AN EXTENSIVE REVIEW OF METRICS FOR EVALUATING IMAGE BINARIZATION ALGORITHMS

Giorgiana Violeta VLĂSCEANU[1]

Cristian AVATAVULUI[2]

Costin-Anton BOIANGIU[3]

Abstract

This comprehensive review paper delves into the essential metrics utilized for the evaluation of image binarization algorithms. Image binarization, a pivotal preprocessing step in many computer vision and image processing systems, poses significant challenges regarding the quality of output. Hence, a diverse range of evaluation metrics has been introduced, each bearing its strengths and limitations. This paper aims to elucidate the fundamental metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio ($PSNR$), F-measure, Pseudo F-measure, and Distance-Reciprocal Distortion Measure (DRD), explicating their definitions, interpretations, advantages, and disadvantages. Furthermore, particular attention is given to the influential Document Image Binarization COntest (DIBCO) standards that have significantly shaped the field of image binarization evaluation. A comparative analysis of these metrics is performed, highlighting their effectiveness, accuracy, and suitability under diverse scenarios. This paper also identifies the existing limitations and proposes potential directions for future research in the realm of image binarization evaluation.

Keywords: Image Binarization, Evaluation Metrics, Mean Squared Error (MSE), Peak Signal-to-Noise Ratio ($PSNR$), F-measure, Pseudo F-measure, Distance-Reciprocal Distortion Measure (DRD), Document Image Binarization COntest (DIBCO)

JEL Classification: C80, C65

1. Introduction

1.1. Background on Image Binarization

Image binarization consists to the process of converting a gray-scale image into binary format, delineating objects of interest from the background [1]. It plays a pivotal role in

[1] PhD student, Eng., Teaching assistant, University Politehnica of Bucharest, Romania, giorgiana.vlasceanu@cs.pub.ro

[2] PhD student, University Politehnica of Bucharest, Romania, cristian.avatavului@stud.acs.pub.ro

[3] PhD, Eng., Professor, University Politehnica of Bucharest, Romania, costin.boiangiu@cs.pub.ro a

document image analysis and other computer vision tasks, serving as a foundational preprocessing step [2]. It simplifies complex images by reducing multilevel intensity information to two levels, foreground and background, thus accentuating regions of interest and making subsequent analyses more manageable. The quality of binarization can significantly impact the performance of these subsequent processes. Numerous algorithms have been proposed to conduct this task, starting from classic methods like Otsu's method [3], Niblack's method [4], and Kapur's entropy-based method [2], up to more recent deep learning-based approaches [5].

1.2. Importance of Evaluation Metrics

The efficacy of the binarization process must be evaluated because the quality of binarization can have a substantial impact on downstream tasks such as object identification, recognition, and tracking. Several evaluation metrics have been proposed to quantify the performance of binarization methods [6][7][8][9]. Metrics such as the Mean Squared Error (MSE), Peak Signal-to-Noise Ratio ($PSNR$), and F-measure provide a quantitative analysis of the binarization output, thus facilitating the comparison and selection of optimal binarization algorithms for different applications. Additionally, metrics like Distance-Reciprocal Distortion Measure (DRD) and Pseudo F-measure have also been introduced to address specific limitations of previous metrics [6][7]. These metrics offer different insights into the binarization performance, including the accuracy of foreground-background separation, noise reduction, and preservation of details.

1.3. Brief Introduction to DIBCO Evaluations

The Document Image Binarization COntest (DIBCO) has been a significant influence in the topic of document image binarization [10][11][12]. DIBCO provides a standard dataset and evaluation methodology, making it possible to compare binarization techniques objectively. Over the years, DIBCO has introduced several novel metrics tailored for document image binarization evaluation, including Pseudo F-measure and Distance-Reciprocal Distortion Measure [6][10]. The insights from DIBCO evaluations have led to creating improved binarization methods and continue to shape the landscape of image binarization research.

1.4. Aim and Structure of the Paper

This study intends to give an in-depth examination of the metrics used to evaluate picture binarization methods, with a special emphasis on DIBCO standards. We first provide an overview of image binarization and the need for various evaluation metrics. We then detail each primary metric used, discussing their definitions, interpretations, advantages, and

drawbacks. Attention is given to DIBCO's approach to image binarization evaluation. We subsequently present a comparative analysis of these metrics, addressing their relative strengths and limitations. Finally, we identify potential future directions in image binarization evaluation metrics.

2. Overview of Image Binarization

2.1. Definition and Utility of Image Binarization**

Image binarization, a critical preprocessing step in many computer vision systems, transforms from a grayscale image or color image a binary image, which consists of only two colors or intensity levels, commonly black and white [1]. This conversion helps segregate the object of interest, usually marked black, from the background, marked white (or the other way around, depending on the application), making it easier to analyze and process the image. Binarization has been a staple in a variety of applications, ranging from document image analysis [2] to object tracking, character recognition, and many more.

2.2. Common Challenges and Issues in Image Binarization

While the concept of binarization seems straightforward, its execution can be fraught with several challenges. Inadequate illumination, shadows, low contrast, noise, and variability of foreground and background intensities all pose significant issues in image binarization [13]. In document analysis, additional complexities such as varying text sizes, faded print, ink seepage, and paper degradation further complicate the binarization process [10]. Thus, a universal binarization method that works effectively under all conditions is yet to be established, which makes the task of binarization a vibrant area of ongoing research [14].

2.3. Existing Solutions and Algorithms for Image Binarization

Numerous algorithms have been developed to address the issues associated with image binarization. Some of the early techniques include Otsu's method, which uses the threshold that minimizes the within-class variance of black and white pixels [3], and Niblack's method, which employs local mean and standard deviation to adaptively select the threshold [4]. Kapur et al. devised an entropy-based method that utilizes the entropy of the histogram for threshold selection [2]. Other notable methods include a recursive thresholding technique proposed by Cheriet et al. [8] and Howe's document binarization technique that automatically tunes the parameters [9]. In the era of deep learning, Tensmeyer and Martinez [5] proposed a fully convolutional neural network approach to binarization, demonstrating superior performance over many traditional methods.

2.4. Importance of Evaluation Metrics for Image Binarization

With the diversity of binarization algorithms, a fair and objective comparison of their performance is crucial, which is where evaluation metrics come into play [6][7][15]. These measures quantify the success rate of binarization methods, facilitating the selection of the optimal method for a given application. They assess how well the algorithm can separate the object of interest from the background and preserve the object's details. Various metrics have been proposed, each providing different insights into the binarization performance. These include Mean Squared Error - MSE, Peak Signal-to-Noise Ratio - $PSNR$, F-measure, Pseudo F-measure, and Distance-Reciprocal Distortion Measure – DRD [6][7]. Furthermore, DIBCO's competitions have shaped the evaluation landscape by introducing novel metrics and a standardized evaluation methodology [10][11][12].

3. Overview of Image Binarization Metrics

3.1. Introduction to the Concept of Metrics

Metrics, in the context of image binarization, are quantitative measures employed to evaluate the performance of binarization algorithms [1]. They are vital tools for discerning the quality of the binarization output and comparing the effectiveness of various binarization methods. These metrics, applied to the binarized images, provide an objective measure of how well an algorithm has performed the task of separating the foreground (object of interest) from the background. They form the basis of robust and objective analysis, thereby driving the evolution of increasingly refined binarization algorithms [15].

3.2. Need for Different Kinds of Metrics

The diverse challenges and intricacies of the binarization process have led to the need for a variety of evaluation metrics. Each metric offers unique insights into the binarization output and addresses different aspects of the binarization process [6][7][15]. For instance, Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio ($PSNR$) are employed to assess the overall difference between the binarized image and the ground truth. In contrast, F-measure, Pseudo F-measure, and Distance-Reciprocal Distortion Measure (DRD) are more focused on the structural preservation and distortion aspects [6][7]. Therefore, different metrics serve different purposes and can collectively provide a holistic assessment of a binarization algorithm's performance.

3.3. General Explanation of Common Metrics

The common metrics used in the evaluation of image binarization are diverse, each offering unique insights.

3.3.1. Mean Squared Error (MSE)

This metric calculates the average of the squared differences between the corresponding pixels in the binarized image and the ground truth, thereby providing a measure of the overall difference between the two images [1].

3.3.2. Peak Signal-to-Noise Ratio ($PSNR$)

$PSNR$ is often used in conjunction with MSE. It is a measure of the peak error and provides an approximation of the perceived reconstruction quality of the binarized image [1].

3.3.3. F-measure

The F-measure is a harmonic means of precision and recall, two commonly used measures in information retrieval and machine learning. Precision is a binarization algorithm's capacity to properly identify foreground pixels, whereas recall measures the algorithm's ability to find all foreground pixels in an image [6].

3.3.4. Pseudo F-measure and Distance-Reciprocal Distortion Measure (DRD)

These metrics were introduced during the DIBCO competitions and aim to address some limitations of the F-measure and to provide a more comprehensive evaluation that takes into account both global and local distortions [7][10].

Collectively, these metrics serve as a reliable means for evaluating and comparing the performance of different image binarization algorithms [6][7][10][11][12].

4. Evaluation Metrics for Image Binarization

The evaluation metrics for image binarization aim to quantitatively assess the performance of various binarization algorithms. They provide an objective measure of how well an algorithm can segregate the object of interest from the background, preserve the details of the object, and minimize noise and distortion. In this chapter, we delve into the specifics of these metrics, elaborating on their definitions, methodologies, advantages, and potential limitations.

We discuss in detail Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio ($PSNR$) which measure the global quality of binarization by assessing the overall difference between the binarized image and the ground truth [1]. Furthermore, we elaborate on the F-measure, a widely used metric that considers both precision and recall to provide a more balanced performance measure [6]. In addition, we delve into the Pseudo F-measure and Distance-Reciprocal Distortion Measure (DRD), two metrics introduced in the Document Image Binarization COntest (DIBCO). These metrics, tailored for this process, are focus on assessing global and local distortions in the binarized image [6][7][10].

We also touch upon other less common but equally significant metrics used in image binarization and their applications. By dissecting these metrics, we aim to provide a comprehensive understanding of how the performance of binarization algorithms is quantified and compared, and how these evaluations inform the selection of the optimal binarization method for specific applications [6][7][10][11][12][15].

4.1. Mean Squared Error (MSE)

The Mean Squared Error (MSE) is a commonly used metric for quantifying the difference between two images, typically a binarized image and its corresponding ground truth [1]. It essentially evaluates the average squared difference between the corresponding pixels of the two images.

MSE can be represented mathematically as follows:

$$MSE = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{N}\left(I_{(i,j)} + K_{(i,j)}\right)^2$$

where $I_{(i,j)}$ and $K_{(i,j)}$are the pixel intensities at location (i, j) in the binarized image and ground truth. M and N are the dimensions of the images.

The interpretation of MSE is straightforward: a smaller MSE indicates a lesser difference between the two images - binarized image and ground truth, signifying a better performance of the binarization algorithm. In other words, a binarization method with a lower MSE has a higher fidelity to the original image [15]. The primary advantage of MSE is its simplicity and ease of computation, which makes it a popular choice in various image processing tasks [15]. Moreover, it provides a global measure of the overall difference between the two images - binarized and ground truth, offering a comprehensive assessment of the binarization quality.

However, MSE also has certain limitations. Firstly, it treats all errors equally, regardless of their spatial distribution or relevance to human perception. As a result, it might not align perfectly with the human visual perception of the quality of binarized images [6]. Secondly, it provides a global measure and can be overly sensitive to extreme values, thereby failing

to capture local distortions effectively [7]. This is particularly relevant in document image analysis where preserving local details, such as text structure, is crucial [10]. Despite these drawbacks, MSE remains a fundamental and widely used metric in the evaluation of image binarization algorithms due to its simplicity and interpretability [1][15].

4.2. Peak Signal-to-Noise Ratio ($PSNR$)

The Peak Signal-to-Noise Ratio - $PSNR$ - is another broadly used metric in image processing, often employed in conjunction with the Mean Squared Error - MSE [1]. The $PSNR$ is a measure of the highest possible power of a signal relative to the power of corrupting noise, providing an approximation of the perceived reconstruction quality of the binarized image.

$PSNR$'s mathematical representation is as follows:

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right)$$

where MAX is the greatest pixel value achievable in the image. In the case of 8-bit grayscale images, the MAX value is 255.

$PSNR$ provides an estimate of the quality degradation due to the noise introduced by the binarization process. A higher $PSNR$ suggests a better quality of the binarized image, corresponding to less noise or distortion introduced by the binarization algorithm [15]. One of the key advantages of $PSNR$ lies in its interpretability: the $PSNR$ value can be intuitively understood as it is measured in decibels (dB). A higher $PSNR$ signifies a greater *signal* relative to the *noise*, indicating a higher image quality [1].

However, similar to MSE, $PSNR$ also has its limitations. While $PSNR$ provides a useful approximation of reconstruction quality, it may not always reflect the subjective quality perceived by the human eye [6]. Certain distortions that are perceptually significant might yield high $PSNR$ values, causing a disconnect between the numerical evaluation and the visual quality of the binarized image. Furthermore, like MSE, $PSNR$ is a global measure and might not be sensitive to local distortions in the binarized image [7]. Nonetheless, due to its relative simplicity and interpretability, $PSNR$ remains a commonly used metric in the field of image binarization [1][15].

4.3. Specificity, Sensitivity (or Recall), and Precision

Specificity, Sensitivity (or Recall), and Precision are fundamental metrics used in binary classification tasks, including image binarization. Together, they provide a comprehensive view of an algorithm's performance [6][15].

4.3.1. Precision

The fraction of true positive predictions (foreground pixels accurately detected) out of all positive predictions made by the algorithm is measured by this statistic. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

Here, TP represents true positives and FP stands for false positives (background pixels incorrectly identified as foreground). Higher Precision indicates the algorithm's reliability in predicting a pixel as part of the foreground, reducing false alarms [6].

4.3.2. Sensitivity (or Recall)

This metric quantifies the proportion of true positive instances successfully identified by the algorithm. Mathematically, it is defined as:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

FN denotes false negatives (foreground pixels incorrectly identified as background). Higher Sensitivity means the algorithm is proficient at detecting all the foreground pixels, reducing missed detections [6].

4.3.3. Specificity

This metric quantifies the proportion of true negative instances (background pixels) that are correctly identified by the algorithm. It is calculated as:

$$\text{Specificity} = \frac{TN}{TP + FP}$$

TN denotes true negatives. A higher Specificity indicates that the algorithm has fewer false alarms for background pixels [15].

Together, these metrics provide a nuanced view of an algorithm's performance, balancing its accuracy for both foreground and background pixels while considering its propensity for false positives (Precision) and false negatives (Sensitivity).

However, each of these metrics captures only one aspect of the performance, which might lead to an incomplete picture. For instance, an algorithm may have high Precision but low Sensitivity, indicating it is overly conservative in predicting foreground pixels, or vice versa [6][15]. Furthermore, these metrics, being global, may not fully capture local distortions in the binarized image, an issue addressed by more complex metrics like the Pseudo F-

measure and the Distance-Reciprocal Distortion Measure (DRD) used in the DIBCO competitions [10][11]. Nonetheless, due to their simplicity and interpretability, Precision, Sensitivity, and Specificity are widely used as starting points for evaluating binary classification algorithms, including those for image binarization [6][15].

4.4. F-measure

The F-measure, frequently referred to as the F-score or F1-score, is a metric that blends the two crucial components of information retrieval: precision and recall [6]. It serves as a harmonic mean of these two values, providing a balanced measure of a binarization algorithm's ability to accurately identify foreground pixels (precision), and its capacity to find all the foreground pixels in the image (recall).

The mathematical representation of the F-measure is given as follows:

$$F - Measure = 2\frac{Precision \cdot Recall}{Precision + Recall}$$

Where Precision is defined as $\frac{TP}{TP+FP}$, and Recall is defined as $\frac{TP}{TP+FN}$. Here, TP denotes true positives, means the foreground pixels correctly identified by the binarization algorithm. FP signifies false positives, for this case the background pixels incorrectly classified as foreground, and FN represents false negatives, the foreground pixels incorrectly classified as background [6].

The interpretation of the F-measure is quite intuitive: a higher F-measure indicates better performance of the binarization algorithm. An F-measure of 1 denotes perfect precision and recall, whereas an F-measure of 0 implies complete failure in both aspects [6].

The primary advantage of the F-measure lies in its ability to provide a balanced evaluation of an algorithm's performance. It ensures that neither precision nor recall is disproportionately emphasized, mitigating the risk of an overly optimistic or pessimistic assessment [6].

However, the F-measure is not without its drawbacks. It assumes equal importance of precision and recall, which may not always be the case in certain applications. Furthermore, like the previously discussed metrics, the F-measure provides a global assessment and may not capture local distortions or structural details effectively [7]. In response to these limitations, metrics like the Pseudo F-measure and Distance-Reciprocal Distortion Measure (DRD) were introduced in the DIBCO competitions [10][11][12]. Despite these caveats, the F-measure remains a widely utilized metric in image binarization due to its interpretability and the balance it offers between precision and recall [6].

4.5. Pseudo F-measure as used in DIBCO

The Pseudo F-measure is a variant of the F-measure metric introduced in the Document Image Binarization COntest (DIBCO) to provide a more comprehensive evaluation of image binarization algorithms [10]. The traditional F-measure considers pixels individually and might not account for local structural distortions in the binarized image. The Pseudo F-measure addresses this by incorporating local pixel context into the metric, essentially measuring the degree to which the binarized image preserves the structure of the original image.

The calculation of the Pseudo F-measure involves the use of a pseudo recall and precision, defined by convolving the binary ground truth and binarization result images with a weighted circular window [10][12]. Then, the Pseudo F-measure is calculated in a similar manner to the traditional F-measure:

$$Pseudo\ F-Measure\ = 2\frac{Pseudo\ Precision\ \cdot Pseudo\ Recall}{Pseudo\ Precision\ + Pseudo\ Recall}$$

The Pseudo Precision and Pseudo Recall are integral components of the Pseudo F-measure and they are calculated using the concepts of convolution and a specific weighted circular window.

First, let's define a binary image B, where $B(i,j)$ represents the pixel at location (i,j). For a grayscale image, $B(i,j)\ = 1$ represents a foreground (or black) pixel, while $B(i,j)\ = 0$ represents a background (or white) pixel.

Now, the binary ground truth image and the binary result image obtained from a binarization algorithm are convolved with a weighted circular window W of radius r. The convolution operation is represented as:

$$C\ =\ B \otimes W$$

where $\otimes$ denotes the convolution operation. The weighted circular window W is defined as:

$$W(x,y) = \frac{1}{2\pi\sigma^2} e^{-\left(\frac{x^2+y^2}{2\sigma^2}\right)}$$

where (x, y) are coordinates in the window, and σ is a parameter related to the size of the window.

The Pseudo Precision and Pseudo Recall are then computed using these convolved images. The Pseudo Precision (PP) is calculated as:

$$PP = \frac{\sum_i \sum_j C_{Binarization(i,j)} \cdot G_{(i,j)}}{\sum_i \sum_j C_{Binarization(i,j)}}$$

where $C_{Binarization(i,j)}$ is the result of convolving the binarization result image with the window W, and $G_{(i,j)}$ is the ground truth binary image.

The Pseudo Recall (PR) is calculated as:

$$PR = \frac{\sum_i \sum_j C_{GroundTruth(i,j)} \cdot B_{(i,j)}}{\sum_i \sum_j C_{GroundTruth(i,j)}}$$

where $C_{GroundTruth(i,j)}$ is the result of convolving the ground truth image with the window W, and $B_{(i,j)}$ is the binarization result binary image [8,14].

These measures capture the degree to which the binarization result matches the local structure of the ground truth image, with a higher Pseudo Precision or Pseudo Recall suggesting a better match to the local structure. In the context of interpretation, a higher Pseudo F-measure indicates a better preservation of local structural information, leading to a higher quality binarized image [10]. One of the primary advantages of the Pseudo F-measure is its ability to evaluate both global and local distortions, making it a more sensitive and comprehensive measure for document image binarization [10][12].

However, the Pseudo F-measure is not without its drawbacks. The computation of Pseudo F-measure is more complex than traditional global metrics due to the requirement of convolving images with a weighted window. Moreover, as with any metric, the Pseudo F-measure might not capture all aspects of image quality and should ideally be used in conjunction with other metrics for a comprehensive evaluation of a binarization algorithm's performance [12]. Despite these limitations, the Pseudo F-measure has been widely adopted in the DIBCO evaluations due to its ability to provide a more detailed assessment of binarization quality, especially in preserving the structural integrity of document images [10][11][12].

4.6. Generalizations for F-measure and Pseudo F-measure

The F-measure or F-score is a popular metric in the field of information retrieval and machine learning, combining precision and recall into a single number. It is defined as the harmonic mean of precision and recall [7]. The F-measure and the Pseudo F-measure may be generalized by the introduction of a parameter, beta (β), to allow for differential weighting of precision and recall. The generic formula for F_β is as follows [11]:

$$F_\beta = \frac{(1+\beta^2)(Precision \cdot Recall)}{(\beta^2 \cdot Precision) + Recall}$$

The β parameter, in this context, regulates the degree of importance that is given to precision over recall [7]. If beta is set to 1, the F-measure becomes the F1-score, meaning that precision and recall are equally important [8]. However, by manipulating the beta parameter, one can adjust the F-measure to favor either precision or recall. A β greater than

1 gives more weight to recall, whereas a beta less than 1 gives more weight to precision [3]. The Pseudo F-measure, employed in the DIBCO competitions, is a variant of the F-measure and also incorporates the beta parameter to achieve a balance between precision and recall. However, unlike the traditional F-measure which utilizes pixel-based precision and recall, the Pseudo F-measure employs region-based precision and recall. The beta parameter plays the same role in the Pseudo F-measure as it does in the traditional F-measure, i.e., to give differential weighting to precision and recall [5].

In DIBCO, the beta parameter is usually set to 0.3, indicating that recall (or in the case of Pseudo F-measure, Pseudo Recall) is considered more important than precision (Pseudo Precision). This is a particularly suitable choice for document image binarization tasks where the priority is to retrieve as much text as possible from the image [15]. Nonetheless, the appropriate value of the beta parameter can differ based on the specific requirements of the task [2]. It should be noted that the selection of beta is critical, and it should reflect the relative importance of precision and recall for the particular problem or application under consideration [13].

4.7. Distance-Reciprocal Distortion Measure (DRD)

The Distance-Reciprocal Distortion Measure (DRD) is a metric specifically designed for document image binarization evaluation in the Document Image Binarization COntest (DIBCO) [11]. The DRD evaluates both the detection error (similar to F-measure) and the distortion error due to misclassification, making it more comprehensive than traditional metrics.

The DRD is defined as the average of the two terms: distortion of false negatives (D_{FN}) and distortion of false positives (D_{FP}).

$$D_{FN} = \sum \frac{\frac{D_{m(i,j)}}{1 + D_{m(i,j)}}}{N_{FN}} \qquad D_{FP} = \sum \frac{\frac{D_{m(i,j)}}{1 + D_{m(i,j)}}}{N_{FP}}$$

$$DRD = \frac{D_{FN} + D_{FP}}{2}$$

In these equations, $D_{m(i,j)}$ denotes the distance from a misclassified pixel at location (i, j) to the nearest correctly classified pixel. N_{FN} and N_{FP} represent the total numbers of false negative and false positive pixels, respectively. This distance measure is reciprocal, meaning that misclassifications far from correctly classified pixels are penalized more heavily [11][12].

A smaller DRD indicates better performance of the binarization algorithm, as it suggests fewer misclassifications and less distortion due to misclassifications. The DRD offers several advantages over traditional metrics. Firstly, it incorporates both detection and

distortion errors, providing a more complete picture of the binarization algorithm's performance. Secondly, by considering the distance of misclassified pixels, it gives a nuanced assessment that heavily penalizes gross misclassifications [11].

However, the DRD also has its limitations. The computation of the DRD is more complex than traditional metrics like MSE or $PSNR$, requiring the calculation of pixel distances. Moreover, DRD might be overly sensitive to minor distortions that have negligible impact on document readability. Lastly, like all other metrics, it might not fully align with human visual perception, necessitating the use of additional metrics for a comprehensive evaluation [12]. Despite these limitations, the DRD is extensively used in the DIBCO evaluations and is recognized for its ability to provide a detailed assessment of binarization quality, especially in the context of document images where both detection and distortion errors significantly affect the readability and subsequent processing of the documents [13,14].

4.8. Jaccard Index

Jaccard similarity coefficient or The Jaccard Index, is a measure used to compare sample set similarity and variety. It is used as an assessment metric in image binarization to quantify the agreement among the binarized output and the ground truth image [14].

The Jaccard Index is defined mathematically as the size of an intersection divided by the dimension of the union of the two groups. For a binarization task, it can be computed as:

$$Jaccard\ Index = \frac{TP}{TP + FP + FN}$$

where TP denotes true positives, FP represents false positives, and FN stands for false negatives. Essentially, this formula calculates the ratio of correctly identified foreground pixels (TP) to all pixels identified as foreground by either the ground truth or the binarization algorithm ($TP + FP + FN$) [14].

A higher Jaccard Index indicates a greater similarity comparing the binarized image and the ground truth, suggesting that the binarization technique is more effective.. The main advantage of the Jaccard Index lies in its simplicity and intuitiveness. It directly relates to the proportion of correctly classified pixels, providing a clear measure of the binarization algorithm's effectiveness [14]. However, the Jaccard Index also has certain limitations. As a global metric, it may overlook local distortions in the binarized image. Also, like any ratio-based metric, it can be sensitive to the class balance in the image. For instance, in an image with many background pixels, a few false positives may not significantly affect the Jaccard Index despite potentially causing noticeable visual artifacts [14].

Even of these limitations, the Jaccard Index is a useful metric that complements other metrics like the F-measure, Pseudo F-measure, and DRD, providing a comprehensive evaluation of the performance of image binarization algorithms [10][11][14].

4.9. Dice Coefficient

The Dice Coefficient is a statistic used to compare the similarity of two samples. It is also known as the Sørensen -Dice index or Dice Similarity Coefficient (DSC). It is used to analyze the agreement between the binarized picture and the ground truth in the context of image binarization [16]. The Dice Coefficient is computed by dividing the size of the intersection of the two sets by the total of their sizes. In terms of image binarization, it can be expressed as:

$$DSC = \frac{2TP}{2TP + FP + FN}$$

where TP represents true positives, FP denotes false positives, and FN stands for false negatives. Essentially, the Dice Coefficient calculates the proportion of correctly identified foreground pixels ($2TP$) against all pixels identified as foreground plus those incorrectly identified as background ($2TP + FP + FN$) [16].

A Dice Coefficient closer to 1 suggests a high similarity between the binarized image and the ground truth, indicating the superior performance of the binarization algorithm. The Dice Coefficient has its strengths in its straightforward interpretability and ease of calculation. It provides a balanced measure of the binarization algorithm's effectiveness by taking into account both the false positives and false negatives [16].

However, the Dice Coefficient, being a global measure, may not capture local errors in the binarized image effectively. Like other ratio-based measures, the Dice Coefficient can be sensitive to the class imbalance in the image. For example, in a predominantly background image, a few misclassified foreground pixels may not significantly affect the Dice Coefficient, despite potentially leading to noticeable visual artifacts [16]. Despite its limitations, the Dice Coefficient is a valuable measure that, alongside metrics like the F-measure, Pseudo F-measure, Jaccard Index, and DRD, provides a well-rounded evaluation of image binarization algorithms [10][11][14][16].

4.10. Matthews Correlation Coefficient (MCC)

The Matthews Correlation Coefficient - MCC, sometimes referred as the phi coefficient, is a binary classification measure that takes consider true and false positives and negatives. It is often considered as a harmonious measure that can be used even though the categories differ greatly distinctive in terms of size [12]. In essence, the MCC is a correlation coefficient between the discovered and anticipated binary classifications. It returns a value

within -1 and +1. In this case, +1 denotes flawless prediction, a coefficient of 0 denotes no better than an arbitrary estimation, and a coefficient of -1 denotes entire disagreement between forecast and observation. The MCC can be calculated based on the elements of a confusion matrix, which are the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN), as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The main advantage of the MCC over other metrics like accuracy, F-score, or the area under the ROC curve (AUC), is that it is a more reliable statistical rate that only produces a high score if the prediction performs effectively in all four confusion matrix sections (TP, FP, TN, FN), as contrasted with other rates, which can yield misleading results even if TP, FP, TN and FN rates are imbalanced. In other words, the MCC considers both the over-prediction and under-prediction of each class and gives a balanced measure of the quality of binary classifications [12].

However, one of the limitations of the MCC is that it does not extend naturally to multiclass classification and does not have a clear interpretation in terms of probabilities or odds ratios. Also, while it does offer a more balanced perspective, it can be more difficult to interpret and explain than simpler statistics such as accuracy or F1 score [12]. Although with these drawbacks, the MCC is widely regarded as a reliable statistical accuracy measure in situations where the classes are imbalanced and provides a good complement to other evaluation metrics used in image binarization evaluations [12].

5. Evaluation Metrics in DIBCO Image Binarization Evaluations

5.1. Overview of DIBCO evaluations

Document Image Binarization Contest (DIBCO) is a benchmarking initiative that provides standard datasets and evaluation methodologies for the field of image binarization [3]. Since its inception in 2009, DIBCO has been instrumental in promoting innovative solutions for document image binarization, which is a crucial preprocessing step in many document image analysis and recognition systems [3].

5.2. Significance and Influence of DIBCO in the Image Processing Community

The significance of DIBCO in the image processing community is far-reaching. As an international competition, it brings together researchers worldwide, fostering a sense of collaboration and competition in advancing image binarization techniques. DIBCO datasets are composed of a wide variety of images, including handwritten and printed texts, historical documents, and texts under different noises and degradations. These diverse

datasets present challenges and opportunities, enabling researchers to develop and test robust binarization algorithms that can handle real-world situations [3].

5.3. Discussion of Metrics used in DIBCO

DIBCO evaluations employ a comprehensive set of metrics to evaluate the performance of binarization algorithms. The F-measure, Pseudo F-measure, DRD, $PSNR$, and MSE are all part of the evaluation toolkit. Furthermore, DIBCO also introduced other measures like the Negative/Positive Rate (NPR), Misclassification Penalty (MP), and Optical Character Recognition (OCR) error to evaluate the binarization results [2][3][5][13][15].

The F-measure used in DIBCO is a harmonic mean of precision and recall, while the Pseudo F-measure introduces weighted factors into the precision and recall calculation, making it more sensitive to certain types of errors [2][5]. DRD measures the average minimum distance between the boundary pixels in the binarized and reference images, providing a unique perspective into the quality of binarization [13]. $PSNR$ and MSE offer measures of the error between the binarized and reference images, each with its own strengths and weaknesses [4].

5.4. Comparison of DIBCO's Approach with Other Evaluation Methods

While other evaluation methods often rely on a single metric or a small set of metrics, DIBCO's approach stands out due to its wide-ranging and comprehensive evaluation using multiple metrics, which is designed to capture different aspects of binarization performance. This multi-metric evaluation approach presents a more complete picture of the binarization algorithm's performance, enabling researchers to identify the strengths and weaknesses of their algorithms and guide their improvements [3].

DIBCO's datasets, evaluation metrics, and methodologies have become a benchmark in the field of image binarization. They have significantly contributed to the development of more effective and efficient image binarization techniques, leading to improved performance in various applications such as document image analysis, OCR, and historical document digitization [1][3][6][7]. The influence and significance of DIBCO in the image processing community continues to grow, with its datasets and evaluation methodologies being widely used and cited in related research [3].

6. Comparative Analysis of Metrics

6.1. Discussion on How Different Metrics Relate to Each Other

In the evaluation of image binarization algorithms, different metrics provide different perspectives on the performance of the algorithms. MSE and $PSNR$, for example, are both

error metrics that quantify the difference between the image binarized and the original one. MSE calculates the mean squared difference between pixel intensities, while $PSNR$ is based on "the ratio between the maximum possible power of a signal and the power of corrupting noise" [4]. As such, they are inversely related; a lower MSE indicates a higher $PSNR$.

Similarly, the notions of True Positives (TP), False Positives (FP), and False Negatives (FN) underpin Precision, Recall, and the F-measure. Precision is the proportion of properly recognized positives in comparison to all identified positives (TP and FP), whereas recall is the proportion of correctly identified positives in comparison to all real positives (TP and FN) [12]. The F-measure is the harmonic mean of Precision and Recall with the goal of balancing these two measures [2]. The Pseudo F-measure, as employed in DIBCO, changes these calculations by taking weights into account [5].

6.2. Comparison of Metrics in Terms of Their Effectiveness, Accuracy, and Usability

In terms of efficacy, accuracy, and usefulness, each measure has advantages and disadvantages. MSE and $PSNR$ are easy to calculate and understand but may not always reflect the perceptual quality of the binarized image [4]. The F-measure and Pseudo F-measure, while they provide a balanced measure, can be sensitive to the choice of the beta parameter [2][5]. DRD, on the other hand, measures the average minimum distance between the boundary pixels in the binarized and reference images and can be useful in situations where boundary preservation is of importance, such as in text recognition tasks. However, the calculation of DRD can be computationally intensive [13].

The usability of these metrics can depend on the specific requirements of the application. In some cases, a simple, easily interpretable metric like MSE or $PSNR$ might suffice, while in other cases, a more sophisticated measure like the Pseudo F-measure or DRD may be necessary [3][2][13].

6.3. Case Studies Demonstrating Different Metrics' Performances in Different Scenarios

Several case studies have illustrated the performance of these metrics in different scenarios. For example, in the DIBCO evaluations, the Pseudo F-measure was found to be particularly effective in identifying algorithms that performed well in preserving text stroke width, an essential feature for text readability and subsequent OCR processing [3][5]. On the other hand, measures like MSE and $PSNR$ have been found to be less effective in this context, as they do not directly consider structural aspects like stroke width [4].

In another study involving historical document image binarization, the F-measure was found to be more effective than $PSNR$ in assessing the quality of the binarization [10]. This shows that the choice of the evaluation metric can significantly impact the perceived

performance of binarization algorithms, and the choice of metric should consider the specific characteristics and requirements of the application domain [10][5][14].

7. Limitations and Challenges in Current Evaluation Metrics

7.1. Identification of Gaps in Current Evaluation Methods

Despite the range of available evaluation metrics for image binarization, there are still gaps in current methods. One of the significant challenges is the lack of consensus on a universal metric that can cater to all types of images and applications [8]. While metrics such as MSE and $PSNR$ are good for quantifying the overall difference between binarized and reference images, they may not capture certain aspects like structural preservation [4]. Conversely, more specialized metrics like DRD and the Pseudo F-measure are sensitive to certain features like boundaries and stroke width but may not be suitable for all types of images [4,9]. Another gap is the heavy reliance on ground truth or reference images for the computation of most of these metrics [8]. While this approach is ideal for benchmarking purposes, it is not always feasible in practical applications where ground truth images may not be available.

7.2. Challenges and Issues in Implementing and Interpreting the Metrics

Implementing and interpreting the metrics presents its own set of challenges. The computation of some metrics, such as the DRD, can be complex and computationally intensive, which may not be practical in some situations [13]. In addition, the interpretation of results can sometimes be ambiguous due to the trade-off nature of certain metrics like Precision and Recall, and their derived metrics such as the F-measure and Pseudo F-measure [2][5].

For metrics like the MSE and $PSNR$, while they provide a straightforward numerical measure of the error, their interpretation in terms of perceptual quality can be non-intuitive [4]. High $PSNR$ or low MSE does not necessarily correlate with a high-quality binarized image, especially when the noise is non-uniform or structured.

7.3. Discussion on How These Challenges Might Be Overcome

To overcome these challenges, future research could focus on developing more sophisticated metrics that balance the trade-off between complexity and effectiveness, as well as developing methods for metric computation that do not heavily rely on ground truth images [8]. One potential approach could be to incorporate machine learning techniques to predict the quality of binarized images based on features learned from a large set of training images [9].

In terms of interpretation, it might be beneficial to provide guidelines or frameworks for interpreting different metrics in different contexts. For instance, a guide on when to use $PSNR$ versus the F-measure based on the specific requirements of the image binarization task could be useful.

Additionally, to address the challenges of interpreting metrics like Precision and Recall, future work could focus on developing more intuitive visualizations and explanations of these metrics to aid in their understanding and use [11][14]. It's also crucial that future research work towards identifying an optimal set of metrics that can be used to evaluate binarization algorithms effectively across a wide range of scenarios.

8. Future Directions and Conclusions

8.1. Suggestions for New Evaluation Metrics or Improvements on Existing Ones

The field of image binarization evaluation metrics is ripe for innovative advancements. While current metrics offer valuable insights, further improvements and new metrics could enhance the precision and practicality of these evaluations [14]. As we grapple with the trade-off between simplicity and sensitivity, one recommendation is the integration of machine learning techniques to enhance metrics [9]. This could potentially lead to metrics that can better adapt to various types of images and binarization tasks.

An alternate approach might be the development of meta-metrics that incorporate multiple existing metrics into a single score [8]. This approach could leverage the strengths of individual metrics while mitigating their limitations. Similarly, current metrics could be refined to be more perceptually relevant. For instance, advancements could be made on MSE and $PSNR$ to make their interpretation more intuitively linked to perceived image quality [4].

8.2. Discussion on the Potential Future of Image Binarization Evaluation

The future of image binarization evaluation holds much potential. As machine learning and AI continue to permeate image processing, we expect these technologies to play a substantial role in enhancing image binarization evaluation [9]. New methodologies could emerge that learn from a diverse range of image data to predict binarization quality, leading to more reliable and adaptive evaluation processes.

Furthermore, as the image binarization field continues to evolve, we anticipate a growing interest in specialized metrics tailored to specific applications, such as document analysis or medical imaging. The continuing development and expansion of benchmark datasets and competitions, like DIBCO, will also be instrumental in driving this research forward [7].

8.3. Summary of Key Findings and Conclusions

In summary, this paper has provided an in-depth review of various evaluation metrics for image binarization, highlighting their respective strengths and weaknesses. The necessity for a balanced evaluation strategy that considers both general and task-specific attributes of the image binarization problem has been stressed.

The paper discussed the significant role of DIBCO in shaping the current understanding and application of these metrics [7]. We examined the limitations and challenges of current metrics and proposed potential directions for future research [13][8]. By combining traditional metric-based evaluations with emerging technologies and methodologies, the field can move towards more accurate, adaptive, and application-specific evaluations of image binarization algorithms [9].

As we move forward, it's vital to continue questioning and refining our approaches to ensure we are effectively evaluating and improving image binarization techniques.

References

[1] Mehmet SEZGIN, Bülent SANKUR - Survey over image thresholding techniques and quantitative performance evaluation - Journal of Electronic Imaging, 13(1), 146-168, 2004, doi: 10.1117/1.1631315

[2] Jagat N. KAPUR, Prasanna K. SAHOO, A. K. WONG - A new method for gray-level picture thresholding using the entropy of the histogram - Computer Vision, Graphics, and Image Processing, 29(3), 273-285, 1985, doi: 10.1016/0734-189X(85)90125-2

[3] Nobuyuki OTSU - A threshold selection method from gray-level histograms - IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62-66, 1979, doi: 10.1109/TSMC.1979.4310076.

[4] Wayne NIBLACK - An introduction to digital image processing - Prentice-Hall, Inc., 1986

[5] Chris TENSMEYER, Tony Martinez - Document Image Binarization with Fully Convolutional Neural Networks - 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 99-104, doi: 10.1109/ICDAR.2017.25

[6] Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - Performance evaluation methodology for historical document image binarization - IEEE Transactions on Image Processing, 22(2), 595-609, 2013, doi: 10.1109/TIP.2012.2219550.

[7] Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - A combined approach for the binarization of handwritten document images. Pattern Recognition Letters, 35, 3-15, 2014, doi: 10.1016/j.patrec.2012.09.026

[8] Mohamed CHERIET, J. N. SAID, Ching Y. SUEN - A recursive thresholding technique for image segmentation - IEEE Transactions on Image Processing, 7(6), 918-921, 1998, doi: 10.1109/83.679444.

[9] Nicholas R. HOWE - Document binarization with automatic parameter tuning - International Journal of Document Analysis and Recognition (IJDAR), 16(3), 247-258, 2013, doi: 10.1007/s10032-012-0192-x

[10] Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - ICFHR 2010 document image binarization contest - 2010 12th International Conference on Frontiers in Handwriting Recognition (pp. 741-746). IEEE, 2010, doi: 10.1109/ICFHR.2010.118

[11] Basilis GATOS, Konstantinos NTIROGIANNIS, Ioannis PRATIKAKIS - DIBCO 2009: Document Image Binarization Contest - 10th International Conference on Document Analysis and Recognition, 2009, pp. 1375-1382, doi: 10.1109/ICDAR.2009.246.

[12] Ioannis PRATIKAKIS, Basilis GATOS, Konstantinos NTIROGIANNIS - ICFHR2012 competition on handwritten document image binarization (H-DIBCO 2012) - 2012 International Conference on Frontiers in Handwriting Italy, 2012, pp. 817-822, doi: 10.1109/ICFHR.2012.216

[13] Bolan SU, Shijian LU, Chew Lim TAN - Binarization of historical document images using the local maximum and minimum - Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10). Association for Computing Machinery, New York, NY, USA, 159–166, 2010, doi: 10.1145/1815330.1815351

[14] E. BADEKAS, N. PAPAMARKOS - Automatic Evaluation of Document Binarization Results - Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2005, Springer, doi: 10.1007/11578079_103

[15] P.T. YAP, P. RAVEENDRAN - Image focus measure based on Chebyshev moments Journal of Electronic Imaging, 6(4), 420-429, 2004, doi: 10.1049/ip-vis:20040395.

Bibliography

E. BADEKAS, N. PAPAMARKOS - Automatic Evaluation of Document Binarization Results - Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2005, Springer, doi: 10.1007/11578079_103

Mohamed CHERIET, J. N. SAID, Ching Y. SUEN - A recursive thresholding technique for image segmentation - IEEE Transactions on Image Processing, 7(6), 918-921, 1998, doi: 10.1109/83.679444.

Basilis GATOS, Konstantinos NTIROGIANNIS, Ioannis PRATIKAKIS - DIBCO 2009: Document Image Binarization Contest - 10th International Conference on Document Analysis and Recognition, 2009, pp. 1375-1382, doi: 10.1109/ICDAR.2009.246.

Nicholas R. HOWE - A Laplacian energy for document binarization - 2011 International Conference on Document Analysis and Recognition, pp. 6-10, China, 2011, doi: 10.1109/ICDAR.2011.11.

Nicholas R. HOWE - Document binarization with automatic parameter tuning - International Journal of Document Analysis and Recognition (IJDAR), 16(3), 247-258, 2013, doi: 10.1007/s10032-012-0192-x

Jagat N. KAPUR, Prasanna K. SAHOO, A. K. WONG - A new method for gray-level picture thresholding using the entropy of the histogram - Computer Vision, Graphics, and Image Processing, 29(3), 273-285, 1985, doi: 10.1016/0734-189X(85)90125-2

Khurram KHURSHID, Imran SIDDIQI, Claudie FAURE, Nicole VINCENT - Comparison of Niblack inspired binarization methods for ancient documents - Document Recognition and Retrieval XVI, 7247, 72470U, 2009, doi: 10.1117/12.805827

G. LEEDHAM, Chen YAN, K. TAKRU, Joie Hadi TAN, Li MIAN - Comparison of some thresholding algorithms for text/background segmentation in difficult document images - International Journal on Document Analysis and Recognition, 6(3), 169-181, 2003, doi: 10.1109/ICDAR.2003.1227784

Shijian LU, Bolan SU, Chew Lim TAN - Binarization of badly illuminated document images through shading estimation and compensation - Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Brazil, 2007, pp. 312-316, doi: 10.1109/ICDAR.2007.4378723.

Wayne NIBLACK - An introduction to digital image processing - Prentice-Hall, Inc., 1986

Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - Performance evaluation methodology for historical document image binarization - IEEE Transactions on Image Processing, 22(2), 595-609, 2013, doi: 10.1109/TIP.2012.2219550.

Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - A combined approach for the binarization of handwritten document images. Pattern Recognition Letters, 35, 3-15, 2014, doi: 10.1016/j.patrec.2012.09.026

Konstantinos NTIROGIANNIS, Basilis GATOS, Ioannis PRATIKAKIS - ICFHR 2010 document image binarization contest - 2010 12th International Conference on Frontiers in Handwriting Recognition (pp. 741-746). IEEE, 2010, doi: 10.1109/ICFHR.2010.118

Nobuyuki OTSU - A threshold selection method from gray-level histograms - IEEE Transactions on Systems, Man, and Cybernetics, 9(1), 62-66, 1979, doi: 10.1109/TSMC.1979.4310076.

Ioannis PRATIKAKIS, Basilis GATOS, Konstantinos NTIROGIANNIS - ICFHR2012 competition on handwritten document image binarization (H-DIBCO 2012) - 2012 International Conference on Frontiers in Handwriting Italy, 2012, pp. 817-822, doi: 10.1109/ICFHR.2012.216

Jaakko SAUVOLA, Matti K. PIETIKÄINEN - Adaptive document image binarization - Pattern Recognition, 33(2), 225-236, 2000, doi: 10.1016/S0031-3203(99)00055-2

Mehmet SEZGIN, Bülent SANKUR - Survey over image thresholding techniques and quantitative performance evaluation - Journal of Electronic Imaging, 13(1), 146-168, 2004, doi: 10.1117/1.1631315

Faisal SHAFAIT, Daniel KEYSERS, Thomas M. BREUEL - Efficient implementation of local adaptive thresholding techniques using integral images - Document Recognition and Retrieval XV, 6815, 681510, 2008, doi: 10.1117/12.767755

Bolan SU, Shijian LU, Chew Lim TAN - Robust document image binarization technique for degraded document images - IEEE Transactions on Image Processing, 22(4), 1408-1417, 2013, doi: 10.1109/TIP.2012.2231089

Bolan SU, Shijian LU, Chew Lim TAN - Binarization of historical document images using the local maximum and minimum - Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10). Association for Computing Machinery, New York, NY, USA, 159–166, 2010, doi: 10.1145/1815330.1815351

Chris TENSMEYER, Tony Martinez - Document Image Binarization with Fully Convolutional Neural Networks - 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 99-104, doi: 10.1109/ICDAR.2017.25

P.T. YAP, P. RAVEENDRAN - Image focus measure based on Chebyshev moments Journal of Electronic Imaging, 6(4), 420-429, 2004, doi: 10.1049/ip-vis:20040395.