Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.9994

Segmentasi Teks Arab Pegon Menggunakan *Histogram Segmentation*

Muhammad Fikri Hidayattullah [1], Sharfina Febbi Handayani [2], Yustia Hapsari [3], M. Ibrahim Hanif [4]
[1,2,4]Prodi D4 Teknik Informatika Universitas Harkat Negeri, Jl. Mataram No. 09 Pesurungan Lor, Kota Tegal, 52147, Indonesia
[3] Prodi Bisnis Digital Universitas Pancasakti Tegal, Jl. Halmahera KM.1 Mintaragen,Kota Tegal, 52121, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-12-19
Revised 2025-12-23
Accepted 2025-12-25

Abstract *– The Pegon script is a cultural heritage of the Indonesian archipelago with high historical value, particularly in Islamic literature in Java, Sunda, and Madura. Efforts to digitize Pegon manuscripts face a number of significant obstacles. The complexity of letterforms, letter-joining patterns, and irregularities in the layout of ancient manuscripts make text extraction much more challenging than modern printed texts. One crucial step in this process is text segmentation. The accuracy of character or word separation will significantly determine the success of the next steps, namely Optical Character Recognition (OCR) and automatic transliteration. Histogram segmentation is known to detect and separate text objects from the background by utilizing pixel intensity distribution. This method is widely recognized in digital image segmentation. Its simplicity and lack of training make it suitable as a first step towards transliterating Arabic-Pegon script. Experimental results show that the model is capable of performing line segmentation well. Meanwhile, the accuracy level of word segmentation was 70% with an Over-Segmentation Rate (OSR) error rate of 0.20 and an Under-Segmentation Rate (USR) of 0.13. The histogram segmentation method proved to be lightweight, quite efficient, and did not require data training unlike the deep learning approach. This research has a significant contribution to the preservation of the Arabic-Pegon script as a treasure of knowledge belonging to the archipelago.*

Keywords: *Histogram Segmentation, Image Processing, Manuscript Digitization,OCR, Pegon Script*

Corresponding Author:

Muhammad Fikri Hidayattullah

Email: fikri@harkatnegeri.ac.id

Abstrak *– Aksara Pegon merupakan salah satu warisan budaya Nusantara yang memiliki nilai sejarah tinggi, terutama dalam literatur Islam di Jawa, Sunda dan Madura. Upaya digitalisasi naskah Pegon menghadapi sejumlah kendala yang tidak sederhana. Kompleksitas bentuk huruf, pola penyambungan antarhuruf, serta ketidakteraturan tata letak tulisan pada manuskrip kuno menjadikan proses ekstraksi teks jauh lebih menantang dibandingkan teks cetak modern. Salah satu tahapan yang sangat penting dalam proses tersebut adalah segmentasi teks. Ketepatan pemisahan karakter atau kata akan sangat menentukan keberhasilan tahap berikutnya, yaitu Optical Character Recognition (OCR) dan transliterasi otomatis. Histogram segmentation dikenal mampu mendeteksi dan memisahkan objek teks dari latar belakang dengan memanfaatkan distribusi intensitas piksel. Metode ini sudah dikenal luas dalam segmentasi citra digital. Cara kerjanya yang sederhana tanpa memerlukan pelatihan, membuat metode ini cocok digunakan sebagai langkah awal menuju tahap transliterasi aksara Arab-Pegon. Hasil eksperimen menunjukkan bahwa model mampu melakukan segmentasi baris dengan baik. Sedangkan tingkat akurasi segmentasi kata sebesar 70% dengan error rate Over-Segmentation Rate (OSR) sebesar 0.20 dan Under-Segmentation Rate (USR) sebesar 0.13. Metode histogram segmentation terbukti ringan, cukup efisien, dan tidak memerlukan pelatihan data seperti pada pendekatan deep learning. Penelitian ini memiliki kontribusi besar terhadap pelestarian aksara Arab-Pegon sebagai khazanah ilmu pengetahuan yang dimiliki oleh nusantara.*

Kata Kunci*: Aksara Pegon, Digitalisasi Naskah, Histogram Segmentation, OCR, Pengolahan Citra*

I. PENDAHULUAN

Aksara Pegon merupakan salah satu bentuk warisan budaya Nusantara yang memiliki nilai sejarah tinggi, terutama dalam literatur Islam di Jawa, Sunda dan Madura [1][2]. Aksara ini menggunakan huruf Arab dengan tambahan beberapa tanda diakritik untuk menyesuaikan dengan fonologi bahasa Jawa dan Melayu [3]. Keberadaan manuskrip dan naskah kuno beraksara Pegon mencerminkan perjalanan intelektual serta perkembangan agama dan budaya di Indonesia sejak abad ke-14. Sayangnya, banyak dari naskah-naskah tersebut masih belum terdigitalisasi dengan baik [4][5][6], sehingga aksesibilitasnya terbatas dan rentan mengalami degradasi fisik seiring berjalannya waktu [7]. Digitalisasi naskah Pegon menjadi upaya strategis dalam melestarikan warisan budaya ini agar dapat diakses oleh generasi mendatang.

Salah satu tantangan utama dalam digitalisasi dan transliterasi naskah Pegon adalah proses segmentasi teks, yang merupakan tahap awal dalam pengolahan citra sebelum dilakukan transliterasi otomatis [8]. Aksara Arab-Pegon memiliki karakteristik yang khas dibandingkan dengan aksara Arab standar. Perbedaan tersebut dapat dilihat dari segi bentuk huruf, pola penyambungan, maupun tata letak penulisannya dalam naskah yang

berbeda dengan aksara Arab standar. Setiap huruf dapat mengalami perubahan bentuk bergantung pada posisinya di awal, tengah, atau akhir kata, serta sering kali saling menyambung secara kompleks. Kondisi ini menjadi semakin menantang ketika teks Pegon ditulis pada naskah kuno yang tidak selalu memiliki keteraturan spasi dan konsistensi bentuk tulisan. Oleh karena itu, proses segmentasi yang mampu memisahkan unit teks secara tepat menjadi tahapan krusial sebelum dilakukan pengenalan karakter atau transliterasi lebih lanjut.

Di bidang pengolahan citra digital, pendekatan berbasis histogram telah lama dimanfaatkan untuk memisahkan objek teks dari latar belakang dengan memanfaatkan distribusi intensitas piksel. Teknik *histogram segmentation* bekerja dengan menganalisis perubahan kepadatan piksel pada arah tertentu untuk mengidentifikasi batas-batas teks. Beberapa penelitian sebelumnya melaporkan bahwa pendekatan ini cukup efektif dalam meningkatkan performa *Optical Character Recognition* (OCR), khususnya pada teks Arab bercetak dengan kualitas citra rendah [9]. Meskipun demikian, sebagian besar studi tersebut masih berfokus pada teks Arab standar [10][11], sementara penerapannya pada aksara Arab Pegon, yang memiliki karakteristik visual dan linguistik berbeda, masih relatif terbatas.

Permasalahan segmentasi menjadi lebih kompleks ketika berhadapan dengan manuskrip Pegon yang mengalami degradasi fisik. Di berbagai sumber, terdapat beberapa faktor semisal tinta yang memudar, bercak pada media tulis, dan tumpang tindih antar goresan huruf sering kali menyebabkan bahwa batas antar kata atau karakter tidak dapat dikenali secara tegas. Faktor-faktor ini berpotensi mengurangi akurasi segmentasi jika tidak diatasi. Berdasarkan pemikiran di atas, maka penelitian ini difokuskan pada tahap awal dalam proses translaterasi arab pegon yaitu pengembangan metode segmentasi berbasis histogram. Dengan harapan bahwa pendekatan ini mampu memisahkan teks dari latar belakang secara lebih konsisten dan mendukung proses translaterasi dan OCR pada tahap berikutnya.

Pengembangan sistem transliterasi otomatis untuk aksara Pegon melibatkan banyak elemen yang memiliki implikasi luas bagi pelestarian budaya serta pengembangan plain studi philologi, sejarah, dan pendidikan Islam. Banyak naskah karya ulama terdahulu ditulis menggunakan aksara Pegon dan memuat nilai-nilai keilmuan yang penting, namun belum sepenuhnya dapat diakses oleh masyarakat luas. Dengan dukungan teknologi segmentasi dan digitalisasi teks, naskah-naskah tersebut dapat dipelajari secara lebih sistematis dan efisien. Selain itu, sistem transliterasi otomatis juga berpotensi menjadi alat bantu akademik bagi peneliti dan pelajar yang belum memiliki kompetensi khusus dalam membaca aksara Pegon.

Seiring dengan perkembangan teknologi pengenalan teks, berbagai pendekatan telah dikembangkan untuk menangani teks Arab, mulai dari metode berbasis *Connected Component Analysis* (CCA) hingga pendekatan berbasis pembelajaran mendalam seperti *Convolutional Neural Networks* (CNN) [12][13]. Meskipun pendekatan berbasis deep learning menunjukkan kinerja yang menjanjikan, metode segmentasi klasik seperti histogram segmentation tetap relevan sebagai fondasi awal, khususnya pada penelitian yang berfokus pada eksplorasi karakteristik data dan keterbatasan sumber daya. Meskipun metode berbasis CNN menunjukkan hasil yang lebih akurat dalam segmentasi karakter Arab [14][15], model ini memerlukan dataset pelatihan yang besar dan daya komputasi tinggi. Sebaliknya, *histogram segmentation* menawarkan pendekatan yang lebih ringan dan efisien dalam mengolah teks Arab Pegon tanpa memerlukan pelatihan dataset yang kompleks [16]. Oleh karena itu, penelitian ini mengusulkan penerapan *histogram segmentation* sebagai solusi yang lebih praktis dan cepat dalam proses segmentasi teks Arab Pegon.

Di sisi lain, upaya digitalisasi teks Pegon memiliki peran strategis dalam konteks pelestarian warisan budaya Nusantara. Perkembangan teknologi yang semakin cepat secara tidak langsung turut memengaruhi pola interaksi generasi muda terhadap aksara-aksara tradisional, yang kini semakin jarang digunakan dalam kehidupan sehari-hari [17]. Kondisi tersebut berpotensi menyebabkan berkurangnya akses dan pemahaman terhadap literatur Pegon yang selama ini menjadi medium penting dalam transmisi pengetahuan keislaman dan kebudayaan lokal.

Kehadiran sistem transliterasi otomatis berbasis teknologi digital membuka peluang baru dalam menjembatani kesenjangan tersebut. Dengan melestarikan transliterasi ke aksara Latin, teks Pegon dapat diakses oleh banyak pihak dan lebih mudah dipelajari dan diakses oleh kalangan umum tanpa menghilangkan kualitas historis tersendiri dari naskah Pegon itu sendiri. Dalam jangka panjang, pendekatan ini berbasis dari pengolahan data dan literasi naskah yang merupakan bagian dari upaya nasional yang lebih besar dalam melestarikan naskah-naskah kuno. Dengan demikian, pendekatan versifier dalam literasi dan digitalisasi Pegon bukan hanya menangani permasalahan data dan informasi yang bersifat teknis, namun juga kultural.

Secara keseluruhan, penelitian ini bertujuan untuk mengembangkan metode segmentasi teks Arab Pegon menggunakan *histogram segmentation* sebagai langkah awal dalam pengembangan sistem transliterasi otomatis. Melalui pendekatan ini, diharapkan dapat tercipta solusi yang lebih akurat dan efisien dalam pengolahan teks Pegon, serta mendukung digitalisasi dan preservasi warisan literasi Nusantara.

II. METODE

Penelitian ini dimulai dari pengumpulan *dataset* aksara Arab-Pegon, pembuatan model untuk segmentasi dan evaluasi terhadap model yang telah dikembangkan.

A. Pembuatan Dataset

Dataset tulisan tangan Arab-Pegon atau pun bentuk lainnya belum tersedia. Oleh karena itu langkah pertama yang dilakukan di dalam penelitian ini adalah melakukan pembuatan *dataset* Arab-Pegon. Pembuatan *dataset* Arab-Pegon bukanlah hal yang mudah. Butuh kerja panjang yang melibatkan banyak pihak. Tim peneliti meminta ke para relawan untuk menulis ulang menggunakan tulisan tangan huruf Pegon seperti pada Gambar 1.

Gambar 1. Formulir Pembuatan *Dataset* Aksara Arab-Pegon

Jumlah huruf Arab-Pegon yang terdapat pada formulir pembuatan *dataset* sebanyak 141 huruf yang diisi oleh 100 relawan yang berbeda. Data yang digunakan pada penelitian ini pada awalnya masih tersedia dalam bentuk media cetak, sehingga tahap pertama dilakukan dengan melakukan proses pemindaian menggunakan perangkat *scanner* untuk mengonversi dokumen fisik menjadi citra digital. Selanjutnya, citra hasil pemindaian distandarisasi ke dalam format warna RGB guna memastikan keseragaman representasi warna sebelum memasuki tahap pemrosesan lanjutan. Setelah itu, citra dikonversi ke dalam skala abu-abu (*grayscale*) agar lebih sesuai untuk kebutuhan analisis citra dan pengolahan berbasis intensitas piksel.

Pada tahap pra-pemrosesan, diterapkan operasi dilasi morfologi menggunakan kernel berukuran 3×3 dengan tujuan mempertegas sekaligus memperluas area berwarna putih pada citra. Untuk menyesuaikan kebutuhan pemisahan objek dan latar belakang, dilakukan operasi inversi warna sehingga area berwarna hitam berubah menjadi putih dan sebaliknya. Proses ini diikuti dengan penerapan adaptive thresholding untuk mengubah citra menjadi citra biner hitam-putih, sehingga konten utama dapat dipisahkan secara lebih jelas dari latar belakang.

Agar bentuk karakter dan struktur tabel semakin menonjol, dilasi kembali diaplikasikan untuk mempertebal objek putih sekaligus menutup celah-celah kecil antar huruf. Selanjutnya, median blur digunakan untuk mereduksi noise serta menyamarkan detail teks sehingga baris-baris pada tabel cenderung menyatu. Hasil

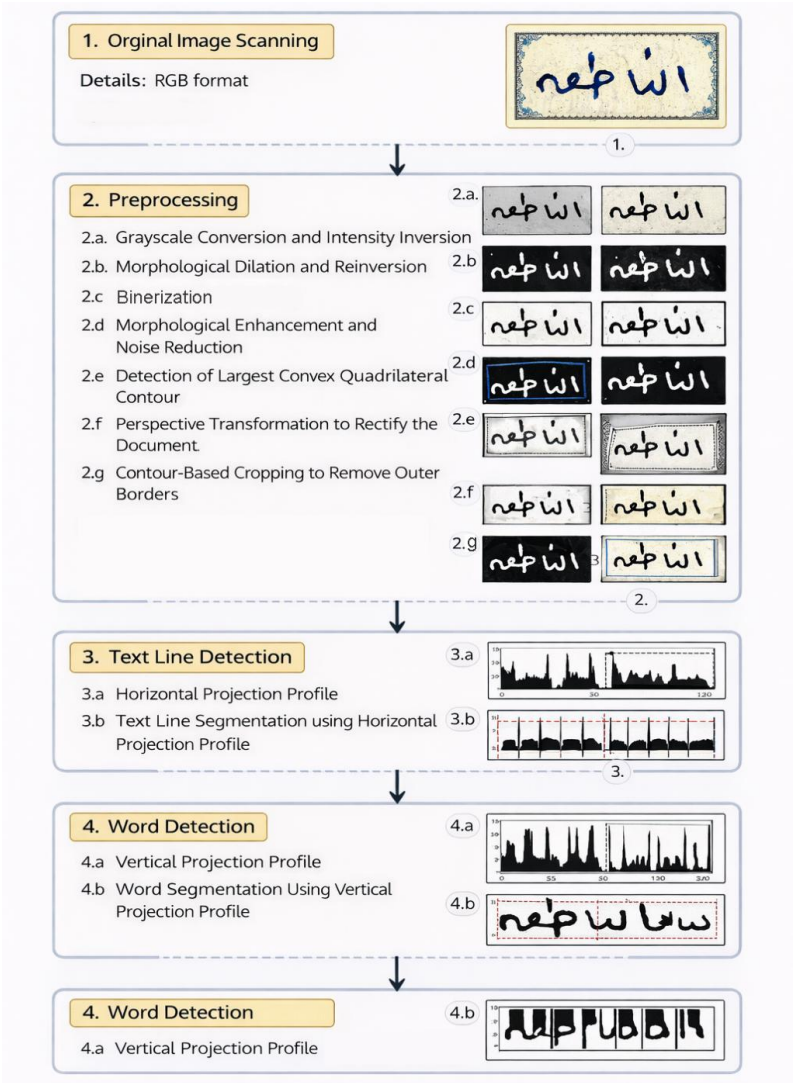
dari proses tersebut kemudian diperkuat kembali melalui dilasi lanjutan untuk menggabungkan area-area putih menjadi satu blok solid yang merepresentasikan posisi tabel secara utuh.

Deteksi kontur berbentuk persegi panjang kemudian diterapkan untuk mengidentifikasi batas tabel, diikuti dengan penggambaran bounding box sebagai sarana evaluasi visual terhadap tingkat presisi deteksi. Berdasarkan hasil tersebut, dilakukan proses pemotongan citra sehingga hanya bagian tabel yang dipertahankan. Untuk mengatasi kemungkinan kemiringan akibat proses pemindaian, diterapkan transformasi perspektif guna mengoreksi orientasi tabel agar tampak tegak lurus dan proporsional.

Tahap selanjutnya adalah pembagian struktur tabel ke dalam grid yang terdiri dari 9 baris dan 7 kolom. Dimensi tinggi dan lebar setiap sel dihitung secara proporsional berdasarkan ukuran total citra tabel. Proses pemotongan kemudian dilakukan secara iteratif dengan menghitung koordinat batas atas, bawah, kiri, dan kanan untuk setiap sel pada grid. Pada setiap sel, area tepi yang tidak relevan dibuang untuk mempertahankan inti objek, kemudian citra diubah ukurannya menjadi dimensi tetap sebesar 75×75 piksel. Seluruh citra hasil ekstraksi dari setiap sel selanjutnya disimpan secara berurutan ke dalam direktori yang telah disiapkan sebagai dataset akhir untuk tahap pemrosesan berikutnya.

B. Pembuatan Model

Model segmentasi dibangun menggunakan pendekatan *histogram-based segmentation* yang memanfaatkan distribusi kepadatan piksel untuk memisahkan unit-unit teks pada aksara Arab-Pegon. Tujuan utama dari tahap ini adalah memperoleh segmentasi huruf dan kata secara akurat sebagai fondasi awal sebelum proses transliterasi dari aksara Arab-Pegon ke teks Latin. Alur lengkap tahapan segmentasi ditunjukkan pada Gambar 2.



Gambar 2. Metode *Histogram Segmentation*

Proses segmentasi diawali dengan akuisisi citra naskah Pegon dalam format RGB melalui proses pemindaian atau dokumentasi digital. Citra asli kemudian diproses pada tahap pra-pemrosesan untuk meningkatkan kualitas visual dan memperjelas struktur teks. Tahapan pra-pemrosesan meliputi konversi citra ke skala abu-abu dan pembalikan intensitas guna menonjolkan perbedaan antara latar belakang dan objek teks. Pada tahap selanjutnya, diterapkan operasi dilasi morfologis yang diikuti dengan pembalikan ulang intensitas citra untuk menegaskan struktur garis serta elemen visual yang bersifat dominan. Langkah ini bertujuan untuk memperjelas perbedaan antara area teks dan latar belakang. Setelah itu, proses binarisasi digunakan untuk mengonversi citra menjadi bentuk biner dengan tingkat kontras yang lebih baik. Untuk menyempurnakan hasil binarisasi, dilakukan operasi morfologis lanjutan guna mereduksi *noise* dan menyambungkan bagian teks yang terputus, sehingga struktur teks menjadi lebih utuh dan siap diproses pada tahap berikutnya.

Untuk memastikan bahwa proses pengolahan citra difokuskan pada area isi dokumen, dilakukan pencarian kontur berbentuk segi empat dengan luas terbesar yang diasumsikan sebagai batas luar halaman naskah. Kontur ini digunakan sebagai acuan untuk mengidentifikasi area relevan, sehingga bagian di luar halaman, seperti border atau latar belakang non-teks, dapat dieliminasi dari proses selanjutnya. Berdasarkan kontur tersebut, diterapkan transformasi perspektif guna merapikan orientasi dokumen sehingga posisi teks menjadi tegak lurus. Selanjutnya, dilakukan pemotongan citra berbasis kontur untuk menghilangkan *border* dan area luar dokumen, sehingga hanya area teks Pegon yang tersisa untuk tahap berikutnya.

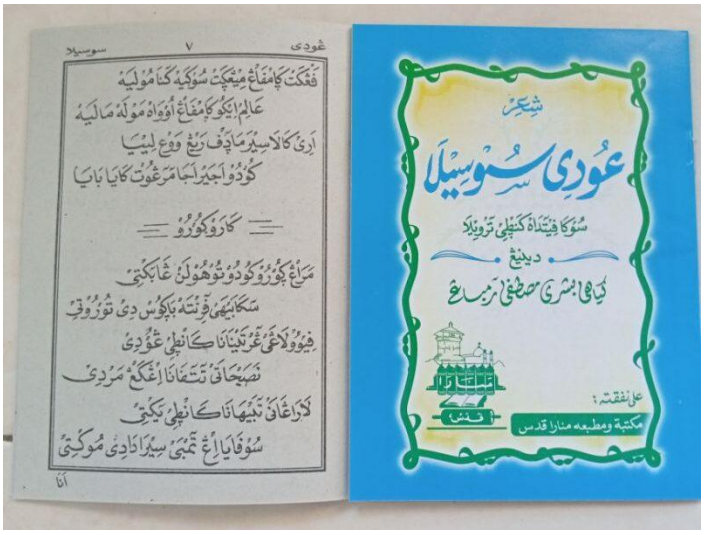
Tahap berikutnya adalah deteksi baris teks (*text line detection*) menggunakan histogram proyeksi horizontal. Histogram ini dihitung berdasarkan jumlah piksel *foreground* pada setiap baris citra. Area dengan kepadatan piksel tinggi diidentifikasi sebagai baris teks, sedangkan area dengan kepadatan rendah dianggap sebagai pemisah antarbaris. Berdasarkan analisis histogram horizontal tersebut, citra dipisahkan menjadi beberapa segmen baris teks secara otomatis.

Setelah baris teks berhasil diperoleh, dilakukan deteksi kata (*word detection*) pada setiap baris menggunakan histogram proyeksi vertikal. Histogram ini merepresentasikan distribusi piksel foreground secara vertikal, sehingga celah antar kata dapat diidentifikasi melalui nilai minimum lokal pada histogram. Titik-titik minimum lokal yang diidentifikasi pada histogram digunakan sebagai penanda batas segmentasi untuk memisahkan kata-kata dalam satu baris teks. Melalui mekanisme ini, setiap baris teks dapat dipilah menjadi unit kata secara otomatis berdasarkan distribusi kepadatan piksel.

Sebagai hasil akhir, tahap segmentasi menghasilkan citra teks Pegon yang telah terpisah pada tingkat kata dan huruf. Pendekatan histogram-based segmentation dipilih karena bersifat deterministik dan tidak memerlukan data latih, sehingga relatif stabil ketika diterapkan pada dataset terbatas. Selain itu, metode ini dinilai mampu menangani karakteristik aksara Arab-Pegon yang memiliki struktur huruf menyambung serta variasi bentuk yang tinggi. Segmentasi yang dihasilkan pada tahap ini menjadi masukan utama untuk proses pengenalan karakter dan transliterasi ke dalam teks Latin pada tahap penelitian selanjutnya.

III. HASIL DAN PEMBAHASAN

Model segmentasi yang sudah selesai dikembangkan akan dicoba secara langsung pada salah satu halaman naskah kitab *Ngudi Susilo*.

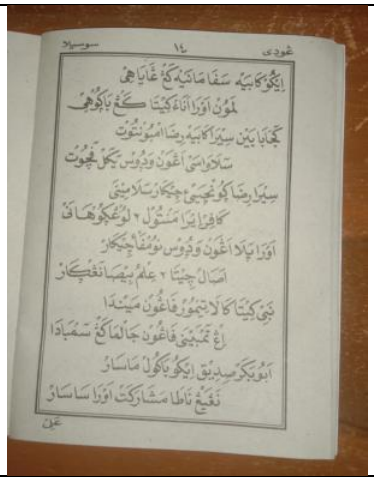


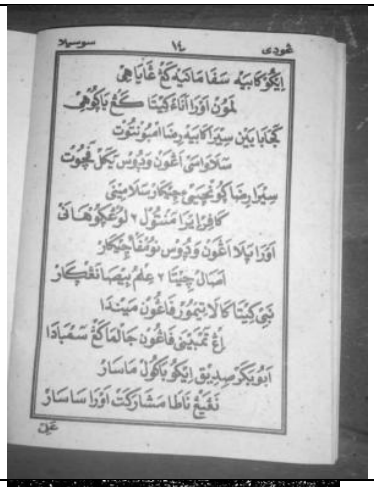
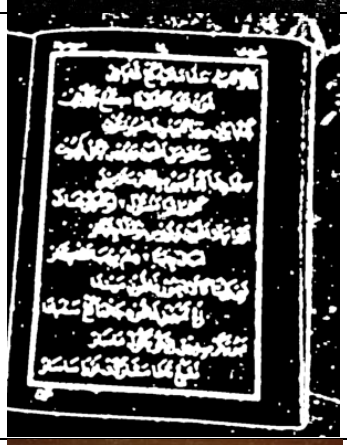
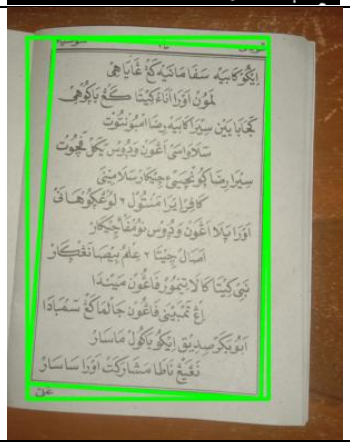
Gambar 3. Kitab *Ngudi Susilo*

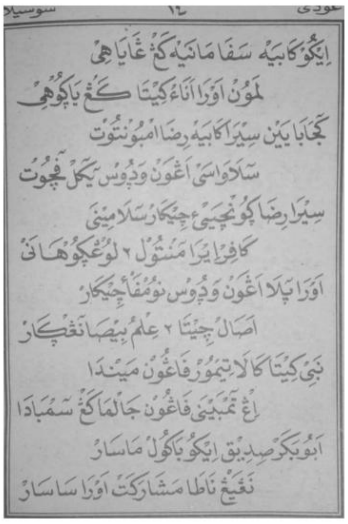
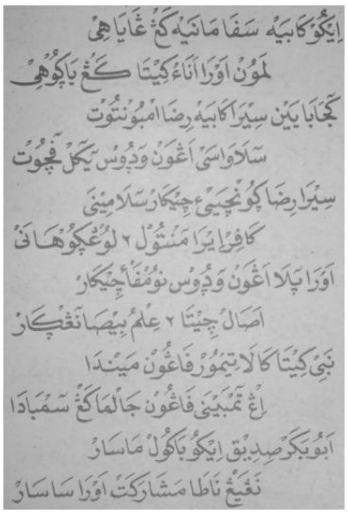
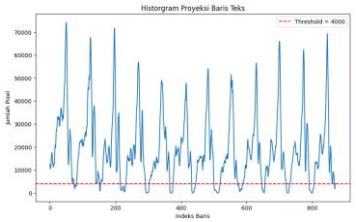
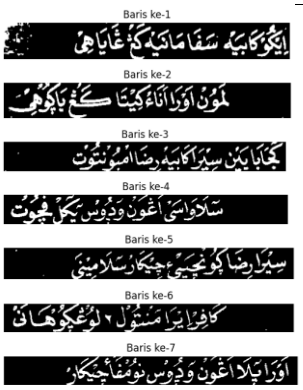
A. Proses Segmentasi

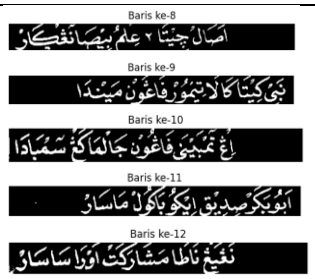
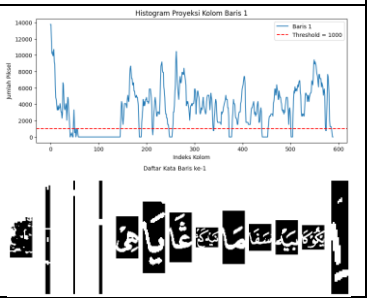
Proses segmentasi meliputi dua tahap utama, yaitu segmentasi baris dan segmentasi kata. Segmentasi baris dilakukan untuk memisahkan baris demi baris yang terdapat pada halaman kitab. Setelah setiap baris terpisahkan, maka langkah selanjutnya adalah memisahkan setiap kalimat yang terdapat pada baris tersebut. Proses inilah yang dinamakan segmentasi kata. Apabila model dapat secara akurat melakukan segmentasi dengan baik, maka dapat digunakan untuk mengembangkan model lanjutan untuk melakukan transliterasi aksara Arab-Pegon. Seluruh rangkaian tahapan segmentasi disajikan pada Tabel 1.

TABEL 1
TAHAPAN SEGMENTASI NASKAH ARAB-PEGON

No.	Tahap	Keterangan	Hasil
1	*Original Images Scanning*	Pemindaian citra asli dalam format RGB untuk diuji ke dalam model. Citra yang dipindai adalah salah satu halaman naskah kitab *Ngudi Susilo*.	
2	*Preprocessing*	Tahap preprocessing meliputi tujuh proses, yaitu: a. Konversi citra ke *grayscale* Konversi citra ke format abu-abu dan pembalikan intensitas dilakukan untuk menyiapkan citra agar fitur-fitur seperti garis tepi lebih mudah dikenali.	

		b. Dilasi morfologis Konversi citra ke format abu-abu dan pembalikan intensitas dilakukan untuk menyiapkan citra agar fitur-fitur seperti garis tepi lebih mudah dikenali.	
		c. *Binarization* *Binerization* digunakan untuk mengubah citra ke format biner, sehingga garis tepi (*border*) menjadi semakin kontras dan mudah dipisahkan dari area teks.	
		d. Operasi morfologis tambahan Operasi morfologis tambahan diterapkan untuk menyambungkan garis yang terputus dan menghilangkan *noise* kecil, agar *border* terbentuk secara utuh dan bisa dideteksi sebagai satu kesatuan.	
		e. *Boundary Contour* Kontur segi empat terbesar dicari karena biasanya mewakili bingkai atau batas luar halaman dokumen, termasuk *border* yang ingin dihilangkan dan melakukan *cropping*.	

		f. Transformasi perspektif Transformasi perspektif dilakukan berdasarkan kontur tersebut untuk merapikan posisi halaman dan membuat *border* tampak tegak lurus terhadap gambar.	
		g. *Cropping* Pemotongan (*cropping*) berbasis kontur digunakan untuk menghapus bagian luar citra, termasuk *border*, sehingga hanya area isi dokumen yang tersisa.	
3.	*Text Line Detection*	Tahap ini digunakan untuk mendeteksi baris naskah. Terdapat dua sub-tahapan di dalamnya, yaitu: a. Histogram proyeksi horizontal Histogram proyeksi horizontal digunakan untuk menganalisis sebaran intensitas piksel secara horizontal.	
		b. Segmentasi baris Hasil segementasi baris dipengaruhi oleh nilai *threshold* histogram.	

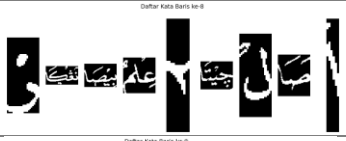
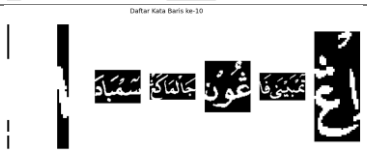
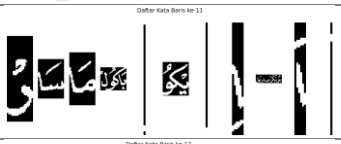
			
4.	*Words Detection*	Setelah silakukan segmentasi baris, maka tahap terakhir dari *histogram segmentation* adalah melakukan segmentasi tiap kata pada baris tersebut.	

B. Evaluasi

Citra yang diuji merupakan halaman naskah kitab dengan aksara Arab-Pegon yang terdiri dari 12 baris kalimat. Metode *histogram segmentation* mampu mensegmentasi seluruh baris dengan tepat seperti yang tampak pada Tabel 1. Namun, untuk mensegmentasi kata masih terdapat ketidak-akuratan. Dari 12 baris kalimat masih dijumpai susunan kata yang tidak tersegmentasi dengan utuh.

TABEL 2
HASIL SEGMENTASI KATA

Baris	Hasil Segmentasi Kata	Kesalahan Segmentasi
Baris ke-1		*Over-segmentation* pada tengah baris, satu kata Pegon terpecah menjadi beberapa segmen akibat ligatur (ekor) rapat dan celah vertikal semu.
Baris ke-2		*Under-segmentation* pada antar kata di sisi kanan baris, dua kata Pegon masih tergabung karena jarak antar kata terlalu sempit.
Baris ke-3		*Under-segmentation* pada antar kata di sisi kanan baris, dua kata Pegon masih tergabung karena jarak antar kata terlalu sempit
Baris ke-4		*Over-segmentation* pada akhir kata, terutama pada huruf Pegon berekor panjang yang terpotong dari kata utamanya.
Baris ke-5		Muncul *noise segmentation* pada area antar kata di tengah baris, menghasilkan segmen kecil non-teks.
Baris ke-6		*Under-segmentation* pada dua kata di bagian tengah baris, batas antar kata tidak terdeteksi akibat sambungan huruf yang rapat.

Baris ke-7		Kesalahan kombinatif: over-segmentation pada awal baris dan *under-segmentation* pada akhir baris.
Baris ke-8		Kesalahan pemotongan pada awal kata di sisi kanan baris, awalan kata terpisah akibat perbedaan intensitas piksel.
Baris ke-9		*Over-segmentation* pada tengah kata di bagian tengah baris, sambungan huruf Pegon terpotong menjadi dua segmen.
Baris ke-10		*Over-segmentation* dominan pada seluruh baris, banyak segmen kecil terbentuk dari goresan vertikal tipis.
Baris ke-11		*False segmentation* pada sisi kiri baris, muncul segmen kosong akibat deteksi minimum lokal palsu.
Baris ke-12		Under-segmentation pada tengah hingga akhir baris, beberapa kata Pegon tergabung karena jarak antar kata sangat rapat.

Evaluasi hasil segmentasi dilakukan dengan membandingkan antara hasil segmentasi dengan teks aslinya, diperoleh hasil seperti pada Tabel 3.

TABEL 3
STATISTIK NILAI SEGMENTASI KATA

Baris	GT (kata asli)	CS (benar)	OS	US
1	7	5	2	0
2	6	5	0	1
3	6	4	1	1
4	5	4	1	0
5	6	5	1	0
6	6	4	0	2
7	7	5	1	1
8	5	4	1	0
9	5	4	1	0
10	6	3	3	0
11	5	3	2	0
12	6	4	0	2
Total	**70**	**49**	—	—

- *Ground Truth* (GT): jumlah kata sebenarnya pada teks Pegon asli.
- *Correctly Segmented Words* (CS): kata yang tersegmentasi tepat 1 banding 1 dengan *ground truth*
- *Over-Segmentation Error* (OS): 1 kata GT terpecah menjadi >1 segmen
- *Under-Segmentation Error* (US): 1 kata GT tergabung menjadi 1 segmen

Akurasi dapat dihitung dengan rumus:

$$Akurasi = \frac{CS}{GT}\ x\ 100\% \qquad (1)$$

$$Akurasi = \frac{49}{70}\ x\ 100\% = 70\%$$

Akurasi segmentasi kata dihitung dengan membandingkan jumlah kata yang tersegmentasi secara benar terhadap jumlah kata *ground truth*. Dari total 70 kata Pegon pada 12 baris teks, sebanyak 49 kata berhasil tersegmentasi dengan tepat, sehingga diperoleh akurasi segmentasi sebesar 70%. Adapun *error rate* (kesalahan segmentasi) dapat dihitung dengan mengevaluasi *Over-Segmentation Rate* (OSR) dan *Under-Segmentation Rate* (USR). OSR dapat dihitung menggunakan formula sebagai berikut:

$$OSR = \frac{OS}{GT} \qquad (2)$$

$$OSR = \frac{14}{70} = 0.20$$

Nilai OSR yang diperoleh sebesar 0.20, sedangkan USR dapat dihitung dengan rumus:

$$USR = \frac{US}{GT} \qquad (3)$$

$$USR = \frac{9}{70} = 0.13$$

Nilai *Over-Segmentation Rate* (OSR) sebesar 0,20 menunjukkan bahwa 20% kata *ground truth* mengalami pemotongan berlebih, di mana satu kata Pegon terpecah menjadi lebih dari satu segmen. Kesalahan *over-segmentation* umumnya ditemukan pada karakter Pegon yang memiliki ligatur rapat serta goresan horizontal yang relatif panjang. Kondisi tersebut menyebabkan terbentuknya minimum lokal semu pada histogram vertikal, sehingga satu kata dapat terpecah menjadi beberapa segmen yang seharusnya tidak dipisahkan.

Di sisi lain, nilai *Under-Segmentation Rate* (USR) sebesar 0,13 menunjukkan bahwa sekitar 13% kata *ground truth* masih tergabung dengan kata lain. Kesalahan ini terjadi ketika batas antar kata tidak teridentifikasi secara jelas, terutama pada teks Pegon dengan jarak antar kata yang sangat sempit dan bentuk huruf yang saling menyambung secara kontinu.

Secara keseluruhan, dominasi kesalahan *over-segmentation* dibandingkan *under-segmentation* mengindikasikan bahwa metode *histogram segmentation* memiliki sensitivitas yang cukup tinggi terhadap variasi ketebalan goresan dan kompleksitas ligatur aksara Pegon. Meskipun demikian, dengan tingkat akurasi segmentasi sebesar 70%, pendekatan ini masih dapat dipandang efektif sebagai metode dasar (*baseline*) untuk segmentasi kata. Hasil ini menunjukkan potensi metode histogram segmentation untuk dikembangkan lebih lanjut, misalnya melalui penyesuaian ambang histogram yang bersifat adaptif atau dengan mengombinasikannya dengan pendekatan pembelajaran mesin guna meningkatkan ketepatan penentuan batas segmentasi.

IV. SIMPULAN

Metode *histogram segmentation* terbukti efektif digunakan untuk melakukan segmentasi teks Arab-Pegon. Hal ini terbukti dari kemampuannya dalam melakukan segmentasi baris dengan sangat akurat. Berdasarkan hasil segmentasi citra uji, seluruh baris berhasil tersegmentasi tanpa kesalahan. Sedangkan untuk segmentasi kata menunjukkan tingkat akurasi sebesar 70% dengan *error rate* pada *Over-Segmentation Rate* (OSR) sebesar 0,20 dan *Under-Segmentation Rate* (USR) sebesar 0,13. Metode ini masih sangat tergantung dengan nilai *threshold* yang ditentukan pada tahap segmentasi baris dan kata. Oleh karena itu diperlukan penelitian lanjutan yang berfokus pada komparasi nilai *threshold* terbaik dan juga menemukan metode penentuan *theshold* secara otomatis (*adaptive threshold*) pada varian teks Arab-Pegon yang berbeda beda.

UCAPAN TERIMAKASIH

Tim peneliti merasa sangat bersyukur atas dukungan penuh dari institusi Universitas Harkat Negeri terhadap penelitian ini. Dukungan berupa fasilitas pendanaan dan juga laboratorium komputer untuk menunjang penelitian. Publikasi ini merupakan luaran dari penelitian institusi yang dikelola di bawah Unit Penelitian dan Pengabdian Masyarakat.

DAFTAR PUSTAKA

[1] Y. Yuliani, "Aksara Tafsir Al-Qur'an Di Priangan:Huruf Pegon Dan Aksara Latin Dalam Karya K.H. Ahmad Sanoesi," *Al-Bayan J. Stud. Ilmu Al- Qur'an dan Tafsir*, vol. 5, no. 1, 2020.

[2] I. Gusmian, "Bahasa dan Aksara Dalam Penulisan Tafsir Al-Qur'an di Indonesia Era Awal Abad 20 M," *Mutawâtir J. Keilmuan Tafsir Hadis*, vol. 5, no. 2, p. 223, Sep. 2015.

[3] R. Rusyadi, *Bahasa Arab Pesantren: Sejarah dan Tradisi Literasi Pegon di Nusantara*, Pertama. Malang: Madza Media, 2021.

[4] I. R. N. Hula, A. Helingo, S. N. A. Jassin, and S. Sarif, "Transcription of Pegon Gorontalo Arabic Orthography, Malay and Arabic Standard: A Contraceptive Linguistic Analysis," *`A Jamiy J. Bhs. dan Sastra Arab*, vol. 11, no. 2, p. 322, 2022.

[5] D. Syafaah, N. B. Rohmah, and U. Rejo, "Pemulihan warisan budaya: desain laboratorium filologi sebagai pusat pembelajaran manuskrip keislaman jawa pesisir," *Humanika*, vol. 31, no. 1, pp. 1–15, 2024.

[6] Elmustian and M. Firdaus, "Filologi, Transformasi Teks, dan Filsafat Pendidikan: Strategi Pelestarian Budaya dalam Konteks Pendidikan Kontemporer," *Indones. Res. J. Educ.*, vol. 4, no. 4, 2024.

[7] E. Purnama, "Pelestarian Koleksi Buku Langka Di Perpustakaan Universitas Gadjah Mada," *J. Multidisipliner Kapalamada*, vol. 2, no. 04, pp. 227–239, 2023.

[8] Y. Ruldeviyani, H. Suhartanto, B. A. Sotardodo, M. H. Fahreza, A. Septiano, and M. F. Rachmadi, "Character recognition system for pegon typed manuscript," *Heliyon*, vol. 10, no. 16, p. e35959, 2024.

[9] N. A. Jebril, H. R. Al-Zoubi, and Q. Abu Al-Haija, "Recognition of Handwritten Arabic Characters using Histograms of Oriented Gradient (HOG)," *Pattern Recognit. Image Anal.*, vol. 28, no. 2, pp. 321–345, 2018.

[10] H. M. Al-Barhamtoshy and S. M. Abdou, "Arabic Manuscripts Alignment, Segmentation, Recognition, and Classification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2025.

[11] T. B. A. Gader and A. K. Echi, "Attention-based CNN-ConvLSTM for Handwritten Arabic Word Extraction," *Electron. Lett. Comput. Vis. Image Anal.*, vol. 21, no. 1, 2022.

[12] S. Alghyaline, "Arabic Optical Character Recognition: A Review," *C. - Comput. Model. Eng. Sci.*, vol. 135, no. 3, pp. 1825–1861, 2023.

[13] I. Saleh Al-Sheikh, M. Mohd, and L. Warlina, "A Review of Arabic Text Recognition Dataset," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 09, no. 01, pp. 69–81, 2020.

[14] M. N. Elagamy, M. M. Khalil, and E. Ismail, "HACR-MDL: Handwritten Arabic Character Recognition Model Using Deep Learning," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 10, no. 1-W1-2023, pp. 123–128, 2023.

[15] S. I. Saleem, A. M. Abdulazeez, and Z. Orman, "A new segmentation framework for arabic handwritten text using machine learning techniques," *Comput. Mater. Contin.*, vol. 68, no. 2, pp. 2727–2754, 2021.

[16] H. A. Al Hamad, L. Abualigah, M. Shehab, K. H. A. Al-Shqeerat, and M. Otair, "Improved linear density technique for segmentation in Arabic handwritten text recognition," *Multimed. Tools Appl.*, vol. 81, no. 20, pp. 28531–28558, Aug. 2022.

[17] A. Mostafa *et al.*, "An End-to-End OCR Framework for Robust Arabic-Handwriting Recognition using a Novel Transformers-based Model and an Innovative 270 Million-Words Multi-Font Corpus of Classical Arabic with Diacritics," pp. 1–31, 2022.