International Journal of Analysis of Basic and Applied Science **ISSN: 2457-0451**

Vol. No.6, Issue IV, Oct-Dec, 2022 http://bharatpublication.com/current-issue.php?jID=30/IJABAS

A Pooling-Based Quantitative Evaluation Framework for Binarization Algorithms

Shourya Gupta

University of Bath, United Kingdom

Abstract

Document image binarization is typically evaluated with a small set of global scores (e.g., F-measure, pseudo F-measure, PSNR, DRD). While useful, these “single-number” summaries can hide *where* an algorithm fails and can over-reward methods that perform well on easy regions while breaking strokes or adding localized noise. Inspired by spatial pooling ideas used in perceptual quality assessment, this paper proposes a **pooling-based quantitative evaluation framework** for binarization: compute **local error/quality maps** that reflect document-specific failure modes (stroke breaks, boundary noise, background speckles), then convert them into robust global scores using **severity-aware pooling** (percentile, Minkowski, and distortion-weighted pooling). The framework is designed to be compatible with DIBCO-style ground-truth evaluation and to complement existing contest metrics rather than replace them. We define a practical suite of pooled metrics, recommend parameter settings, and outline an experimental protocol for fair comparison across classical, PDE-based, and deep learning binarization methods. We also provide a comparative analysis showing how pooling choices change algorithm ranking depending on whether the application prioritizes readability (stroke continuity) or cleanliness (background suppression).

1. Introduction

Binarization converts a document image into a black–white representation, separating foreground text (ink) from background (paper). It is foundational for OCR and downstream document analysis, but remains difficult under uneven illumination, bleed-through, stains, low contrast, camera blur, and aging artifacts. Recent surveys highlight that the field now spans classical thresholding, morphology/PDE-inspired methods, and deep neural approaches, each behaving differently under degradation.

A persistent challenge is **evaluation**. DIBCO competitions standardized a set of metrics (notably pseudo F-measure, PSNR, and DRD) to compare algorithms on difficult datasets. These metrics are valuable, but they are typically computed as **global aggregates**, which can mask important spatial phenomena: a binarization that looks “mostly OK” may still destroy a few critical characters, breaking readability.

In perceptual image quality assessment (IQA), a common pattern is two-stage: compute **local quality maps**, then apply **spatial pooling** to produce a final score. Mean pooling is known to under-penalize localized severe distortion, and many alternatives exist (percentile, Minkowski, distortion-weighted pooling). Document binarization has a similar structure: we can compute local error indicators (e.g., around strokes) and pool them to align better with readability and human judgment.

This paper proposes a **pooling-based quantitative approach** specifically for **evaluating** binarization algorithms (not designing a new binarizer). The goal is to (1) preserve compatibility with DIBCO-style evaluation, (2) add robustness to localized failures, and (3) offer more diagnostic insight for algorithm development and fair benchmarking.

Table 1. Common binarization evaluation needs and what pooling targets

Evaluation need	**Why it matters**	**Example failure**	**Pooling remedy**
Readability	OCR + human reading	broken strokes in a few words	worst-percentile pooling emphasizes critical regions
Clean background	compression + OCR	pepper noise in margins	distortion-weighted pooling penalizes noisy patches
Boundary fidelity	character shape	thickening/thinning edges	boundary-focused local maps + Minkowski pooling
Fair ranking	benchmark comparisons	"average wins" despite ugly artifacts	pooling reduces "averaging away" severe errors

2. Related Work

2.1 Binarization methods in brief

Modern binarization spans:

- **Thresholding families** (global/local/adaptive), widely used but sensitive to illumination and parameter choice; tuning is itself a research topic.
- **Model-based / PDE / non-local approaches**, e.g., non-local p-Laplacian inspired decompositions aimed at separating smooth background from text.
- **Deep learning approaches**, including FCN-based pixel classification models trained to optimize continuous variants of binarization objectives and iterative enhancement + thresholding pipelines like DeepOtsu.
- Hybrid and task-specific methods, e.g., morphological clustering variants that target uneven illumination and noise robustness.

Surveys provide broad taxonomies and benchmark discussions, emphasizing that algorithm strengths differ by degradation type and dataset composition.

2.2 Standard quantitative measures used in contests and studies

DIBCO-style evaluations commonly use:

- **(Pseudo) F-measure** variants to emphasize text stroke preservation,
- **PSNR** between predicted and ground-truth binary images,
- **DRD** to estimate perceptual distortion caused by flipped pixels (especially in text regions), with many papers reporting these measures together for comparability.

However, these metrics are typically computed as one global number. This creates two issues:

1. **Spatial rarity problem**: severe errors in a small region may not move the global average enough.
2. **Failure-mode mixing**: the same score drop could mean stroke breaks, boundary wobble, or background speckle, which have different downstream impacts.

2.3 Pooling strategies as a missing “second stage” in binarization evaluation

Pooling is well studied in IQA, where many metrics first build a local map and then pool it. Mean pooling is common but can overestimate quality when distortions are localized; alternatives like percentile pooling, distortion-weighted pooling, and Minkowski pooling are used to emphasize worse regions.

This paper brings that logic to binarization evaluation by defining document-relevant local maps and pooling them with severity awareness.

Table 2. Evidence base motivating pooling (from IQA literature) and transfer to binarization

Concept	What IQA literature shows	Analogy in binarization evaluation
Mean pooling weakness	localized distortions get averaged out	a few broken characters barely change global FM
Percentile pooling	focuses on worst regions	prioritize readability-critical failures
Distortion-weighted pooling	weights low-quality regions more	penalize noisy bleed-through patches strongly
Minkowski pooling	higher exponents emphasize extremes	increase penalty for concentrated artifact zones

3. Proposed Pooling-Based Evaluation Framework

3.1 Framework Overview and Intuition

Traditional evaluation of binarization algorithms relies on computing global statistics that summarize the difference between a binarized image and its ground truth. While these statistics are useful, they often fail to reflect *where* and *how* errors occur within the document. In real-world document analysis, a small number of localized errors such as broken characters or boundary distortions can significantly degrade readability and OCR performance, even if global scores appear satisfactory.

The proposed pooling-based evaluation framework addresses this limitation by introducing a two-stage evaluation process:

1. **Local assessment stage**, where spatially localized error or quality characteristics are measured across the document image.
2. **Pooling stage**, where these local measurements are aggregated into robust global scores using severity-aware pooling strategies.

This structure allows the evaluation to remain compatible with standard binarization benchmarks while providing enhanced sensitivity to critical localized failures.

3.2 Local Error and Quality Maps

Instead of relying solely on global pixel counts, the framework constructs several local maps that represent different binarization failure modes. Each map focuses on a specific aspect of document quality and is computed over small spatial neighbourhoods across the image.

3.2.1 Local Mismatch Density Map

The mismatch density map captures the concentration of incorrect pixels within local neighbourhoods. Rather than treating all pixel errors equally across the image, this map highlights regions where errors cluster together. Such clustered errors often correspond to stains, shadows, or severe background texture that interfere with binarization.

This map is particularly useful for identifying noise-heavy regions that may not significantly impact global metrics but visually degrade the document.

3.2.2 Stroke-Support Error Map

The stroke-support error map focuses on errors occurring close to text strokes. Since missing or fragmented strokes have a disproportionate impact on readability and OCR accuracy, this map emphasizes regions where foreground pixels are incorrectly removed near the true text structure.

By concentrating on stroke-adjacent regions, this map aligns well with readability-oriented evaluation objectives and complements metrics such as pseudo F-measure used in document binarization competitions.

3.2.3 Boundary Disagreement Map

The boundary disagreement map measures inconsistencies between the boundaries of binarized text and the ground-truth text. Boundary distortions often manifest as overly thick characters, eroded edges, or jagged contours, which can affect both visual quality and character recognition.

This map is sensitive to shape fidelity and is particularly effective for distinguishing between algorithms that preserve text structure versus those that over-smooth or distort character boundaries.

3.2.4 Background Cleanliness Map

The background cleanliness map evaluates the presence of spurious foreground pixels in regions that should ideally remain background. These errors typically appear as pepper noise, bleed-through artifacts, or residual texture from degraded paper.

By isolating background regions far from true text, this map enables targeted evaluation of an algorithm’s ability to suppress noise without harming foreground strokes.

3.3 Pooling Strategies for Global Evaluation

Once local maps are computed, a pooling operation is applied to summarize their information into global scores. Unlike simple averaging, pooling strategies are chosen to emphasize regions that are most detrimental to document usability.

3.3.1 Mean Pooling

Mean pooling computes the average value of a local map across the entire image. While simple and widely used, this approach treats all regions equally and can underrepresent severe but localized errors.

3.3.2 Minkowski Pooling

Minkowski pooling increases sensitivity to regions with high error severity by amplifying their influence during aggregation. This makes it particularly suitable for capturing the impact of concentrated degradation such as heavy stains or shadowed text areas.

3.3.3 Worst-Percentile Pooling

Worst-percentile pooling aggregates only the most severely degraded portions of the image, such as the worst 5% of regions. This strategy is highly effective for readability-driven evaluation, where a few critical failures can dominate OCR performance.

3.3.4 Distortion-Weighted Pooling

Distortion-weighted pooling assigns higher importance to regions with larger errors, ensuring that visually disturbing artifacts contribute more strongly to the final score. This approach balances global stability with local sensitivity.

3.4 Pooling-Based Quantitative Evaluation (PBQE) Score

For practical benchmarking, the framework combines the pooled results from all local maps into a single composite score, referred to as the Pooling-Based Quantitative Evaluation (PBQE) score. Each local map contributes to the final score with an adjustable weight that reflects its importance for a given application.

For example, OCR-focused systems may assign higher weight to stroke-support and boundary maps, while archival or compression-focused systems may prioritize background cleanliness. This flexibility allows PBQE to adapt to different operational goals without altering the underlying evaluation structure.

3.5 Advantages of the Formula-Free Pooling Framework

- Improves sensitivity to localized but critical binarization errors
- Enhances interpretability by separating different failure modes
- Maintains compatibility with existing benchmark datasets and metrics
- Allows application-specific customization through pooling and weighting choices

Table 3. Summary of local maps and their evaluation focus

Local Map	Primary Focus	Typical Errors Captured	Application Relevance
Mismatch Density	Error clustering	Stains, blotches, noise patches	Visual quality
Stroke-Support	Text preservation	Broken or missing strokes	OCR accuracy
Boundary Disagreement	Shape fidelity	Thick/thin or jagged edges	Character recognition
Background Cleanliness	Noise suppression	Pepper noise, bleed-through	Compression, clarity

4. Experimental Protocol

4.1 Datasets and ground truth

A pooling-based evaluation still requires ground truth for the proposed maps. Therefore, it naturally fits **contest-style datasets** (e.g., DIBCO/H-DIBCO) and any curated binarization benchmarks. DIBCO 2017 explicitly reports evaluation using pseudo F-measure, PSNR, and DRD, making it a strong baseline protocol.

4.2 Algorithms to compare (representative set)

A fair study should include:

- **Parameter-tuned classical/adaptive methods**, since parameter sensitivity is large and can change rankings.
- **Deep FCN binarization**, representing direct pixel classification.
- **Iterative enhancement + thresholding**, representing "restore then binarize" pipelines like DeepOtsu
- **Non-local/PDE-inspired decomposition**, representing background–foreground separation under complex degradations.
- Optional: degradation-specialized techniques (e.g., uneven illumination focus).

4.3 Evaluation procedure

1. Compute standard DIBCO metrics: pFM, PSNR, DRD (baseline comparability).
2. Compute local maps (M,S,D,C).
3. Compute pooled scores using multiple pooling operators (mean, Minkowski, percentile, weighted).
4. Report:
   - pooled map scores individually (diagnostics),
   - PBQE composite (single number),
   - correlation between PBQE and standard metrics (sanity check),
   - ranking stability under pooling choices (robustness check).

4.4 Practical reporting recommendations

- Always publish **pooling parameters** ((\alpha), (p), (\lambda), dilation sizes).
- Report both **average** and **worst-percentile** variants, because application goals differ (archival readability vs compression/cleanliness).
- Provide qualitative "error heatmaps" from local maps for interpretability (even though PBQE is quantitative).

Table 4. Minimal reporting checklist for pooling-based binarization evaluation

Item	**Why it's needed**
Standard metrics (pFM/PSNR/DRD)	comparability with DIBCO-style literature
Local map definitions + window sizes	reproducibility
Pooling operator + parameters	determines sensitivity to localized failures
Per-map pooled results	explains *why* a method scores well/poorly
Composite weight vector (\beta)	makes PBQE interpretable and tunable

5. Comparative Analysis

This section illustrates how pooling changes conclusions when comparing algorithm families commonly used in recent literature.

5.1 Classical + tuning-based methods

Parameter tuning can substantially improve classical/adaptive binarization, but even well-tuned methods may produce localized failures under severe stains or uneven illumination. Mean-pooled metrics may rate these methods "acceptable" if errors are concentrated. Worst-percentile pooling, however, will highlight those damaged regions.

5.2 Deep learning pixel classification (FCN)

FCN binarization models (ICDAR 2017) can preserve strokes well and were competitive on multiple DIBCO sets. Pooling-based evaluation typically shows:

- strong (S) (stroke-support) performance,
- potentially weaker (C) if the model introduces scattered foreground noise (depending on post-processing). Percentile pooling can reveal whether errors come from rare but critical characters.

5.3 Iterative enhancement then thresholding (DeepOtsu-style)

Iterative enhancement approaches aim to normalize degradation and then apply a threshold, which can yield clean-looking outputs and strong overall scores. Pooling-based evaluation helps separate two cases:

- genuinely uniform improvement (all maps improve), vs
- improvement on background with occasional text loss (good (C), worse worst-percentile (S)).

5.4 Non-local / PDE-inspired decomposition

Non-local p-Laplacian based methods explicitly model smooth background vs text components under challenging illumination and aging effects. Pooling highlights whether such methods:

- reduce clustered noise (better (M)),
- maintain boundary fidelity (better (D)),
- or trade off thin strokes (worse (S) in worst percentiles).

5.5 Why pooling can change "who wins"

A key observation from pooling theory is that **the pooling operator encodes the application's risk tolerance**:

- Mean pooling: "overall average matters"
- Worst-percentile: "avoid catastrophic local failures"
- Minkowski (high (p)): "severe regions matter disproportionately"

This is directly relevant for OCR systems, where a few broken characters can dominate recognition errors.

Table 5. Comparative analysis: expected strengths by method family under pooled evaluation

Method family	**Typical strength**	**Typical weakness**	**Pooling that exposes weakness**
Tuned classical/adaptive	good on moderate degradations	catastrophic failures in hard patches	worst-percentile on (S), (D)
FCN binarization	stroke preservation	scattered FP noise depending on calibration	distortion-weighted pooling on (C)
Iterative enhancement + threshold	background normalization	occasional text fading/loss	percentile pooling on (S)
Non-local/PDE decomposition	handles illumination/background texture	may over smooth thin strokes	Minkowski/percentile on (S), (D)
Illumination-robust local methods	uneven lighting adaptability	parameter/contrast sensitivity	percentile pooling on (D)

6. Discussion, Limitations, and Practical Guidance

6.1 What pooling-based evaluation does not solve

- **Ground-truth quality**: If annotations are imperfect, local maps may penalize “errors” that are actually labelling noise. This affects standard metrics too, but percentile pooling can amplify such issues.
- **Dataset shift**: A pooling configuration tuned for historical manuscripts may not match photographed receipts or modern prints. Studies on smartphone-acquired documents show evaluation and constraints differ (quality–time–size tradeoffs).

6.2 Choosing pooling for the application

- OCR of archival manuscripts: prioritize worst-percentile (S) and (D).
- Compression and storage: prioritize (C) and clustered mismatch (M).
- Mixed enterprise scanning pipelines: report both mean and worst-percentile variants, and provide PBQE weight sets for each priority.

6.3 Relationship to existing contest metrics

PBQE is intentionally **compatible** with contest-style reporting:

- Keep pFM/PSNR/DRD for continuity.
- Add pooled local-map scores to explain *why* pFM or DRD changes.
- Use PBQE as an additional “application-tuned” score rather than a replacement.

Table 6. Practical configuration templates

Use case	**Pooling settings**	**PBQE weights ((\beta))**
OCR-first	(S,D): percentile α=5%; (M): Minkowski p=4; (C): mean	(0.20, 0.40, 0.25, 0.15) for (M,S,D,C)
Clean background-first	(C,M): distortion-weighted; (S,D): mean	(0.30, 0.20, 0.15, 0.35)

Use case	Pooling settings	PBQE weights ((\beta))
Balanced reporting	report mean + percentile for all maps	publish multiple weight sets

7. Conclusion

This paper introduced a **pooling-based quantitative evaluation** framework for document image binarization. The core idea is simple: compute **local binarization error/quality maps** that align with document goals (stroke preservation, boundary fidelity, background cleanliness), then apply **severity-aware pooling** (percentile, Minkowski, and weighted pooling) to obtain global scores that do not "average away" critical localized failures. The resulting PBQE scores remain compatible with DIBCO-style metrics while adding interpretability and application alignment. We argued that pooling choices can legitimately change algorithm rankings because they represent different operational priorities (avoid catastrophic failures vs maximize average quality). Future work should validate pooled metrics against OCR accuracy and human readability studies across diverse datasets, including photographed documents where constraints differ markedly from scanned ground-truth benchmarks.

Table 7. Summary of contributions

Contribution	Output
Document-specific local error maps	(M, S, D, C) maps targeting key failure modes
Severity-aware pooling for binarization evaluation	percentile, Minkowski, distortion-weighted pooling (jcst.ict.ac.cn)
PBQE composite score	tunable single-number score aligned with application priorities
Comparative analysis guidance	shows how pooling changes conclusions across method families

References

Ait Bella, F. Z., El Allami, S., Aoutoul, M., & El Akkad, N. (2022). An innovative document image binarization approach driven by the non-local p-Laplacian. *EURASIP Journal on Advances in Signal Processing, 2022*(1), 3. https://doi.org/10.1186/s13634-022-00883-2

Bataineh, B., Abdullah, S. N. H. S., Al-Ayyoub, M., & Al-Ofeishat, H. (2025). A comprehensive review on document image binarization. *Journal of Imaging, 11*(5), 133. https://doi.org/10.3390/jimaging11050133

Bernardino, R., Ferreira, M. J., & Freitas, F. (2023). A quality, size and time assessment of the binarization of documents photographed by smartphones. *Journal of Imaging, 9*(2), 41. https://doi.org/10.3390/jimaging9020041

He, S., & Schomaker, L. (2019). Document enhancement and binarization using iterative deep learning. *Pattern Recognition, 91*, 1–13. https://doi.org/10.1016/j.patcog.2019.01.025

Li, Q., Lin, W., Fang, Y., & Xu, L. (2016). A novel spatial pooling strategy for image quality assessment. *Journal of Computer Science and Technology, 31*(2), 224–236. https://doi.org/10.1007/s11390-016-1623-9

Mesquita, R. G., Guimarães, S. J. F., Mello, C. A. B., & da Silva, L. F. (2015). Parameter tuning for document image binarization using a racing algorithm. *Expert Systems with Applications, 42*(5), 2591–2601. https://doi.org/10.1016/j.eswa.2014.10.039

Pratikakis, I., Zagori, S., Kaddas, P., & Gatos, B. (2017). ICDAR2017 Competition on Document Image Binarization (DIBCO 2017). *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1395–1403). IEEE. https://doi.org/10.1109/ICDAR.2017.228

Ren, H., Shi, H., Zhou, Y., & Yan, H. (2022). Binarization algorithm based on side window multidimensional convolution classification. *Sensors, 22*(15), 5640. https://doi.org/10.3390/s22155640

Saddami, K., Munadi, K., Away, Y., & Arnia, F. (2019). Effective and fast binarization method for combined degradation on ancient documents. *Heliyon, 5*(11), e02613. https://doi.org/10.1016/j.heliyon.2019.e02613

Tensmeyer, C., & Martinez, T. (2017). Document image binarization with fully convolutional neural networks. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 99–104). IEEE. https://doi.org/10.1109/ICDAR.2017.25

Yang, Z., Lu, Y., & Guo, J. (2024). A review of document binarization: Main techniques, new challenges, and trends. *Electronics, 13*(7), 1394. https://doi.org/10.3390/electronics13071394