. To track domain knowledge, participants are asked during the survey if they have expertise in a field related to historical paper analysis or image processing. Domain knowledge is tracked because experts in different fields related to this research may have different expectations for the algorithms' performance and may have differing familiarity with watermark images.

5 Results

The evaluation of the binarization algorithms is executed both quantitatively, discussed in Section 5.1, and quantitatively, discussed in Section 5.2.

5.1 Quantitative Results

The quantitative data shows that, across metrics, the results are fairly poor. This can be seen by comparing them to other

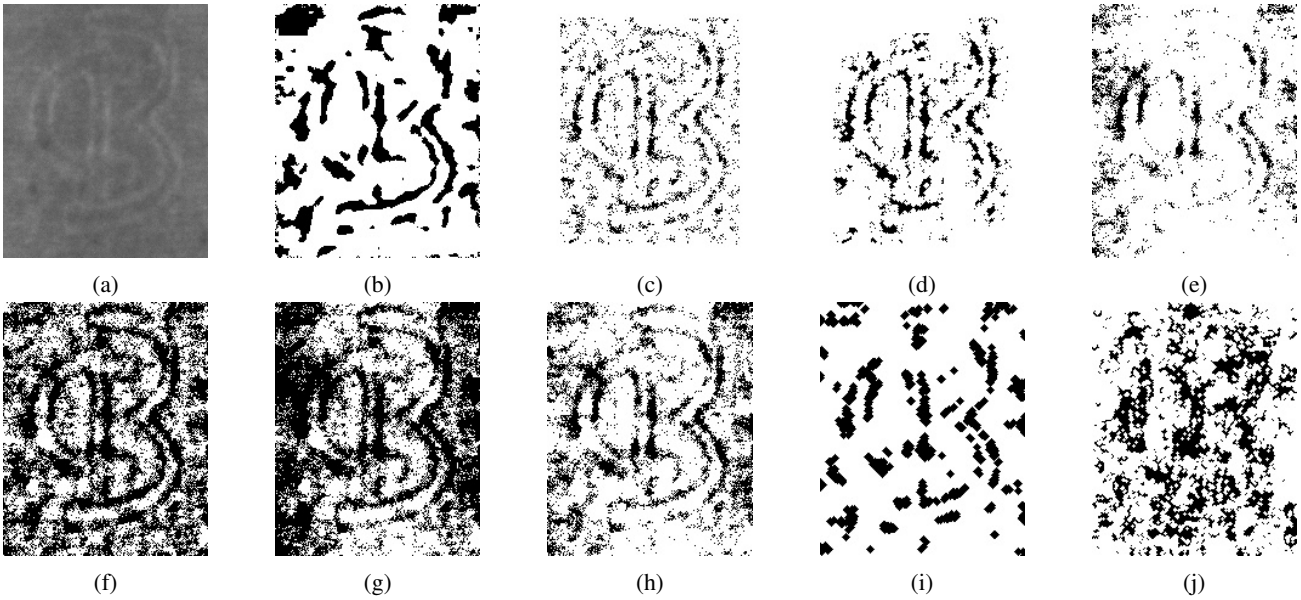


Figure 7: Binarization results of a test watermark image (a), produced using different algorithms: (b) Gatos' algorithm [20], (c) Kamel's algorithm [25], (d) the proposed algorithm, (e) Mello's algorithm [22], (f) Niblack's algorithm [12], (g) Otsu's algorithm [11], (h) Rao's algorithm [16], (i) the watermark prototype's algorithm [8], (j) Su's algorithm [17]. These images are an example of what participants are presented in the survey.

		F1 Score ($\times 10^{-2}$)		PSNR		NRM ($\times 10^{-2}$)		MPM ($\times 10^{-2}$)	
Category	Algorithm	Mean	Std. Div.	Mean	Std. Div.	Mean	Std. Div.	Mean	Std. Div.
Local	Su et al. [17]	13.95	5.81	8.38	2.48	37.23	7.27	5.46	3.62
	Kamel and Zhao [25]	20.46	8.63	**12.16**	2.65	37.88	8.41	**1.83**	2.50
	Gatos [20]	20.20	6.92	6.68	0.84	21.51	5.96	10.03	2.70
	Proposed Algorithm	24.21	8.54	**11.47**	1.97	34.38	7.35	**1.82**	1.98
	Niblack [12]	12.17	4.09	3.16	0.18	27.14	2.19	24.83	1.32
	Watermark Prototype [8]	11.49	4.11	2.84	0.65	29.20	3.54	27.38	4.29
Hybrid	Rao et al. [16]	20.50	6.87	6.40	0.68	**19.24**	4.50	9.18	2.73
Global	Mello and Costa [22]	**30.72**	8.57	9.62	0.62	**19.13**	6.93	2.47	1.18
	Otsu [11]	15.50	6.52	4.15	1.21	21.55	4.93	16.24	5.47

Table 1: Quantitative results for the nine selected algorithms. Algorithms in italics are the baseline algorithms.

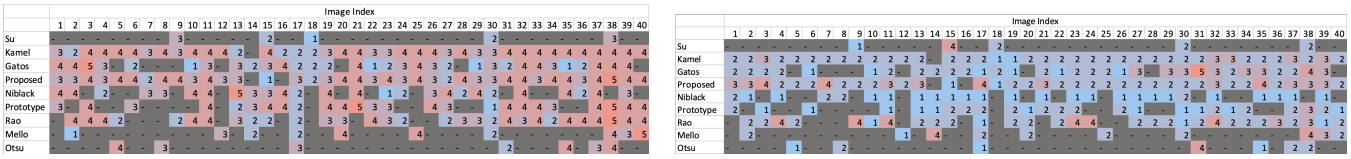
specialized binarization algorithms, like those in [10]. For example, the MPM is an order of magnitude worse than most algorithms presented in [10]. It can also be seen that there is little agreement between metrics (Table 1). Mello's algorithm performs best for F1 and NRM, Kamel's algorithm performs best for PSNR, and the proposed algorithm performs best for MPM. This also shows that metrics disagree on the performance of local and global algorithms, since Mello is a global algorithm, and Kamel is a local algorithm. Baseline algorithms perform worse than all of their specialized counterparts on all metrics, with the exception of the NRM.

When comparing the proposed algorithm to Kamel – the algorithm it seeks to improve – the proposed algorithm is better for the F1 score, the NRM, and the MPM, but worse for the PSNR (Table 1). The proposed algorithm does not significantly improve performance. This may be caused by poorly generated low and high detail images. The algorithm will not work effectively if the low detailed image does not identify any watermark pixels correctly or if the foreground pixels are too sparse and the window radius too small. Likewise, if the low detailed image contains too many misclassified foreground pixels, the result will be a poorly isolated watermark. Thus, the proposed algorithms stands to be improved.

To understand better the discrepancy in results between metrics, the metrics can be analyzed. The MPM penalizes misclassifications that are closer to the ground truth less than those further from the ground truth. As a result, images that have less defined watermark outlines and fewer overall misclassifications are favored over images with more defined watermarks and more misclassifications. The results in Figure 7 align with this expectation, since the MPM ranks Kamel's algorithm highly, and in the figure it's algorithm produces a sparse watermark outline with less misclassified foreground. The F1 score, PSNR, and NRM all use pixel-wise differences, but each prioritizes different things. For example, the PSNR penalizes the proportion of misclassifications overall,

Category	Algorithm	Percentage (%) of Overall Selection	Percentage (%) of Mode for an Image
Local	Su et al. [17]	0.90	0.00
	Kamel and Zhao [25]	**34.65**	**45.00**
	Gatos [20]	10.90	7.50
	Proposed Algorithm	**30.97**	32.50
	Niblack [12]	7.43	5.00
	Watermark Prototype [8]	5.28	0.00
Hybrid	Rao et al. [16]	6.94	0.00
Global	Mello and Costa [22]	2.08	0.00
	Otsu [11]	0.83	0.00

Table 2: The percentage of an algorithm's binarized result being chosen as the best across users and images, and the percentage that an algorithm is the mode for an image (excluding ties). Algorithms in italics are the baseline algorithms.



(a) (b)

Figure 8: Heat maps showing the median Likert ratings for each algorithm corresponding to a certain image. Likert ratings for statement 1 are depicted in (a) and statement 2 in (b). Columns are images, and rows are algorithms. Colors range from strongly Disagree in blue (represented as 1) to Strongly Agree in red (represented as 5). Gray represents no choice for that image. Note that numbers represent ordinal categories and are used for brevity, they should not be treated as numerical.

whereas the NRM tends to be more forgiving if one class, for example foreground, is heavily overclassified. This can be seen in Figure 7, where Otsu, which has a high NRM and low PSNR, misclassifies foreground frequently. On the other hand, Kamel's algorithm has a good PSNR but a poor NRM, and has a more even distribution between false positives and false negatives (Fig. 7).

As regards answering the research question, these results show that none of the algorithms are particularly effective, especially when compared to results from [10]. It is difficult to determine which algorithms perform better, since the metrics often contradict each other. Additionally, the quantitative results were generated using synthetic data, which is not the same as the original watermark images. Synthetic data does not account for tearing or staining that watermark scans often contain. Thus, considering the qualitative results is imperative to gain a better understanding of which algorithms might be best in practice.

5.2 Qualitative Results

For the qualitative evaluation, 36 participants completed the survey. 17 participants had expertise in a field related to image processing, 12 had expertise in a field related to historical paper, 7 had neither, and no participants had both. The exact user population size is unknown. However, since users would be historians and researchers, the population is safely assumed to be larger than 100 individuals. Under this assumption, the survey's sample size is too small for results to be statistically significant and applicable to a broader population [34]. Thus, the qualitative results serve only as informal observations.

According to the survey results (Table 2), Kamel's algorithm was chosen most frequently from the binarized results, followed by the proposed algorithm. However, there is no algorithm that is chosen the majority of the time. The survey also shows that all local algorithms, barring Su's algorithm, are chosen more frequently than global algorithms. Additionally, the proposed algorithm is chosen less frequently than Kamel's algorithm. This may be because, for some images, the proposed algorithms removes noise from Kamel's result at the expense of the watermark outline, as evident in Figure 7d. However, the proposed algorithm is chosen more than the seven other algorithms. These results align with the MPM and PSNR metrics. Interestingly, Mello's algorithm is chosen relatively infrequently among participants, despite being high ranking with the F1 and NRM (Table 1). This indicates that these metrics may not fully align with human perception.

Rating	Percentage (%) for Statement 1	Percentage (%) for Statement 2
Strongly Agree	12.01	1.32
Agree	**45.56**	12.08
Neutral	20.90	16.74
Disagree	18.61	**48.75**
Strongly Disagree	2.92	21.11

Table 3: The percentage of a rating being chosen across users and across images for the first statement: "The image I selected contains the complete watermark", and the second statement: "The image I selected contains only the watermark and none of the non-watermark background". Percentages are measured overall, without considering which algorithm the rating corresponds to.

Most participants agreed that the complete watermark was shown and most participants disagreed that no misclassified foreground was present (Table 3, Fig. 8). Specialized local algorithms, particularly Kamel, Gatos, and the proposed algorithm, tend to contain less misclassified foreground, and are selected more often than the other algorithms (Fig. 8b). However, there is still variation across ratings, even within a specific algorithm (Fig. 8). Participant responses show that the best binarized results isolate the watermark the majority of the time and that results still frequently contain noticeable misclassified foreground.

Gwett's AC_1 agreement coefficient is used to measure how much participants agree with each other [35]. Gwett's AC_1 applies specifically to several raters rating several categorical or ordinal items, and is therefore more applicable to this data [35] than other metrics. Gwett's AC_1 is less sensitive to the prevalence of categories than other similar metrics [35]. This is important because of the element of subjective preference inherent in survey data. The coefficient results show that, according to benchmarks used in [36], the algorithm selection question has fairly poor agreement and both Likert statements have good agreement (Table 4).

Question	AC_1	95% Confidence Interval
Best Algorithm Selection	0.261	[0.224, 0.298]
Likert Statement 1	0.669	[0.627, 0.710]
Likert Statement 2	0.677	[0.638, 0.717]

Table 4: The AC_1 coefficient and the 95% confidence interval for each survey question. Likert statement one states: "The image I selected contains the complete watermark", and statement two states: "The image I selected contains only the watermark and none of the non-watermark background". Likert statement agreement is measured overall, without considering which algorithm the rating corresponds to.

When considering how the selected algorithms compare to each other, the survey results show that Kamel is chosen most frequently as the best result (Table 2). However, due to participant disagreement, any conclusion about which algorithm is best can be tentative at best. In addition, there is substantial agreement among participants that the best binarization results shows the complete watermark, and substantial disagreement that the result removes all non-watermark background. Thus, the thresholding algorithms can be considered only somewhat effective, since the watermark is often present, but is usually not completely isolated. As stated above, these observations only relate to the participants and cannot be generalized to a broader population due to the small sample size.

6 Responsible Research

Reproducibility and transparency are key components to responsible research, since they ensure that the results of this research can be verified. To ensure that this research is as reproducible as possible, the implementation of the algorithms used to produce the results are available on a GitLab repository in the TUDelft EWI GitLab instance[4]. Transparency in the watermark data being used is more difficult, because the watermark data cannot be made publicly available. To mitigate the impact of this, various stages of binarization applied to the watermark images are seen in this paper to make the steps in the research process transparent. The research data will be published publicly to 4TU.ResearchData[5] to ensure transparency in the data analysis process. The image drawing dataset [33], is also available online, and is licensed under a Creative Commons Attribution 4.0 International License[6]. Thus, several steps have been taken to ensure that the data and algorithms used throughout this research are openly available for reproduction or further research.

To gather qualitative results a human research study was conducted. When conducting a human research study, it is necessary that all participant data is gathered ethically and consensually. To ensure this was the case, a proposal was sent to the Human Research Ethics Committee (HREC) in the Delft University of Technology. In order to ensure that all aspects of the study were thoroughly considered and evaluated, a consent form, data management plan, and HREC checklist. These forms help ensure that the study followed GDPR and that participants were consensually participating. This research therefore also seeks to be ethical by ensuring the safety of those participating in the human research study.

7 Conclusions and Future Work

This paper seeks to answer how effective thresholding algorithms are when applied to raw watermark images. To this end, nine thresholding algorithms are applied to watermark images and evaluated. The results show that no algorithm manages to completely separate the watermark from its surroundings. Qualitatively, participants tended to answer that the best binarized results contain the watermark, but misclassify non-watermark pixels as foreground. Although the qualitative results cannot be generalized to a larger population, they align with the quantitative results in demonstrating that the selected algorithms fail to completely isolate the watermark.

This paper also seeks to compare how different thresholding algorithms perform when applied to watermark images. Specialized algorithms tend to perform better than the baseline algorithms. Out of the nine thresholding algorithms studied, the logical adaptive algorithm proposed by Kamel and Zhao [25], followed by the proposed algorithm, tend to perform better than others. However, there is remarkable disagreement both among participants and metrics on this matter, so no definitive conclusion can be made. To answer the research question, the selected thresholding techniques are not effective to a significant extent when applied to degraded watermark images.

Further research can improve watermark isolation. For example, qualitative evaluations can be conducted on a larger

[4]https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Skrodzki_Castaneda/alantink-Automated-processing-of-scanned-historic-watermar

[5]https://data.4tu.nl/

[6]https://creativecommons.org/licenses/by/4.0/

sample, so that results could be generalized to the user group. Another area for future work would be to analyze the effectiveness of applying non-thresholding algorithms to watermark binarization, for example by using machine learning. Additionally, it would be valuable to combine these thresholding algorithms with denoising techniques. Ultimately, these thresholding algorithms do not effectively isolate watermarks, but there is much room for further investigation.

References

[1] L. Müller, "Understanding paper: Structures, watermarks, and a conservator's passion," https://harvardartmuseums.org/article/understanding-paper-structures-watermarks-and-a-conservator-s-passion, 2021.

[2] E. Heawood, "The use of watermarks in dating old maps and documents," *The Geographical Journal*, vol. 63, no. 5, pp. 391–410, 1924. [Online]. Available: http://www.jstor.org/stable/1781227

[3] N. Chaki, S. H. Shaikh, and K. Saeed, *Exploring Image Binarization Techniques*, ser. Studies in Computational Intelligence. New Delhi: Springer India, 2014, vol. 560. [Online]. Available: https://link.springer.com/10.1007/978-81-322-1907-1

[4] Jyotsna, S. Chauhan, E. Sharma, and A. Doegar, "Binarization techniques for degraded document images — A review," in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*. Noida, India: IEEE, Sep. 2016, pp. 163–166. [Online]. Available: http://ieeexplore.ieee.org/document/7784945/

[5] Y. Yang and H. Yan, "An adaptive logical method for binarization of degraded document images," *Pattern Recognition*, vol. 33, no. 5, pp. 787–807, May 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320399000941

[6] I. Pratikakis, K. Zagoris, X. Karagiannis, L. Tsochatzidis, T. Mondal, and I. Marthot-Santaniello, "ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019)," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. Sydney, Australia: IEEE, Sep. 2019, pp. 1547–1556. [Online]. Available: https://ieeexplore.ieee.org/document/8978205/

[7] C. R. Johnson, "Decision Trees for Watermark Identification in Rembrandt's Etchings," *Journal of Historians of Netherlandish Art*, vol. 12, no. 2, Jun. 2020. [Online]. Available: https://jhna.org/articles/decision-trees-for-watermark-identification-in-rembrandts-etchings

[8] D.-M. Banta, S. Kho, A. N. Lantink, A.-R. Marin, and V. Petkov, "A watermark recognition system: An approach to matching similar watermarks," http://resolver.tudelft.nl/uuid:e8dfbd63-ae54-4159-b786-d1d8c64dc827, 2023.

[9] H. A. A. Al Faleh Al Hiary, "Paper-based watermark extraction with image processing," Ph.D. dissertation, University of Leeds, Leeds, United Kingdom, 2008.

[10] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "DIBCO 2009: document image binarization contest," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 14, no. 1, pp. 35–44, Mar. 2011. [Online]. Available: http://link.springer.com/10.1007/s10032-010-0115-7

[11] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan. 1979. [Online]. Available: http://ieeexplore.ieee.org/document/4310076/

[12] W. Niblack, *An introduction to digital image processing*. Englewood Cliffs, N.J: Prentice-Hall International, 1986.

[13] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, Feb. 2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320399000552

[14] K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of Niblack inspired binarization methods for ancient documents," K. Berkner and L. Likforman-Sulem, Eds., San Jose, CA, Jan. 2009, p. 72470U. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.805827

[15] E. Kavallieratou, "A binarization algorithm specialized on document images and photos," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*. Seoul, South Korea: IEEE, 2005, pp. 463–467 Vol. 1. [Online]. Available: http://ieeexplore.ieee.org/document/1575589/

[16] A. V. S. Rao, G. Sunil, N. V. Rao, T. K. Prabhu, L. Reddy, and A.S.C.S.Sastry, "Adaptive binarization of ancient documents," in *2009 Second International Conference on Machine Vision*. Dubai, United Arab Emirates: IEEE, 2009, pp. 22–26.

[17] Bolan Su, Shijian Lu, and Chew Lim Tan, "Robust Document Image Binarization Technique for Degraded Document Images," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1408–1417, Apr. 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6373726/

[18] B. Su, S. Lu, and C. L. Tan, "Binarization of historical document images using the local maximum and minimum," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. Boston Massachusetts USA: ACM, Jun. 2010, pp. 159–166. [Online]. Available: https://dl.acm.org/doi/10.1145/1815330.1815351

[19] Q. Chen, Q.-s. Sun, P. Ann Heng, and D.-s. Xia, "A double-threshold image binarization method based on edge detector," *Pattern Recognition*, vol. 41, no. 4, pp. 1254–1267, Apr. 2008. [Online].

Available: https://linkinghub.elsevier.com/retrieve/pii/S003132030700413X

[20] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognition*, vol. 39, no. 3, pp. 317–327, Mar. 2006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320305003821

[21] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 13, no. 4, pp. 303–314, Dec. 2010. [Online]. Available: http://link.springer.com/10.1007/s10032-010-0130-8

[22] C. A. B. Mello and A. H. M. Costa, "Image Thresholding of Historical Documents Using Entropy and ROC Curves," in *Progress in Pattern Recognition, Image Analysis and Applications*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, A. Sanfeliu, and M. L. Cortés, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, vol. 3773, pp. 905–916, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/11578079_93

[23] T. Pun, "Entropic thresholding, a new approach," *Computer Graphics and Image Processing*, vol. 16, no. 3, pp. 210–239, Jul. 1981. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0146664X81900381

[24] C. A. Mello and L. A. Schuler, "Tsallis entropy-based thresholding algorithm for images of historical documents," in *2007 IEEE International Conference on Systems, Man and Cybernetics*. Montreal, QC, Canada: IEEE, Oct. 2007, pp. 1112–1117. [Online]. Available: http://ieeexplore.ieee.org/document/4413767/

[25] M. Kamel and A. Zhao, "Extraction of Binary Character/Graphics Images from Grayscale Document Images," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 3, pp. 203–217, May 1993. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1049965283710151

[26] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A Modified Adaptive Logical Level Binarization Technique for Historical Document Images," in *2009 10th International Conference on Document Analysis and Recognition*. Barcelona, Spain: IEEE, 2009, pp. 1171–1175. [Online]. Available: http://ieeexplore.ieee.org/document/5277601/

[27] W. Xiong, L. Zhou, L. Yue, L. Li, and S. Wang, "An enhanced binarization framework for degraded historical document images," *EURASIP Journal on Image and Video Processing*, vol. 2021, no. 1, p. 13, Dec. 2021. [Online]. Available: https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-021-00556-4

[28] N. R. Howe, "A Laplacian Energy for Document Binarization," in *2011 International Conference on Document Analysis and Recognition*. Beijing, China: IEEE, Sep. 2011, pp. 6–10. [Online]. Available: http://ieeexplore.ieee.org/document/6065266/

[29] C. Wolf and D. Doermann, "Binarization of low quality text using a Markov random field model," in *Object recognition supported by user interaction for service robots*, vol. 3. Quebec City, Que., Canada: IEEE Comput. Soc, 2002, pp. 160–163. [Online]. Available: http://ieeexplore.ieee.org/document/1047819/

[30] S. K. Bera, S. Ghosh, S. Bhowmik, R. Sarkar, and M. Nasipuri, "A non-parametric binarization method based on ensemble of clustering algorithms," *Multimedia Tools and Applications*, vol. 80, no. 5, pp. 7653–7673, Feb. 2021. [Online]. Available: http://link.springer.com/10.1007/s11042-020-09836-z

[31] S. Bhowmik, R. Sarkar, B. Das, and D. Doermann, "*GiB* : A ${G}$ ame Theory ${I}$ nspired ${B}$ inarization Technique for Degraded Document Images," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1443–1455, Mar. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8517161/

[32] M. Valizadeh and E. Kabir, "An adaptive water flow model for binarization of degraded document images," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 16, no. 2, pp. 165–176, Jun. 2013. [Online]. Available: http://link.springer.com/10.1007/s10032-012-0182-z

[33] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–10, Aug. 2012. [Online]. Available: https://dl.acm.org/doi/10.1145/2185520.2185540

[34] A. M. Adam, "Sample Size Determination in Survey Research," *Journal of Scientific Research and Reports*, pp. 90–97, Jun. 2020. [Online]. Available: https://journaljsrr.com/index.php/JSRR/article/view/1154

[35] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 29–48, May 2008. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/10.1348/000711006X126600

[36] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples," *BMC Medical Research Methodology*, vol. 13, no. 1, p. 61, Dec. 2013. [Online]. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-61

A Appendix: Hyper-parameters

The table below outlines which values were used for the hyper-parameters in each selected algorithm.

Algorithm	Parameter	Value
Su et al. [17]	Window for contrast image	7
	Gamma	0
	Morphological structure	3×3 cross
Kamel and Zhao [25]	Window size	5
	Threshold T	-0.03
Gatos [20]	Sauvola k-value	0.005
	Background estimation window radius	25×25
	q	0.6
	$p1$	0.5
	$p2$	0.8
	Shrink window radius	3×3
	k_{sh}	21
	k_{sw}	9
	k_{sw1}	21
Proposed Algorithm	Refer to Appendix B	-
Niblack [12]	Window size	41
	k	0.1
Watermark Prototype [8]	Window size	45
	k	0.01
	Morphological structure	3×3 cross
	Morphological closing iterations	3
Roa et al. [16]	Window size	31
	k	0.01
Mello and Costa [22]	Refer to [22]	-
Otsu [11]	-	-

Table 5: List of algorithms and the values of their hyper-parameters that were used during calculation.

B Appendix: Proposed Algorithm

This appendix contains a more detailed version of Algorithm 1, shown in Section 4.3.

Algorithm 2 Proposed Algorithm

Input: The grayscale input image I
Output: The binary image $result$

$T \leftarrow -0.03$
$stroke\_width \leftarrow 5$
$morph\_structure \leftarrow 3 \times 3$ cross

$initial\_binarized \leftarrow$
$initial\_binarized \leftarrow$ Kamel$(I, T, stroke_width)$
Apply morphological opening to $initial\_binarized$ using $morph\_structure$

$fraction\_binarized \leftarrow$ fraction of foreground pixels in $initial\_binarized$.
if $fraction\_binarized > 0.2$ **then**
    $T_{\text{high}} = T - 0.01$
    $T_{\text{low}} = T - 0.015$
else if $fraction\_binarized < 0.05$ **then**
    $T_{\text{high}} = T + 0.005$
    $T_{\text{low}} = T$
else
    $T_{\text{high}} = T$
    $T_{\text{low}} = T - 0.005$
end if

$high\_detail \leftarrow$
$initial\_binarized \leftarrow$ Kamel$(I, T_{\text{high}}, stroke_width)$
$low\_detail \leftarrow$ Kamel$(I, T_{\text{low}}, stroke_width)$
Apply morphological opening to $high\_detail$ and morphological closing to low_detail using $morph\_structure$.

$fraction\_low \leftarrow$ fraction of foreground pixels in low_detail.
if $fraction\_low < 0.02$ **then**
    $window\_radius \leftarrow 15$
else if $fraction\_low < 0.1$ **then**
    $window\_radius \leftarrow 10$
else
    $window\_radius \leftarrow 5$
end if

$result \leftarrow$ Create a white image, same shape as input
for all foreground pixels in low_detail **do**
    $window_{\text{high}} \leftarrow$ window of $high\_detail$ centered at
    pixel with radius $window\_radius$.

    Add foreground pixels in $window_{\text{high}}$ to
    corresponding region in the $result$ image.
end for
if fraction of foreground pixels in $result > 0.5$ **then**
    Apply morphological opening to $result$ using $morph\_structure$
else
    Apply morphological erosion to $result$ using $morph\_structure$
end if