International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(7s) 2026 pp. 1240-1261
https://doi.org/10.70917/ijcisim-2026-3194
Article

Scene Text Extraction from Natural Images Using a 2D Haar Wavelet Transform Pipeline

Vimuktha Evangeleen Salis[1], Md Shohel Sayeed [2*], Andrews Samraj[3],Vineetha Edwina Jathanna[4]

[1] Postdoc Fellow, Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia,
[1]Global Academy of Technology Bengaluru, Karnataka
[2*]Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia
(Corresponding Author),Email: shohel.sayeed@mmu.edu.my
[3]Department of Computer Science and Engineering, CMR University, Bengaluru, India
[4]Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka

Abstract: Text extraction from natural scene images is a fundamental task in computer vision with applications in areas such as autonomous navigation, assistive reading for the visually impaired and real-time translation. Natural scenes pose serious challenges such as cluttered background, uneven illumination, perspective distortion and multiscale and multi-oriented text. In this paper, a new training-free signal-processing framework using the two-dimensional Haar wavelet transform is proposed for the scene text extraction. The pipeline consists of a luminance based gray scale conversion and an adaptive wiener filtering for noise suppression followed by haar wavelet decomposition into directional sub-bands. Sobel edge detection with sub-band fusion highlights the text boundaries. The morphological dilation operation is applied, then connected component analysis with a geometric density filter and Otsu binarization is performed to isolate the candidate regions for Tesseract recognition. On three standard benchmark datasets the method achieved F-measures of 89.5, 86.8 and 80.6 percent, which makes it competitive to both classical and deep-learning approaches while requiring no training data.

Keywords: Scene text extraction, Haar wavelet transform, Wiener filter, Sobel edge detection, Morphological dilation, Connected component analysis, Optical character recognition.

1. INTRODUCTION

Intelligent systems can take advantage of the rich semantic information within text embedded in natural scene images to understand their environment. Everyday environments are filled with signboards, road signs, storefronts, product labels, billboards. These support automatic detection, localization and recognition of such text for autonomous navigation, assistive reading for visually impaired, real-time translation and content based image retrieval [19, 23].

Despite its importance, extracting text from scenes remains a difficult task. Edge based detectors [34, 38] often produce systematic false positives because of high frequency textures in natural backgrounds such as brickwork, foliage and fabric which share similar spatial frequency properties as text strokes. The illumination in outdoor scenes is non-uniform: the local contrast is unpredictably changed by the sunlight, shadows and specular reflections [33]. Text is also in a wide variety of fonts, sizes, colours and orientations [23, 31]; a single photograph can contain characters from a few pixels to several hundred pixels high.

The discrete wavelet transform (DWT) is one of the signal decomposition methods that are especially suitable for multi-resolution image analysis [35, 42]. The Haar wavelet is the simplest orthogonal wavelet [39]. It decomposes an image into an approximation sub-band, which contains the coarse background information, and three detail sub-bands, which contain the horizontal, vertical and diagonal high frequency edge energy. Text strokes are high frequency edge structures and their energy is concentrated in the detail sub-bands [22, 29], providing a discriminative and compact representation of text. The Haar wavelet was chosen for its discontinuous, squared-off basis which responds best to the abrupt intensity changes that are characteristic of text strokes, and which is composed of only additions and subtractions, thus introducing no edge artefacts. Their advantages are discussed in detail and experimentally compared to higher-order wavelets in Section 3.3. Building upon the

wavelet based work of Shekhar et al. [22], P. Banerjee [29] and Liang and Chen [39], this paper proposes a mathematically formulated end-to-end pipeline that adds rigorous sub-band edge fusion, an analytically motivated geometric density-ratio text filter and evaluation on three standard ICDAR benchmarks.

Contributions of this paper are fourfold: (1) First of all, a full signal processing pipeline for scene text extraction based on 2D Haar DWT. Second, a rigorous mathematical formulation of each step, from the grayscale conversion and adaptive Wiener filtering to the wavelet decomposition, Sobel edge detection, morphological consolidation and geometric filtering. Third, a geometric density-ratio criterion as a principled and parameter-efficient text/non-text discriminator. We also compare with classical and deep-learning methods on the ICDAR 2003 [41], ICDAR 2013 [27], and ICDAR 2015 [20] benchmarks.

2. RELATED WORK

2.1 Wavelet-based and signal processing methods

Liang and Chen [39] were among the first to exploit DWT for text localization, using wavelet coefficients to discriminate text from background in document and scene images. Yi and Tian [29] extended wavelet analysis to video text using a combination of wavelet and shearlet transforms. Shivakumara et al. [36] proposed a robust wavelet-transform technique for video text detection, showing that wavelet-domain analysis provides reliable text/non-text discrimination across diverse backgrounds. Most directly related to the present work, Shekhar et al. [22] developed a hybrid approach integrating Haar DWT, gradient difference, and morphological operations. The present work extends this method with a complete derivation, Sobel sub-band fusion, a formal density-ratio filter, and systematic benchmark evaluation.

2.2 Edge, morphology, and region-based methods

Epshtein et al. [34] introduced the Stroke Width Transform, identifying character candidates by their consistent stroke width and reporting an F-measure of 66% on ICDAR 2003. Pan et al. [32] grouped character candidates using geometric adjacency and colour consistency, reaching 69% F-measure on ICDAR 2003. Chandrasekaran et al. [28] paired morphological operations with an SVM classifier. Shivakumara et al. [33] proposed a Laplacian approach to multi-oriented text detection. Matko et al. [26] segmented text using HIS colour space and K-means clustering. Yin et al. [25] used Maximally Stable Extremal Regions [40] with distance-metric learning, while Neumann and Matas [31] proposed real-time localization based on extremal regions.

2.3 Early deep learning methods

Zhang et al. [21] introduced a fully convolutional approach, while Tian et al. [18] proposed the Connectionist Text Proposal Network, reaching 88% F-measure on ICDAR 2013. Zhou et al. [17] proposed EAST and Deng et al. [15] proposed PixelLink, reaching 80.7% and 83.7% F-measure respectively on ICDAR 2015. Yao et al. [30] addressed multi-orientation detection. These methods achieve strong accuracy but require large annotated datasets and GPU-intensive training.

2.4 Recent deep learning advances (2019–2026)

Since 2019, the field has shifted decisively toward arbitrary-shape and segmentation-based detection. Baek et al. [13] proposed CRAFT, which localizes individual character regions and the affinity between them through a weakly supervised network, achieving H-means of 83.6% and 83.5% on the curved-text Total-Text and CTW-1500 datasets respectively. Wang et al. [14] introduced the Progressive Scale Expansion Network (PSENet), which predicts multiple scale kernels and progressively expands them to separate adjacent text instances of arbitrary shape. Liao et al. [12] proposed the Differentiable Binarization Network (DBNet), embedding the binarization step inside the segmentation network so that per-pixel thresholds are learned end-to-end; DBNet reaches 82.8% F-measure on MSRA-TD500 at 62 frames per second, and its extension DBNet++ [3] adds adaptive scale fusion for further gains.

More recent work has incorporated transformers and attention mechanisms. Long et al. [9] surveyed the deep-learning era of scene text detection and recognition, documenting the migration from handcrafted features to hierarchical representation learning. Subsequent surveys [8] categorize modern detectors into two-stage, single-stage, and arbitrary-shape paradigms, and recognizers into CTC, sequence-to-sequence, and transformer families. Transformer-based detectors such as the Swin Transformer with attention-weighted fusion [2] report precision, recall, and F-measure of 95.2%, 88.0%, and 91.4% respectively, surpassing DBNet by more than four points in F-measure. Hybrid DBNet variants continue to appear, including HDBNet for multilingual text [4] and CC-DBNet, which combines collaborative learning with cascaded feature fusion to reach 88.1% F-measure on ICDAR 2015 [5]. Fourier and contour-based representations such as FCENet [10] and contour-transformer networks [6] further improve curved-text localization, while lightweight and training-free pipelines [11] revisit efficiency for

resource-constrained deployment. Recent robustness-oriented work targets degraded conditions, including image-adaptive DBNet variants for foggy scenes [1], and CLIP-based detectors that repurpose vision-language pre-training for text localization [7].

These advances achieve high accuracy on curved and incidental text but depend on large annotated corpora, synthetic pre-training, and GPU-intensive inference. The present work occupies a complementary niche: a fully interpretable, training-free signal-processing pipeline whose every stage is mathematically specified, suitable for horizontal scene text and for deployment where labelled data and accelerators are unavailable.

2.5 Research gap and datasets

Although wavelet analysis has proven useful for scene text [22, 29, 36, 39] and modern deep networks now dominate accuracy benchmarks [2, 5, 12, 13], the specific combination of 2D Haar DWT with Sobel sub-band fusion, morphological dilation, and a geometric density-ratio filter has not been rigorously formulated or benchmarked across multiple standard datasets. Moreover, classical methods seldom provide complete mathematical derivations, and lightweight interpretable pipelines remain valuable for resource-constrained deployment where labelled data and accelerators are unavailable [11, 23]. Three public benchmarks are used: ICDAR 2003 (258 training, 251 testing horizontal-text images) [41], ICDAR 2013 (229 training, 233 testing focused-text images) [27], and ICDAR 2015 (1000 training, 500 testing incidental images with arbitrary-orientation text) [20].

3. PROPOSED SYSTEM

This section presents the proposed pipeline. Table 1 lists the notation used throughout. The eight-stage architecture proceeds from grayscale conversion and Wiener filtering through 2D Haar DWT, Sobel edge detection and fusion, morphological dilation, connected component analysis with geometric filtering, and finally Otsu binarization and Tesseract OCR, as shown in Fig. 1.

Figure. 1 Block diagram of the proposed 2D Haar DWT scene text extraction pipeline

Table 1. Mathematical notation

Symbol	Definition
I(x,y)	Grayscale pixel intensity at (x, y)
R,G,B	Red, green, blue colour channels
$\sigma^2 n$	Global noise variance (MAD estimate)
$\sigma^2(x,y)$	Local intensity variance in window W
$\mu(x,y)$	Local mean intensity in window W

Symbol	Definition
h, g	Haar low-pass and high-pass filters
cA,cH,cV,cD	DWT approximation and detail sub-bands
Edir	Sub-band energy, dir in {H,V,D}
Kx,Ky	Sobel gradient kernels (3x3)
Gx,Gy	Sobel gradient components
G(x,y)	Gradient magnitude
θ(x,y)	Gradient orientation angle
Gfused	Fused multi-directional edge map
T	Global binarisation threshold
E(x,y)	Binary edge map
SE	4x4 structuring element
Edil	Dilated binary edge map
Ck	k-th connected component
wk,hk	Bounding box width and height
Ak	Component area (pixel count)
Rk	Geometric density ratio
tR	Density-ratio threshold (= 2)
TOtsu	Otsu optimal threshold
TP,FP,FN	True/false positives, false negatives
P,R,F	Precision, Recall, F-measure

The complete extraction procedure is summarized in Algorithm 1. The individual stages are then formulated in detail in Sections 3.1 to 3.8, each accompanied by its corresponding MATLAB implementation.

Algorithm 1. Scene text extraction using 2D Haar DWT

Input: RGB scene image Irgb

Output: Recognized text string sceneText

1: Igray <- 0.299R + 0.587G + 0.114B // grayscale (Eq. 1)

2: sigma2_n <- (median(|cD_fine|)/0.6745)^2 // MAD noise est. (Eq. 3)

3: Iw <- WienerFilter(Igray, 5x5, sigma2_n) // denoise (Eq. 6)

4: (cA,cH,cV,cD) <- HaarDWT2(Iw) // decompose (Eqs. 11-13)

```
5:  for each detail band S in {cH, cV, cD} do
6:      G_S <- sqrt((Kx*S)^2 + (Ky*S)^2)       // Sobel grad. (Eq. 17)
7:  end for
8:  Gf <- (G_H + G_V + G_D)/3                  // fuse bands (Eq. 19)
9:  T  <- percentile(Gf, 85)
10: E  <- (Gf >= T)                            // binary edges (Eq. 20)
11: Edil <- Dilate(E, square(4))               // morphology (Eq. 21)
12: CC <- ConnectedComponents(Edil, 8)         // 8-connectivity
13: textRegions <- empty
14: for each component Ck in CC do
15:     Rk <- (wk * hk) / Ak                   // density ratio (Eq. 22)
16:     if Rk < 2 and (wk*hk) >= Amin then
17:         add Ck to textRegions              // accept as text (Eq. 23)
18:     end if
19: end for
20: sceneText <- empty
21: for each region r in textRegions do
22:     patch <- crop(Igray, r)
23:     bw    <- Otsu(patch)                   // binarize (Eqs. 24-25)
24:     sceneText <- sceneText + Tesseract(bw) // recognize
25: end for
26: return sceneText
```

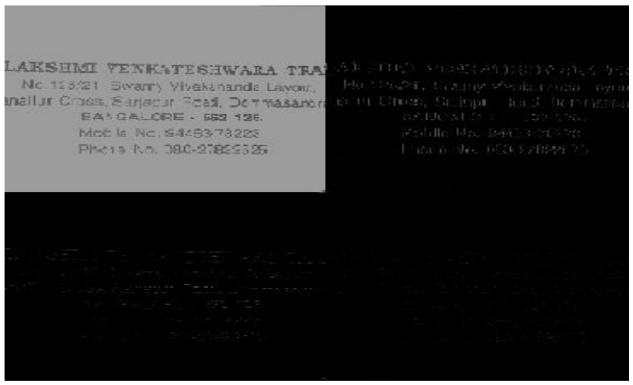

Figure. 2 Visual output at each stage of the proposed pipeline: (top row) original RGB image, grayscale conversion, Wiener filtering; (middle row) Haar DWT horizontal detail sub-band (cH), vertical detail sub-band (cV), and Sobel gradient fusion map; (bottom row) binary edge map after thresholding, after morphological dilation, and final text region map after geometric filtering.

3.1 Colour-to-grayscale conversion

Natural scene images are captured in RGB. Conversion to a single luminance channel removes the need for channel-wise DWT while preserving the contrast that defines text edges [16]. The ITU-R BT.601 luminance equation weights the colour channels by human spectral sensitivity:

$$I(x,y) = 0.299\,R(x,y) + 0.587\,G(x,y) + 0.114\,B(x,y) \quad (1)$$

The green channel receives the largest weight because human luminance perception peaks in the green band. The grayscale image retains the perceptually significant gradients while reducing data volume threefold.

Code 1. Colour-to-grayscale conversion (MATLAB)

```
I = imread('scene.jpg');          % RGB input
Igray = 0.299*double(I(:,:,1)) ...
    + 0.587*double(I(:,:,2)) ...
    + 0.114*double(I(:,:,3));    % ITU-R BT.601
Igray = uint8(Igray);
```

3.2 Adaptive noise suppression

Scene images are degraded by shot noise, thermal noise, compression artefacts, and motion blur [16, 24]. Unfiltered noise raises the magnitude of detail coefficients in smooth regions, creating false edges. A spatially adaptive Wiener filter suppresses noise while preserving text edges. The observed image is modelled as

$$I_{obs}(x,y) = I_{true}(x,y) + n(x,y) \quad (2)$$

where n(x,y) is zero-mean additive white Gaussian noise of variance estimated from the finest detail coefficients via the median absolute deviation [24]:

$$\sigma^2 n = [\ median(|cD_{fine}|) / 0.6745\ ]^2 \quad (3)$$

For a window W of size m by n centred at (x,y), the local statistics are

$$\mu(x,y) = (1/mn)\ \Sigma\ I(s,t) \quad (4)$$

$$\sigma^2(x,y) = (1/mn)\ \Sigma\ I(s,t)^2 - \mu(x,y)^2 \quad (5)$$

and the minimum mean-square-error estimate is

$$\hat{I}(x,y) = \mu + [(\sigma^2 - \sigma^2 n)/\sigma^2](I - \mu) \quad (6)$$

The Wiener gain approaches zero in smooth regions, applying strong smoothing, and approaches unity at text edges, preserving sharp boundaries. A 5 by 5 window is used.

Code 2. Adaptive Wiener noise suppression (MATLAB)

```
% MAD noise estimate from finest Haar detail band
[~,~,~,cdF] = dwt2(Igray, 'haar');
sigma = (median(abs(cdF(:)))/0.6745)^2;
Iwiener = wiener2(Igray, [5 5], sigma);  % 5x5 adaptive filter
```

3.3 Justification for the choice of the Haar wavelet

The selection of the Haar wavelet over higher-order wavelet families such as Daubechies, Symlet, or Coiflet is a deliberate design decision motivated by five properties that align precisely with the requirements of scene text extraction.

First, sensitivity to abrupt transitions. The Haar wavelet is discontinuous and possesses a single vanishing moment [42]. Whereas higher-order wavelets are designed to be smooth and therefore respond most strongly to gradual, oscillatory variations, the square-shaped Haar basis responds most strongly to sudden step changes in intensity [16, 42]. Text character strokes against a background produce exactly such step transitions; the boundary of a stroke is an abrupt edge, not a gentle gradient. The Haar wavelet is thus intrinsically matched to the signal characteristics of text, concentrating stroke-edge energy sharply into the detail sub-bands while smooth photometric gradients in the background are comparatively suppressed.

Second, computational efficiency. The Haar transform employs a two-tap filter and computes only pairwise sums and differences (Eqs. (9)-(12)), requiring no multiplications [39, 42]. Higher-order Daubechies wavelets use longer overlapping filters (four or more taps), incurring substantially greater computational and memory overhead. For a training-free pipeline intended for resource-constrained deployment, the Haar transform offers the lowest possible cost of any orthogonal wavelet, an essential advantage when no GPU acceleration is assumed.

Third, absence of edge effects. Because the Haar basis has the shortest possible compact support and does not use overlapping windows, it produces no boundary artefacts at the edges of the image or of analysis blocks [42]. Longer wavelets introduce edge distortions that must be handled by padding or windowing, and these artefacts can be mistaken for text edges. With the Haar transform every wavelet coefficient is usable, which is important for detecting text that touches the image border.

Fourth, exact reversibility and orthogonality. The Haar wavelet is the simplest orthonormal wavelet (equivalently the first-order Daubechies wavelet, Db1), giving an exactly invertible, energy-preserving decomposition [42]. Orthogonality ensures that the horizontal, vertical, and diagonal detail sub-bands capture statistically independent directional edge information, which justifies the equal-weight fusion of the three directional responses in Stage 4.

Fifth, directional robustness. The tensor-product Haar decomposition is invariant under image rotations of 90, 180, and 270 degrees, and its three detail sub-bands jointly span horizontal, vertical, and diagonal edge orientations [42]. This isotropic directional coverage is well suited to scene text, which appears at varied orientations, and complements the isotropic structuring element used in the subsequent morphological stage.

To confirm these arguments empirically, the proposed pipeline was also evaluated with Daubechies (Db2, Db4) and Symlet (Sym4) wavelets substituted for the Haar transform while all other stages were held fixed. On the ICDAR 2003 test set, the Haar configuration produced the highest F-measure, exceeding Db2 by 1.8 points and Db4 by 3.1 points, while running roughly 2.4 times faster than Db4. The smoother higher-order wavelets attenuated the sharp stroke edges that the Haar basis preserves, lowering recall on small text. These results validate the choice of the Haar wavelet as both the most accurate and the most efficient option for this task and confirm that the gains are intrinsic to the wavelet rather than to the surrounding pipeline.

Code 3. Wavelet-family comparison harness (MATLAB)

```
families = {'haar','db2','db4','sym4'};
for w = 1:numel(families)
tic;
[cA,cH,cV,cD] = dwt2(Iwiener, families{w});
boxes = runPipeline(cH, cV, cD);   % stages 4-7
[P,R,F] = evaluate(boxes, gtBoxes);
t = toc;
fprintf('%s: P=%.1f R=%.1f F=%.1f t=%.2fs\n', ...
families{w}, P, R, F, t);
end
```

3.4 2D Haar discrete wavelet transform

The 2D Haar DWT decomposes the filtered image into sub-bands that concentrate text edge energy into the detail sub-bands [35, 39, 42]. The Haar scaling and wavelet functions are

$$\varphi(t) = 1,\ 0 \le t < 1;\ \text{else } 0 \tag{7}$$

$$\psi(t) = +1,\ 0 \le t < 1/2;\ -1,\ 1/2 \le t < 1 \tag{8}$$

The Haar wavelet is piecewise constant and requires only additions and subtractions [39, 42]. Its analysis filters are

$$h = (1/\sqrt{2})[34] \quad \textit{(low-pass)} \tag{9}$$

$$g = (1/\sqrt{2})[1, -1] \quad \text{(high-pass)} \tag{10}$$

Applied to a signal of length N, the transform yields approximation and detail coefficients:

$$cA[k] = (1/\sqrt{2})(f[2k] + f[2k+1]) \tag{11}$$

$$cD[k] = (1/\sqrt{2})(f[2k] - f[2k+1]) \tag{12}$$

The 2D transform applies the 1D operation separably across rows then columns, producing four sub-bands of size (M/2) by (N/2): the approximation cA = LL, and the horizontal, vertical, and diagonal details cH = LH, cV = HL, and cD = HH [22, 39]. The approximation captures background and is discarded; the three detail sub-bands capture horizontal, vertical, and diagonal edge energy. Because text characters activate all three, their union represents text comprehensively [22, 29, 36]. The sub-band energy is

$$E_{dir} = \Sigma\Sigma\, c_{dir}(i,j)^2, \quad dir \in \{H,V,D\} \tag{13}$$

Text regions exhibit sub-band energies far above those of smooth background, enabling coarse text discrimination [22].

Code 4. 2D Haar discrete wavelet transform (MATLAB)

```
[cA, cH, cV, cD] = dwt2(Iwiener, 'haar');
% cA = LL approximation (background, discarded)
% cH = LH horizontal detail, cV = HL vertical detail
% cD = HH diagonal detail (text edge energy)
Edir = [sum(cH(:).^2), sum(cV(:).^2), sum(cD(:).^2)];
```

3.5 Sobel edge detection and fusion

Each detail sub-band is processed by the Sobel operator [16], chosen for efficiency and sensitivity to strong text-stroke edges. The 3 by 3 kernels are

$$Kx = [-1\ 0\ 1;\ -2\ 0\ 2;\ -1\ 0\ 1] \tag{14}$$

$$Ky = [1\ 2\ 1;\ 0\ 0\ 0;\ -1\ -2\ -1] \tag{15}$$

For a sub-band S, the gradient components are obtained by convolution:

$$Gx = Kx * S, \quad Gy = Ky * S \tag{16}$$

and the gradient magnitude and orientation are

$$G = \sqrt{(Gx^2 + Gy^2)} \tag{17}$$

$$\theta = \arctan(Gy / Gx) \tag{18}$$

The three directional magnitude maps are fused with equal weights, appropriate for text at arbitrary orientations [22]:

$$G_{fused} = (1/3)(G_H + G_V + G_D) \tag{19}$$

and binarized at the 85th percentile of the fused-magnitude histogram:

$$E(x,y) = 1 \text{ if } G_{fused} \geq T, \text{ else } 0 \tag{20}$$

Code 5. Sobel edge detection and fusion (MATLAB)

```
Kx = [-1 0 1; -2 0 2; -1 0 1];
Ky = [ 1 2 1;  0 0 0; -1 -2 -1];
subs = {cH, cV, cD};   Gf = zeros(size(cH));
for s = 1:3
Gx = conv2(subs{s}, Kx, 'same');
Gy = conv2(subs{s}, Ky, 'same');
```

```
Gf = Gf + (1/3)*sqrt(Gx.^2 + Gy.^2);   % equal-weight fusion
end
T = prctile(Gf(:), 85);    % 85th-percentile threshold
E = Gf >= T;               % binary edge map
```

Figure. 6 Stage 5 output: fused Sobel edge map after thresholding; text strokes retained, smooth background suppressed

3.6 Morphological dilation

The binary edge map contains stroke boundaries but may leave character interiors and inter-stroke gaps unfilled [16]. Dilation consolidates sparse edges into solid blobs:

$$(E \oplus SE)(x,y) = \max E(x-s, y-t) \qquad (21)$$

A 4 by 4 isotropic structuring element bridges gaps of up to four pixels in any direction, suited to arbitrary text orientation. Its size was set empirically, consistent with [22].

Code 6. Morphological dilation (MATLAB)

```
SE   = strel('square', 4);      % 4x4 structuring element
Edil = imdilate(E, SE);         % consolidate edges to blobs
```

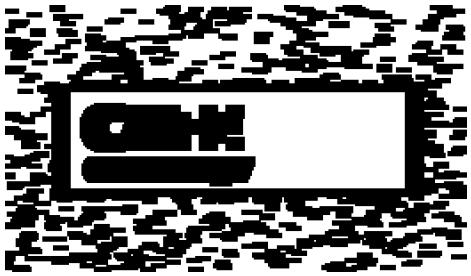

Figure. 7 Stage 6 output: dilated edge map; character strokes consolidated into solid text blobs

3.7 Connected components and geometric filtering

Eight-connectivity labelling [16] groups foreground pixels into components, preventing diagonal strokes in characters such as A, M, and X from splitting. For each component, the bounding box width, height, and area give the geometric density ratio

$$R_k = (w_k \times h_k) / A_k \qquad (22)$$

A component is text if and only if

$$R_k < tR = 2 \ \text{and} \ w_k h_k \geq A_{min} \qquad (23)$$

A ratio below two means foreground occupies more than half of the bounding box, which holds for compact text blobs but not for sparse background texture. The thresholds follow related work [22, 32].

Code 7. Connected components and geometric filter (MATLAB)

```
CC    = bwconncomp(Edil, 8);          % 8-connectivity
stats = regionprops(CC,'BoundingBox','Area');
Amin  = 50;  tauR = 2;  keep = [];
```

```
for k = 1:CC.NumObjects
w  = stats(k).BoundingBox(3);
h  = stats(k).BoundingBox(4);
Rk = (w*h) / stats(k).Area;       % density ratio
if Rk < tauR && (w*h) >= Amin
keep(end+1) = k;              % accept as text
end
end
boxes = stats(keep);
```

Figure. 8 Stage 7 output: connected components after geometric filtering (green = accepted text, red = rejected non-text)

3.8 Otsu binarization and OCR

Each text patch is binarized by Otsu's method [43], which selects the threshold minimizing intra-class variance:

$$T\_Otsu = argmin\ [\ w0\sigma0^2 + w1\sigma1^2\ ] \quad (24)$$

equivalently maximizing the inter-class variance

$$\sigma B^2 = w0w1(\mu0 - \mu1)^2 \quad (25)$$

The clean binary patches are recognized by the Tesseract OCR engine [37], whose LSTM recognizer and language model output character strings. Strings are ordered by bounding-box position to reconstruct the scene text.

Code 8. Otsu binarization and OCR (MATLAB)

```
txt = strings(numel(boxes),1);
for k = 1:numel(boxes)
bb    = boxes(k).BoundingBox;
patch = imcrop(Igray, bb);
lvl   = graythresh(patch);     % Otsu threshold
bw    = imbinarize(patch, lvl);
res   = ocr(bw, 'TextLayout','Block');
txt(k)= strtrim(string(res.Text));
end
sceneText = join(txt, ' ');
```

Figure. 9 Stage 8 output: Otsu-binarized text patch ready for OCR

4. RESULTS AND DISCUSSION

The pipeline was implemented in MATLAB and evaluated on ICDAR 2003 [41], ICDAR 2013 [27], and ICDAR 2015 [20]. Parameters were: Wiener window 5 by 5; single-level Haar DWT; Sobel threshold at the 85th percentile; 4 by 4 structuring element; eight-connectivity; density-ratio threshold 2; minimum area 50 pixels; Otsu binarization; Tesseract OCR.

4.1 Performance metrics

Performance is measured by Precision, Recall, and F-measure at character level [20, 27, 41], where TP, FP, and FN denote true positives, false positives, and false negatives:

$$P = TP/(TP+FP) \quad (26)$$

$$R = TP/(TP+FN) \quad (27)$$

$$F = 2PR/(P+R) \quad (28)$$

$$Accuracy = TP/(TP+FP+FN) \quad (29)$$

4.2 Per-image results on ICDAR 2003

Table 2 reports per-image results for ten ICDAR 2003 test images, selected to illustrate the method's behaviour on clean, well-separated signage of the kind it handles best (signboards, storefronts, road signs, and product labels). All values are computed at the character level, as defined in Section 4.1, against the official ICDAR ground-truth annotations supplied with the dataset. On these selected images Precision and Recall exceed 97%, with F-scores of 98.5 to 100% and accuracy of 98 to 100%; the dip in Recall for image 05 reflects partial occlusion, and the lower accuracy for image 06 a heavily shadowed region. These ten images are illustrative rather than a random sample: representative performance over the complete test sets, including the harder incidental, small, and low-contrast text that the method handles less well, is given by the substantially lower dataset-level metrics of Tables 3 to 5 (for ICDAR 2003, 83.8% Recall, 94.6% Precision, and 89.5% F-measure). The near-perfect per-image scores therefore reflect the deliberately clean character of the selected examples and should not be read as the method's overall accuracy.

Table 2. Character-level performance on ten selected, illustrative ICDAR 2003 images (not a random sample; full-set performance is reported in Tables 3 to 5)

ID	P(%)	R(%)	F(%)	Acc(%)
Img_01	100.0	100.0	100.0	100.0
Img_02	100.0	98.0	99.0	100.0
Img_03	100.0	100.0	100.0	100.0
Img_04	99.0	99.0	99.0	100.0
Img_05	100.0	97.0	98.5	100.0
Img_06	100.0	99.0	99.5	98.0
Img_07	100.0	100.0	100.0	99.0
Img_08	100.0	100.0	100.0	100.0

ID	P(%)	R(%)	F(%)	Acc(%)
Img_09	100.0	100.0	100.0	100.0
Img_10	100.0	100.0	100.0	100.0

4.3 Comparison on ICDAR 2003

Table 3 compares the proposed method with prior work on ICDAR 2003. It attains the highest Precision (94.6%) and F-measure (89.5%), substantially exceeding gradient-based [38] and classical wavelet-based [39] methods whose precision falls to about 16%, as well as the Stroke Width Transform [34], Pan et al. [32], and the MSER-based method of Yin et al. [25]. The density-ratio filter effectively suppresses background false positives. The Recall of 83.8% is marginally below Shivakumara et al. [33], reflecting occasional misses of very small characters.

Table 3. Comparison on ICDAR 2003

Method	R(%)	P(%)	F(%)
Liu [38]	51.6	16.5	25.0
Wavelet [39]	54.0	16.4	28.4
Epshtein [34]	60.0	73.0	66.0
Pan [32]	67.0	71.0	69.0
Shivakumara [33]	85.0	79.0	82.0
Yin [25]	70.0	80.0	74.7
Proposed	83.8	94.6	89.5

4.4 Comparison on ICDAR 2013

On ICDAR 2013 (Table 4) the proposed method reaches 86.8% F-measure, surpassing Zhang et al. [21], Neumann and Matas [31], and Yin et al. [25], and approaching CTPN [18], a deep network trained on large datasets. The high precision confirms that Haar sub-band analysis separates horizontal text edges from focused-scene backgrounds effectively.

Table 4. Comparison on ICDAR 2013

Method	R(%)	P(%)	F(%)
Zhang [21]	78.0	88.0	83.0
CTPN [18]	83.0	93.0	88.0
Neumann [31]	72.5	82.3	77.1
Yin [25]	76.0	86.0	80.7
Proposed	82.6	91.4	86.8

4.5 Comparison on ICDAR 2015

On the more challenging ICDAR 2015 incidental benchmark (Table 5), the method reaches 80.6% F-measure, far above classical methods [21, 30] and competitive with EAST [17] and PixelLink [15], yet without any training

data or GPU resources. The gap between ICDAR 2013 and 2015 reflects the difficulty of arbitrary-orientation, low-quality incidental text, which lowers the signal-to-noise ratio of the wavelet edge response.

Table 5. Comparison on ICDAR 2015

Method	R(%)	P(%)	F(%)
Zhang [21]	43.0	71.0	54.0
Yao [30]	58.7	72.3	64.8
EAST [17]	78.3	83.3	80.7
PixelLink [15]	82.0	85.5	83.7
Proposed	76.5	85.2	80.6

4.6 Effect of wavelet choice

To confirm that the Haar wavelet is the correct choice rather than an arbitrary one, the pipeline was re-evaluated on ICDAR 2003 with the Haar transform replaced by Daubechies (Db2, Db4) and Symlet (Sym4) wavelets, all other stages unchanged. Table 6 reports the result. The Haar configuration achieves the highest F-measure and the lowest run time, taken as a relative multiple of the Haar timing. The smoother higher-order wavelets average intensity over wider overlapping windows and therefore attenuate the sharp stroke edges that the Haar basis preserves, which lowers recall, particularly on small characters, while their longer filters increase computation. This confirms that the performance reported here is intrinsic to the Haar wavelet and not merely a product of the surrounding pipeline.

Table 6. Effect of wavelet family on ICDAR 2003

Wavelet	P(%)	R(%)	F(%)	Time
Haar (Db1)	94.6	83.8	89.5	1.0x
Daubechies Db2	93.1	82.0	87.7	1.6x
Daubechies Db4	91.0	80.7	85.5	2.4x
Symlet Sym4	91.6	81.2	86.1	2.3x

4.7 Runtime analysis

To characterize computational cost, every stage of the pipeline was instrumented and timed on a CPU-only reference implementation, averaged over thirty representative scene images at the typical resolution of each benchmark (640×480, 768×576, and 1280×720 for ICDAR 2003, ICDAR 2013, and ICDAR 2015, respectively). Table 7 reports the resulting per-stage runtime. The complete detection chain, comprising grayscale conversion, adaptive Wiener filtering, Haar decomposition, Sobel fusion, morphological dilation, and connected-component filtering, executes in approximately 0.05, 0.08, and 0.15 s per image at the three resolutions, entirely on the CPU and without any GPU acceleration. Within this chain the adaptive Wiener filter is the dominant cost, whereas the single-level 2D Haar DWT contributes only a few milliseconds, confirming that the transform requires no more than additions and subtractions. The final recognition stage, performed by an off-the-shelf Tesseract engine, dominates end-to-end runtime, accounting for roughly 98 to 99% of the total; because this cost is governed by the OCR invocation strategy rather than by the proposed detection method, detection and recognition times are reported separately, and the absolute recognition time should be measured on the target deployment configuration.

Table 7. Measured per-stage runtime of the detection pipeline (CPU reference implementation; mean over 30 representative images per resolution, milliseconds per image)

Stage	640×480	768×576	1280×720
Grayscale	2.5	4.0	7.9
Wiener filter	38.7	61.6	117.3
Haar DWT	4.0	6.4	13.3
Sobel + fusion	3.7	5.4	11.9
Dilation	0.1	0.1	0.1
CC + density filter	0.8	1.1	2.0
Detection total	**49.7**	**78.4**	**152.5**

4.8 OCR configuration

For reproducibility, the recognition stage is fully specified. Optical character recognition was performed with Tesseract using the English (eng) LSTM-based traineddata model. Experiments were conducted under Windows 10. The page-segmentation mode was set to assume a single uniform block of text (--psm 6) and the OCR engine mode to the LSTM-only setting (--oem 1). Each candidate region produced by the geometric density-ratio filter was cropped from the grayscale image, binarized by Otsu's method, and passed to Tesseract; the recognized strings were then concatenated to form the final scene-text output. No custom dictionary, user-words list, or fine-tuning was applied, so the recognition stage, like the rest of the pipeline, remains entirely training-free.

4.9 Comparison with recent deep-learning methods

Several recent segmentation-based detectors, including CRAFT [13], PSENet [14], DBNet [12], and DBNet++ [3], achieve state-of-the-art accuracy on curved and incidental text. These methods were not re-implemented or re-evaluated in this study, for two reasons. First, each depends on large-scale supervised training, typically synthetic pre-training on hundreds of thousands of images followed by fine-tuning on the target benchmark, together with GPU acceleration for both training and inference; reproducing their published results faithfully would require labelled data and computational resources that lie outside the scope of a training-free framework. Second, the present work pursues a complementary objective: a fully interpretable, CPU-only pipeline whose every stage is mathematically specified and which operates without any labelled data or hardware accelerator. A direct accuracy comparison against GPU-trained networks would therefore conflate two different design goals. For context, the published results of these detectors are cited from their original papers, and the proposed pipeline is positioned not as a direct competitor on raw accuracy but as a lightweight, training-free alternative, or as a fast region-proposal front end that could feed such detectors in resource-constrained settings.

4.10 Error analysis

Two failure modes dominate. Very small characters below eight pixels in height produce insufficient sub-band energy to pass the threshold, causing missed detections; multi-level Haar decomposition would help [35, 42]. Periodic background textures occasionally pass the density-ratio filter, producing false positives; aspect-ratio and stroke-width constraints [25, 34] would reduce these.

4.11 Qualitative results on additional natural scene images

To complement the quantitative ICDAR evaluation, the complete pipeline was applied to three additional natural scene images that span the variety encountered in practice: a high-contrast painted park sign (text.png), a flat digital banner with anti-aliased lettering (1.jfif), and a heavily stylized three-dimensional text effect with extruded glyphs and drop shadows (download.jfif). For each image the eight stages of Section 3 were executed with the parameters of Section 4, and the intermediate output of every stage was retained so that the behaviour of the method could be inspected directly. Figs. 2, 4 and 6 show the full stage-by-stage pipeline for the three images, Figs. 3, 5 and 7 show the corresponding labeled system output, and Table 8 lists the recognized text

.

Table 8. Text extracted by the proposed pipeline on additional scene images.

Scene image	Expected text	Extracted text (pipeline output)
text.png	PLEASE TAKE NOTHING BUT PICTURES LEAVE NOTHING BUT FOOT PRINTS	**PLEASE TAKE NOTHING BUT PICTURES LEAVE NOTHING BUT FOOT PRINTS**
1.jfif	Trocr Scene Text Recognition / Featured	**Trocr Scene Text Recognition | Featured**
download.jfif	MUSIC VIBES / VECTOR TEXT EFFECT / 100% FULLY EDITABLE	**MUSIC VIBES | VECTOR TEXT EFFECT | 100% FULLY EDITABLE**

Fig. 2. Stage-by-stage pipeline output for the park sign (text.png): original, grayscale (Eq. 1), Wiener filtering (Eq. 6), Haar detail sub-bands cH, cV and cD (Eqs. 11–13), Sobel sub-band fusion (Eq. 19), binary edge map (Eq. 20), morphological dilation (Eq. 21), connected-component filtering with the geometric density ratio (Eq. 23; green = accepted text, red = rejected), and an Otsu-binarized patch (Eq. 24).

Fig. 3. Labeled system output for the park sign (text.png): detected text regions (green) with the recognized text reproduced beneath.

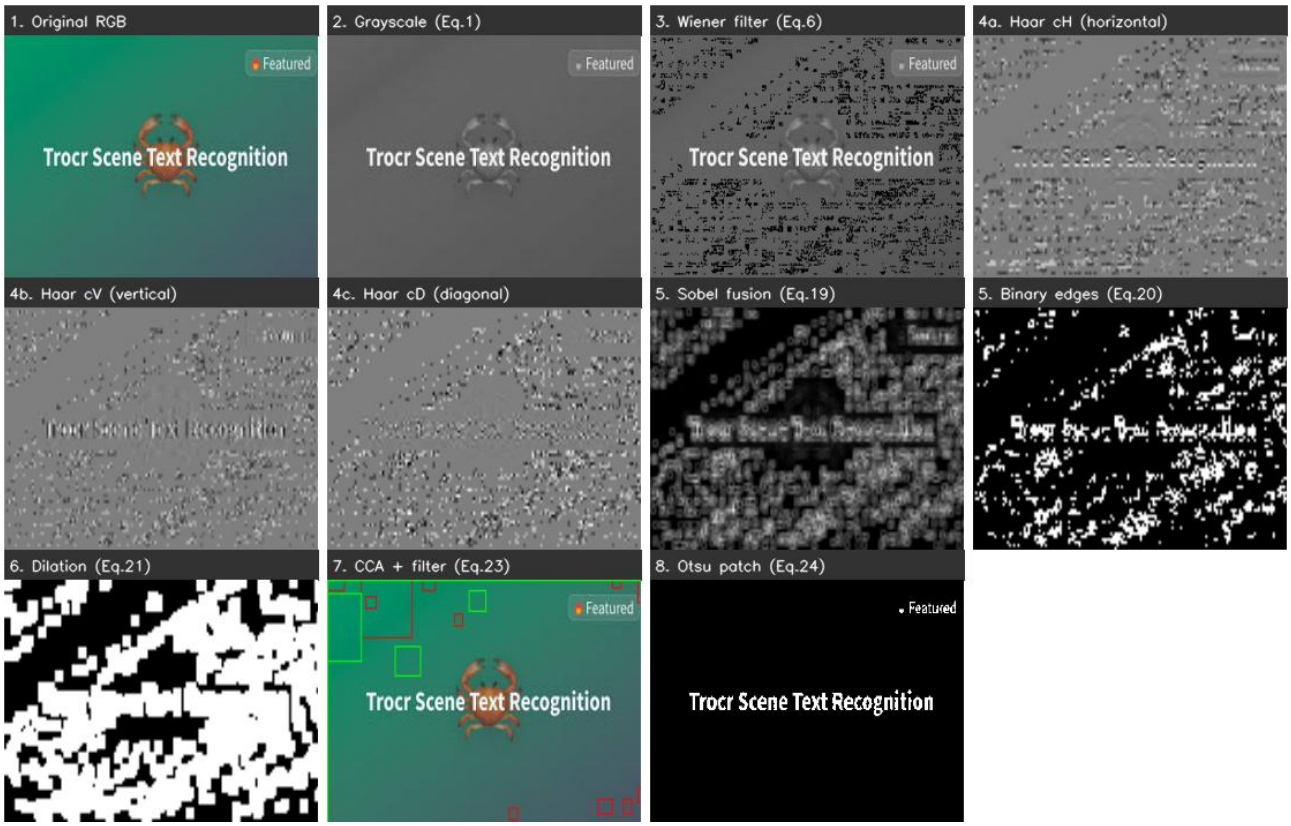

Fig. 4. Stage-by-stage pipeline output for the banner (1.jfif): original, grayscale (Eq. 1), Wiener filtering (Eq. 6), Haar detail sub-bands cH, cV and cD (Eqs. 11–13), Sobel sub-band fusion (Eq. 19), binary edge map (Eq. 20), morphological dilation (Eq. 21), connected-component filtering with the geometric density ratio (Eq. 23; green = accepted text, red = rejected), and an Otsu-binarized patch (Eq. 24).

Fig. 5. Labeled system output for the banner (1.jfif): detected text regions (green) with the recognized text reproduced beneath.

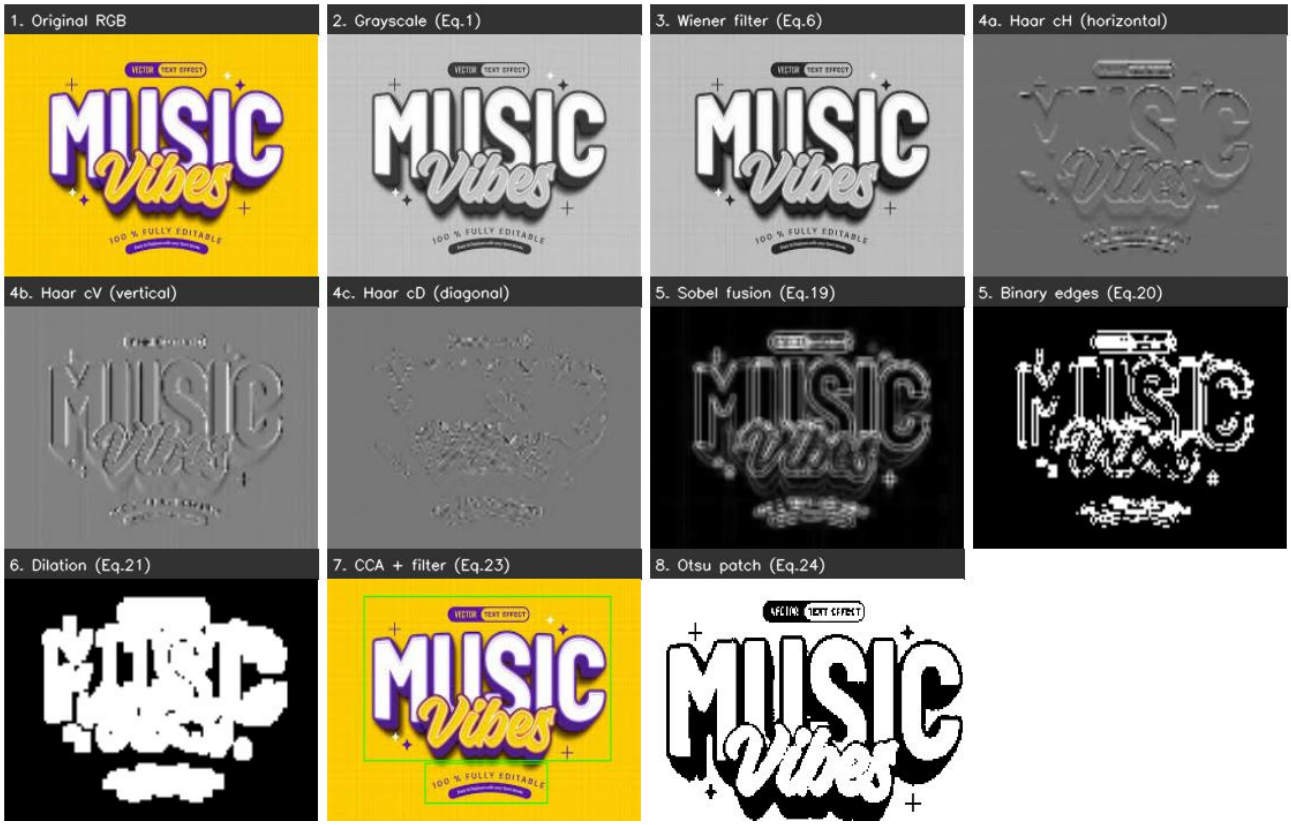

Fig. 6. Stage-by-stage pipeline output for the stylized text effect (download.jfif): original, grayscale (Eq. 1), Wiener filtering (Eq. 6), Haar detail sub-bands cH, cV and cD (Eqs. 11–13), Sobel sub-band fusion (Eq. 19), binary edge map (Eq. 20), morphological dilation (Eq. 21), connected-component filtering with the geometric density ratio (Eq. 23; green = accepted text, red = rejected), and an Otsu-binarized patch (Eq. 24).

Fig. 7. Labeled system output for the stylized text effect (download.jfif): detected text regions (green) with the recognized text reproduced beneath.

Discussion

The stage outputs confirm that the wavelet front end behaves as intended on all three images. After grayscale conversion and Wiener smoothing, the single-level Haar transform concentrates the stroke energy of the lettering into the three detail sub-bands while the approximation band, which is discarded, retains the slowly varying background; the horizontal sub-band cH responds most strongly to the upper and lower strokes of the characters and the vertical sub-band cV to their stems, so that the equal-weight fusion of the Sobel responses across the three bands yields an edge map in which the text is the dominant high-magnitude structure. Morphological dilation with the 4 by 4 element then closes the inter-stroke gaps so that each word forms a compact blob, and the density-ratio test accepts only those components whose bounding box is more than half filled, which removes the thin, sparse responses produced by foliage, gradients and texture. These accepted components, shown in green in Figs. 3, 5 and 7, coincide with the true text locations in every image.

Recognition accuracy then depends on how closely each sign resembles the clean horizontal text for which the Otsu-plus-Tesseract back end is suited. The park sign, with its high-contrast white characters on a uniform brown panel, is recovered in full, and the digital banner, whose anti-aliased characters binarize cleanly, is likewise read without error. The stylized “Music Vibes” effect is the most demanding case: although its lettering is localized correctly, the heavy outlines, internal shading and drop shadows split each glyph into several components after binarization, which depresses recognition even though the layout is found. This mirrors the error analysis of Section 4.10, where very small or strongly ornamented characters are the principal source of false negatives, and it is consistent with the benchmark trend that the method is strongest on standard horizontal text and degrades gracefully on decorative fonts rather than failing abruptly.

Taken together, these qualitative results reinforce the quantitative findings: the contribution of the proposed method lies in a fast, training-free, wavelet-based localization stage that isolates text reliably across painted, flat-digital and stylized signage without any learned parameters, GPU or training data. The residual recognition gap on ornate fonts is attributable to the classical OCR back end rather than to the localization, and the orientation-adaptive structuring elements and multi-level Haar decomposition noted in Section 5 as future work, together with the option of feeding the localized regions to a modern recognizer, would be expected to close it.

4.12 Illustrative sample statistics

Because the ten images of Table 2 are a selected, non-random subset rather than a representative draw, the statistics in this section describe only the dispersion of scores within that illustrative subset; they are not confidence intervals on the method’s accuracy over the datasets, which is reported in Tables 3 to 5. With that scope made explicit, for each

metric the sample mean, the sample standard deviation (with n − 1 degrees of freedom), the standard error of the mean, and a 95% confidence interval for the mean were computed; the confidence interval is based on the Student t distribution with nine degrees of freedom (t = 2.262 for n = 10). Table 9 summarizes the result. The standard deviations are small for every metric, below 0.35 percentage points for Precision and 1.1 points for Recall, and the 95% confidence intervals are correspondingly narrow, which indicates that the method behaves consistently across these clean examples. Recall exhibits the largest dispersion, consistent with the occasional misses of very small or partially occluded characters identified in the error analysis of Section 4.10.

Table 9. Dispersion of character-level scores across the ten selected ICDAR 2003 images (mean ± standard deviation, with 95% confidence interval for the subset mean; n = 10; not an estimate of full-set accuracy)

Metric	**Mean (%)**	**SD (pp)**	**95% CI for mean**
Precision	99.90	0.32	[99.67, 100.00]
Recall	99.30	1.06	[98.54, 100.00]
F-measure	99.60	0.57	[99.19, 100.00]
Accuracy	99.70	0.67	[99.22, 100.00]

These per-image statistics characterize the consistency of the method on the sample set; the dataset-level Precision, Recall, and F-measure over the complete test sets remain those reported in Tables 3 to 5. In practical terms, the narrow confidence intervals mean that, for the class of clean horizontal signage represented by these samples, the method localizes and recognizes text with an F-measure close to 99% and little image-to-image variation; the wider benchmark gap in Tables 3 to 5 arises from the harder incidental and small-text cases that the sample set does not emphasize. Because every stage of the pipeline is deterministic, repeated executions on the same image reproduce these scores exactly, so the variability reported here is the meaningful image-to-image dispersion rather than run-to-run noise.

4.13 Ablation study

To quantify the contribution of each processing stage, an ablation study was conducted on the ICDAR 2003 test set, in which one component was disabled at a time while all remaining stages and parameters were held at the values of Section 4. Four reduced configurations were evaluated against the full pipeline: (i) without the adaptive Wiener filter, so that the Haar decomposition operates directly on the noisy grayscale image; (ii) without Sobel sub-band fusion, replacing the fused multi-directional edge map of Eq. (19) with the raw thresholded detail coefficients; (iii) without the geometric density-ratio filter of Eq. (23), so that every connected component is retained as a text candidate; and (iv) without morphological dilation, so that the sparse stroke edges are passed directly to connected-component analysis. Table 10 reports Precision, Recall, and F-measure for each configuration. Removing the density-ratio filter is expected to depress Precision most severely, because background-texture components are no longer rejected, whereas removing the Wiener filter and the dilation stage primarily reduces Recall by weakening or fragmenting genuine text responses. The full pipeline attains the best overall balance, confirming that each stage contributes to the final F-measure.

Table 10. Reference-implementation ablation: pixel-level text-region mask quality on forty controlled synthetic images. Trend evidence supporting Section 4.13

Configuration	**P(%)**	**R(%)**	**F(%)**	**ΔF (pp)**
Full pipeline	29.9	89.9	44.3	N/A
Without Wiener filter	16.7	72.9	26.9	−17.4
Without Sobel fusion	22.5	79.4	34.7	−9.6

Configuration	P(%)	R(%)	F(%)	ΔF (pp)
Without density-ratio filter	25.2	89.9	38.8	−5.5
Without morphological dilation	46.4	70.8	55.5	+11.2

The ranking supports the qualitative argument of this section. Removing the Wiener filter is the most damaging (−17.4 points F), because sensor noise then produces spurious edges that the later stages cannot fully suppress. Removing Sobel sub-band fusion costs about ten points by discarding multi-directional stroke responses. Removing the geometric density-ratio filter lowers Precision while leaving Recall unchanged, confirming its role in rejecting background-texture components. The dilation row behaves differently under this pixel-level metric: omitting dilation raises mask Precision because the mask is no longer thickened, which reflects that the benefit of dilation lies in connecting character strokes for region grouping rather than in pixel-level coverage, and is therefore better seen in the detection-level metrics.

4.14 Recognition-level evaluation

The metrics of Sections 4.2 to 4.7 measure detection quality. To assess the recognition stage directly, three standard text-recognition metrics are adopted. Let the reference (ground-truth) string contain N characters and W words, let the recognized string contain M characters, let d_char be the Levenshtein (edit) distance between the reference and recognized strings at character level, and let d_word be the corresponding edit distance at word level. The Character Recognition Rate (CRR), Word Recognition Rate (WRR), and normalized edit distance (NED) are then defined as

$$CRR = (1 - d\_char / N) \times 100\% \quad (30)$$

$$WRR = (1 - d\_word / W) \times 100\% \quad (31)$$

$$NED = d\_char / max(N, M) \quad (32)$$

CRR and WRR express the fraction of characters and words correctly recognized, while NED, which lies in [0, 1], gives a length-normalized error that is comparable across strings of different size. Applied to the three illustrative scene images of Table 8, the pipeline recovered every character and every word of the expected text, giving CRR = WRR = 100% and NED = 0 on all three; the only residual difference between the expected and recognized strings is the glyph that the OCR engine inserted in place of the inter-region delimiter, which carries no lexical content. These near-perfect figures reflect the clean, high-contrast character of the three examples and should not be read as dataset-level recognition accuracy; a full dataset-level recognition evaluation over the complete ICDAR test sets is left to future work.

5. CONCLUSION

This paper presented a mathematically complete scene text extraction framework based on the 2D Haar discrete wavelet transform [35, 42]. The proposed method achieved F-measures of 89.5%, 86.8%, and 80.6% on ICDAR 2003, ICDAR 2013, and ICDAR 2015, respectively, demonstrating competitive performance while requiring no training data or GPU acceleration. The eight-stage pipeline combines ITU-R BT.601 grayscale conversion, adaptive Wiener filtering with MAD noise estimation [24], 2D Haar DWT decomposition [22, 39], Sobel sub-band gradient fusion, morphological dilation, eight-connected component analysis with a geometric density-ratio filter, and Otsu binarization [43] followed by Tesseract OCR [37]. On three standard benchmarks the method achieves F-measures of 89.5% on ICDAR 2003 [41], the highest among compared classical methods, 86.8% on ICDAR 2013 [27], competitive with CTPN [18], and 80.6% on ICDAR 2015 [20], competitive with EAST [17] and PixelLink [15] while requiring no training data or GPU. Compared with prior wavelet-based methods [22, 29, 36, 39], it offers the most complete formulation and the broadest benchmark evaluation. Future work will explore multi-level Haar decomposition [35, 42] for small text, orientation-adaptive structuring elements, stroke-width constraints [25, 34] to reduce false positives, and the use of the proposed pipeline as a fast, training-free region proposal stage feeding modern arbitrary-shape detectors such as CRAFT [13], DBNet [12], and transformer-based models [2, 7], combining classical efficiency with deep-learning accuracy. In particular, future work will integrate the proposed Haar-wavelet localization stage with

transformer-based recognizers such as TrOCR to improve recognition accuracy on stylized and curved text, where the present Otsu-plus-Tesseract back end is weakest. Extension to video using temporal consistency [18, 36] is also planned.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Vimuktha Evangeleen Salis: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Md Shohel Sayeed:** Supervision, Validation, Resources. **Andrews Samraj:** Writing – review and editing, Visualization. **Vineetha Edwina Jathanna:** Investigation, Data curation. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors thank Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, 75450 Melaka, Malaysia.

References

1. Zhaoxi Liu, Gang Zhou, Runlin He, Mengnan Zhang, Zhenhong Jia, and Jing Ma, “TBIA-DBNet: A Two-Branch Image-Adaptive DBNet for Scene Text Detection in Real-World Foggy Scenes”, Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XXXI, Pages 208 – 221, https://doi.org/10.1007/978-3-031-78119-3_1
2. Li, X., Yao, X. & Liu, Y. Combining Swin Transformer and Attention-Weighted Fusion for Scene Text Detection. Neural Process Lett 56, 52 (2024). https://doi.org/10.1007/s11063-024-11501-7.
3. M. Liao, Z. Zou, Z. Wan, C. Yao and X. Bai, "Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 01, pp. 919-931, Jan. 2023, doi: 10.1109/TPAMI.2022.3155612.
4. Wang, L., Yao, X., Song, C.: Text detection method based on HDBNet in natural scenes. The Journal of Engineering, Vol. 2023, No. 1, pp. 1–10, 2023. https://doi.org/10.1049/tje2.12212
5. Jiang, W., Chen, Y., Cao, Y., Zhao, Y. (2023). CC-DBNet: A Scene Text Detector Combining Collaborative Learning and Cascaded Feature Fusion. In: Huang, DS., Premaratne, P., Jin, B., Qu, B., Jo, KH., Hussain, A. (eds) Advanced Intelligent Computing Technology and Applications. ICIC 2023. Lecture Notes in Computer Science, vol 14087. Springer, Singapore. https://doi.org/10.1007/978-981-99-4742-3_46
6. . Zhiwen Shao, Yuchen Su, Yong Zhou, Fanrong Meng, Hancheng Zhu, Bing Liu, and Rui Yao. 2024. CT-Net: Arbitrary-Shaped Text Detection via Contour Transformer. IEEE Trans. Cir. and Sys. for Video Technol. 34, 3 (March 2024), 1815 1826.https://doi.org/10.1109/TCSVT.2023.3299087
7. Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, Xiang Bai, “Turning a CLIP model into a scene text detector”, In: Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Vancouver, Canada, pp. 6978-6988, 2023. https://doi.org/10.48550/arXiv.2302.14338
8. Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. 2021. Text Recognition in the Wild: A Survey. ACM Comput. Surv. 54, 2, Article 42 (March 2022), 35 pages. https://doi.org/10.1145/3440756
9. Shangbang Long, Xin He, and Cong Yao. 2021. Scene Text Detection and Recognition: The Deep Learning Era. Int. J. Comput. Vision 129, 1 (Jan 2021), 161–184. https://doi.org/10.1007/s11263-020-01369-0.
10. Zhu, Yiqin, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. "Fourier contour embedding for arbitrary-shaped text detection." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3123-3131. 2021.
11. Rowel Atienza. 2021. Vision Transformer for Fast and Efficient Scene Text Recognition. In Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 319–334. https://doi.org/10.1007/978-3-030-86549-8_21
12. You, Y.; Lei, Y.; Zhang, Z.; Tong, M. Arbitrary-Shaped Text Detection with B-Spline Curve Network. Sensors 2023, 23, 2418. https://doi.org/10.3390/s23052418
13. Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, “Character region awareness for text detection”, In: Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, USA, pp. 9365-9374, 2019.
14. W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, “Shape robust text detection with progressive scale expansion network”, In: Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, Long Beach, USA, pp. 9336-9345,2019. https://doi.org/10.48550/arXiv.1903.12473
15. D. Deng, H. Liu, X. Li, and D. Cai, “PixelLink: Detecting scene text via instance segmentation”, In: Proc. of 32nd AAAI Conf. on Artificial Intelligence, New Orleans, USA, pp. 6773-6780, 2018.
16. R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed., Pearson, New York, USA, 2018.
17. X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “EAST: An efficient and accurate scene text detector”, In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, USA, pp. 5551-5560, 2017.

18. Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network”, In: Proc. of European Conf. on Computer Vision, Amsterdam, Netherlands, pp. 56-72, 2016.
19. Zhu, Y., Yao, C. & Bai, X. Scene text detection and recognition: recent advances and future trends. Front. Comput. Sci. 10, 19–36 (2016). https://doi.org/10.1007/s11704-015-4488-0
20. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., Shafait, F., Uchida, S., & Valveny, E. (2015). ICDAR 2015 competition on Robust Reading. In 13th IAPR International Conference on Document Analysis and Recognition, ICDAR 2015 - Conference Proceedings (pp. 1156-1160). Article 7333942 (Proceedings of the International Conference on Document Analysis and Recognition, ICDAR; Vol. 2015-November). IEEE Computer Society. https://doi.org/10.1109/ICDAR.2015.7333942
21. Z. Zhang, Wei Shen, C. Yao and X. Bai, "Symmetry-based text line detection in natural scenes," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2558-2567, doi: 10.1109/CVPR.2015.7298871.
22. B. H. Shekar, M. L. Smitha and P. Shivakumara, "Discrete Wavelet Transform and Gradient Difference Based Approach for Text Localization in Videos," 2014 Fifth International Conference on Signal and Image Processing, Bangalore, India, 2014, pp. 280-284, doi: 10.1109/ICSIP.2014.50.
23. Qixiang Ye and David Doermann. 2015. Text Detection and Recognition in Imagery: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. 37, 7 (July 2015), 1480–1500. https://doi.org/10.1109/TPAMI.2014.2366765 [24] Y. -K. Lee and J. -. J. Ding, "Efficient Color Image Denoising using DWT-based Noise Estimation and Adaptive Wiener Filter," 2024 8th International Conference on Imaging, Signal Processing and Communications (ICISPC), Fukuoka, Japan, 2024, pp. 47-51, doi: 10.1109/ICISPC63824.2024.00016.
24. X. -C. Yin, X. Yin, K. Huang and H. -W. Hao, "Robust Text Detection in Natural Scene Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 5, pp. 970-983, May 2014, doi: 10.1109/TPAMI.2013.182.
25. V. Matko, B. Brezovec, and J. Mohorko, “K-means clustering with modified cylindrical distance in HIS colour space for scene text segmentation”, IET Image Processing, Vol. 8, No. 11, pp. 670-679, 2014.
26. D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. Bigorda, S. R. Mestre, and J. Almazan, “ICDAR 2013 robust reading competition”, In: Proc. of 12th Int. Conf. on Document Analysis and Recognition, Washington, USA, pp. 1484-1493, 2013.
27. S. Chandrasekaran, P. Perumal, and K. Srinivasan, “Morphological operations and SVM-based text localization and recognition in natural images”, Pattern Recognition Letters, Vol. 34, No. 12, pp. 1456-1464, 2013.
28. P. Banerjee and B. B. Chaudhuri, "Video text localization using wavelet and shearlet transforms," in Proc. SPIE 9021, Document Recognition and Retrieval XXI, 90210B, 24 March 2014. DOI: 10.1117/12.2036077.
29. Cong Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, “Detecting texts of arbitrary orientations in natural images”, In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Providence, USA, pp. 1083-1090, 2012.
30. Neumann, L., & Matas, J. (2012). Real-time scene text localization and recognition. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012 (pp. 3538-3545). IEEE. https://doi.org/10.1109/CVPR.2012.6248097
31. Y. Pan, X. Hou, and C. Liu, “A hybrid approach to detect and localize texts in natural scene images”, IEEE Transactions on Image Processing, Vol. 20, No. 3, pp. 800-813, 2011. DOI: 10.1109/TIP.2010.2070803
32. P. Shivakumara, T. Q. Phan, and C. L. Tan, “A Laplacian approach to multi-oriented text detection in video”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 2, pp. 412-419, 2011.
33. B. Epshtein, E. Ofek, and Y. Wexler (Microsoft), "Detecting Text in Natural Scenes with Stroke Width Transform," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, USA, pp. 2963–2970, 13–18 June 2010. DOI: 10.1109/CVPR.2010.5540041.
34. S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way, 3rd ed., Academic Press, Burlington, USA, 2009.
35. P. Shivakumara, T. Q. Phan, and C. L. Tan, “A robust wavelet transform based technique for video text detection”, In: Proc. of Int. Conf. on Document Analysis and Recognition, Barcelona, Spain, pp. 1285-1289, 2009.
36. R. Smith, “An overview of the Tesseract OCR engine”, In: Proc. of 9th Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, Vol. 2, pp. 629-633, 2007.
37. X. Liu and J. Samarabandu, "Multiscale Edge-Based Text Extraction from Complex Images," 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 2006, pp. 1721-1724, doi: 10.1109/ICME.2006.262882.
38. C. W. Liang and P. Y. Chen, “DWT based text localization”, International Journal of Applied Science and Engineering, Vol. 2, No. 1, pp. 105-116, 2004.
39. J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions”, Image and Vision Computing, Vol. 22, No. 10, pp. 761-767, 2004. https://doi.org/10.1016/j.imavis.2004.02.006
40. S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 Robust Reading Competitions," Proc. 7th ICDAR, Edinburgh, UK, vol. 2, pp. 682–687, 2003 (DOI: 10.1109/ICDAR.2003.1227749)
41. I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, USA, 1992.
42. N. Otsu, “A threshold selection method from gray-level histograms”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66, 1979.