Sistem Deteksi Elemen Teks pada Naskah Sunda Kuno

Erick Paulus[1], Mira Suryani[2], Rudi Rosadi[3], Akik Hidayat[4]
Departemen Ilmu Komputer Universitas Padjadjaran Bandung
[1]erick.paulus@unpad.ac.id, [2]mira.suryani@unpad.ac.id, [3]rudi.rosadi@unpad.ac.id, [4]akik@unpad.ac.id

Abstrak *– Naskah Sunda kuno yang merupakan warisan budaya dalam bentuk tulisan tangan memiliki kekayaan informasi yang berkaitan dengan kehidupan kebudayaan Sunda antara abad ke-8 sampai abad ke-16. Berbagai upaya sudah dilakukan untuk mengkonservasi benda budaya tersebut, seperti dijitalisasi naskah menjadi citra dijital. Oleh karena itu, sistem deteksi elemen teks diperlukan oleh OCR untuk mengenali karakter atau kata yang terdapat pada naskah Sunda kuno. Penelitian ini dirancang untuk mengujicoba sistem deteksi elemen teks yang terdiri dari tahapan binerisasi, pemberian label elemen teks, dan pemotongan elemen teks. Data naskah sunda yang dipakai dalam penelitian analisa dokumen ini adalah sebanyak 10 sample lontar. Tujuan penelitian ini adalah untuk mengukur keakuratan sistem deteksi elemen teks. Hasil percobaan menunjukan sistem deteksi memiliki keakuratan sebesar 90% untuk kualitas citra dengan iluminasi seragam. Namun metode deteksi ini masih bergantung dari hasil proses binerisasi.*

Kata kunci*: deteksi elemen teks, binerisasi, connected component, naskah sunda kuno*

1. Pendahuluan

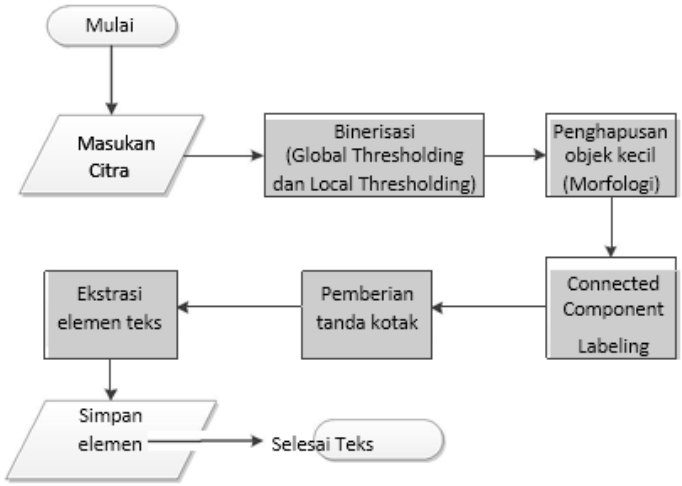
Berdasarkan sudut pandang kebudayaan, naskah merupakan salah satu warisan budaya kebendaan (*tangible cultural heritage*) yang bersifat kongkrit (*material culture*), dan juga mengandung teks yang dapat dikelompokkan sebagai salah satu warisan budaya nonkebendaan (*intangible cultural heritage*) yang bersifat abstrak. Naskah Sunda kuno yang ditulis sekitar abad ke-8 sampai abad ke-16 menceritakan kehidupan kebudayaan Sunda yang berkaitan dengan adat-istiadat, sastra, keagamaan, mitologi, ilmu pengetahuan, paririmbon/mujarobat, dan pendidikan. Sekarang ini, beberapa naskah sunda kuno tersebut sudah berbentuk versi dijital berupa citra. Selain itu, banyak masyarakat Sunda juga yang semakin tertarik untuk memahami isi informasi yang terkandung dalam naskah sunda kuno tersebut. Oleh karena itu, analisa citra dokumen sejarah menjadi penting sehingga mempermudah masyarakat atau para peneliti untuk mengeksplorasi informasi yang terkandung di dalam naskah tersebut.

Salah satu tahapan dalam analisa citra dokumen adalah deteksi elemen teks yang tertulis pada naskah tersebut. Namun, kondisi fisik lontar yang beragam menjadi tantangan tersendiri ketika sistem mendeteksi elemen teks. Beberapa penelitian melalukan pemrosesan citra (*image processing*) seperti penyaringan derau, perbaikan kualitas citra, normalisasi, perbaikan kemiringan dan binerisasi dilakukan sebelum analisa citra (*image anaysis*) [1][2][3]. Proses binerisasi dilakukan sebagai tahapan pra proses sebelum dilakukannya segmentasi teks dan Optical Character Recognation [4]. Septiarini melakukan profil proyeksi terhadap karakter cetak (*printed document*) dengan latar belakang putih dengan diawali proses binerisasi dengan *global thresholding* [1]. Keberhasil 90% yang dicapai oleh Septiarini tidak berlaku untuk dokumen naskah kuno. Kesiman membuat langkah peningkatan kualitas citra dengan memetakan citra asli naskah Bali ke nilai *lacunarity* sebelum dilakukan segmentasi area teks [2]. Selain itu, penerapan metode *mathematical morphology* dengan operasi erotion sangat baik digunakan pada tahap pre-processing. Setelah kualitas citra semakin baik, Profil proyeksi horizontal dan vertikal digunakan untuk memisahkan elemen baris dan elemen teks. Perkembangan kinerja metode connected component labelling (CCL) yang semakin meningkat dapat dijadikan salah satu alternatif dalam proses segmentasi elemen teks[5]. Oleh karena itu, peneliti melakukan uji eksperimen metode CCL dalam mendeteksi elemen teks. Adapun yang termasuk elemen teks

pada aksara sunda kuno adalah berupa aksara suara (vokal), aksara ngalagena, dan rarangken. Posisi rarangken bisa di atas huruf, di bawah huruf, dan sejajar huruf.

2. Metode Penelitian

Penelitian ini merupakan hasil uji eksperimen metode deteksi elemen teks citra naskah Sunda kuno yang memiliki karakteristik unik. Diagram alir proses deteksi elemen teks diantaranya adalah proses binerisasi, penghapusan objek kecil, connected component labeling, Pemberian tanda kotak, ekstraksi elemen teks dan simpan potongan elemen teks tersebut (lihat gambar 1). Data yang digunakan dalam penelitian ini adalah beberapa data lontar yang diperoleh dari Perpustakaan Nasional Republik Indonesia(PNRI) dan Situs Kabuyutan Ciburuy, Garut.



Gambar 1. Diagram Alir Deteksi Elemen Teks

2.1 Proses Binerisasi

Menurut Yahya, ada tiga kelompok cara meningkatkan kualitas citra, yaitu menggunakan metode binerisasi/thresholding, menggunakan metode hibrid binerisasi dengan metode lainnya, dan menggunakan metode non-binerisasi[6]. Fungsi binerisasi adalah mentransformasi citra keabuan menjadi citra biner. Nilai citra biner hanya ada dua, yaitu nilai 0 (putih) untuk mewakili elemen non teks atau 1 (hitam) untuk merepresntasikan elemen teks [7]. Proses binerisasi citra dokumen merupakan tahapan awal dalam analisa dan memiliki tujuan untuk memisahkan elemen teks (*foreground text*) dari elemen non teks (*background document*) [8]. Penelitian lainnya menyebutkan bahwa hasil binerisasi yang semakin baik akan membuat pengenalan karakter lebih akurat [4].

2.2 Thresholding

Konversi citra digital menjadi citra biner ditentukan oleh besarnya nilai ambang batas (threshold). Berbagai algoritma binerisasi untuk naskah dokumen bersejarah telah diteliti. Khususnya tahapan thresholding, penentuan nilai ambang batas ada yang bersifat global dan lokal[9]. Metode Otsu merupakan salah satu metode thresholding yang paling efektif dengan menerapkan analisa clustering menggunakan distributi Gaussian dari pixel [10] [11]. Metode ini sesuai untuk citra dengan kualitas baik, iluminasi seragam, dan kontras yang tinggi. Kekurangan metode ini adalah terdapat difusi garis yang disebabkan oleh aspek iluminasi tidak

ISBN: 978-602-60581-0-2

seragam. Persamaan 1 merupakan rumus metode Otsu dengan T adalah nilai ambang batas, f(i,j) adalah fungsi citra asli, dan h(i,j) adalah citra biner.

$$h(i,j) = \begin{cases} 1\ (white), & if\ f(i,j) \geq T \\ 0\ (black), & otherwise \end{cases} \quad (1)$$

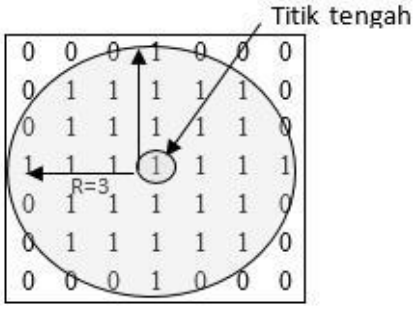
Namun untuk kasus lontar sunda kuno, penentuan nilai ambang batas lebih baik ditentukan secara lokal untuk setiap sub window. Salah satu metode local thresholding yang relatif baik adalah Metode Sauvola [12]. Formula Metode Sauvola direpresentasikan pada persamaan 2. Parameter m dan s adalah nilai rata-rata dan standar deviasi dari nilai intensitas warna pada sub window, dan parameter k adalah nilai konstanta [-1,0) yang dipilih untuk mendapatkan kualitas binerisasi terbaik. Nilai k kurang dari -0.2 cocok untuk mendeteksi obejek hitam, selebihnya cocok untuk mendeteksi objek putih.

$$T = m \times \left[1 + k\left(1 - \frac{s}{R}\right)\right] \quad (2)$$

2.3 Penghapusan Objek Kecil

Pada umumnya, hasil binerisasi dari naskah sunda kuno tersebut masih terdapat derau yang menempel pada teks atau noda. Pada penelitian ini, metode morfologi open (pada Matlab menggunakan fungsi *bwareopen*) digunakan untuk menghapus derau atau noda kecil yang tidak melekat pada teks dengan cara menghapus objek *connected component* yang lebih kecil dari nilai batas. Karena beragamnya kondisi citra, maka penentuan nilai batas tidak dapat dilakukan secara otomatis. Penentuan nilai batas ini juga dipengaruhi ukuran resolusi citra. Berdasarkan ujicoba, nilai batas terbaik yang dipakai pada penelitian ini adalah 50.

Naskah Sunda kuno tertulis pada lempir daun lontar yang disusun dan dirangkai menggunakan tali pada bagian tengah lontar sebagai penyambung antar lempir. Lubang kecil sebagai jalur pengikat antar lempir akan terlihat hitam ketika proses akusisi data dilakukan. Oleh karena itu, noda lubang hitam sebisa mungkin dihapus supaya menghasilkan deteksi elemen teks yang lebih baik. Metode penghapusannya menggunakan metode morfologi closing dengan elemen struktur morfologi bentuk disk. Ilustrasi struktur elemen morfologi dapat dilihat pada gambar 2.



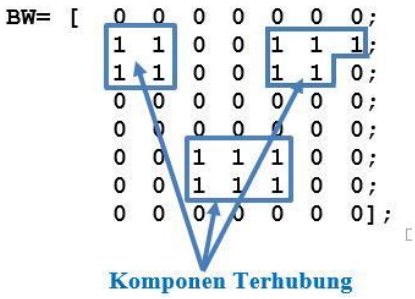
Gambar 2. Struktur elemen morfologi disk

2.4 Connected Component Labelling

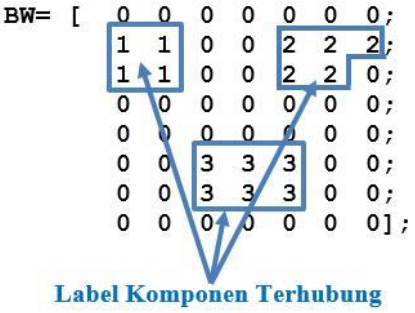
Komponen terhubung (*connected component*) pada citra biner adalah sekelompok pixel yang membentuk koloni yang saling terhubung. Gambar 3 mengilustrasikan citra biner yang memiliki tiga komponen terhubung. Kemudian, label komponen terhubung adalah proses

ISBN: 978-602-60581-0-2

identifikasi komponen terhubung pada suatu citra dan pemberian nilai label unik ke masing-masing kelompok pixel (lihat gambar 4). Fungsi bwlabel pada Matlab sudah ditingkatkan kinerjanya menjadi fungsi bwconncomp sehingga komputasi dapat dilakukan lebih fleksibel dan hemat memori[13].



Gambar 3. Ilustrasi kompenen terhubung (*connected components*)



Gambar 4. Ilustrasi Label kompenen terhubung (*labeled connected components*)

3. Hasil dan Analisis

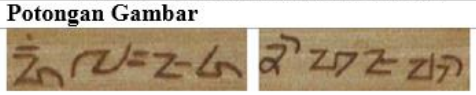
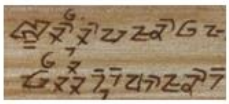
Pada bagian ini dipaparkan hasil eksperimen yang dilakukan mulai dari penerapan skema binerisasi sampai proses deteksi elemen teks dengan menggunakan Label komponen terhubung. Data yang digunakan pada penelitian ini diperoleh dari situs Kabuyutan Ciburuy Garut dan Perpustakaan Nasional Republik Indonesia. Dataset naskah sunda kuno dikelompokan berdasarkan kondisi citra (lihat tabel 1 dan 2).

Tabel 1. Karakteristik Citra Lontar

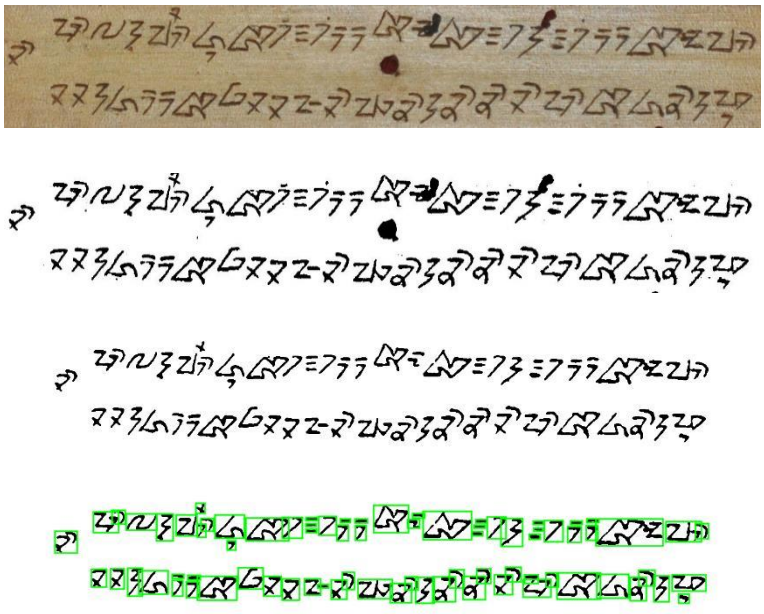
Kondisi	**Keterangan**
Baik	memiliki tingkat kecerahan degradasi yang bagus dan tanpa titik, noda, bercak, bayangan
Noda atau spot	Memiliki noda titik, noda akibat tinta, bercak dan terdapat sedikit atau banyak derau pada gambar latar belakang
Retakan	Terdapat efek retakan pada lontar sehingga terbentuk seperti garis yang dapat mengakibatkan aksara saling terhubung
Bayangan	Lontar berubah warna akibat usia atau kelembapan yang tinggi
Iluminasi akibat akuisisi data	Adanya iluminasi akibat pencahayaan saat prosess akuisisi data. Contoh terang pada bagian tengah dan gelap pada bagian tepi

ISBN: 978-602-60581-0-2

Tabel 2. Contoh potongan Citra Lontar

Kondisi	Potongan Gambar
Baik	
Noda atau spot	
Retakan	
Bayangan	
Iluminasi akibat akuisisi data	

Selanjutnya, 10 sample lontar diuji sesuai dengan skema deteksi yang sudah dijelaskan pada bagian 2. Langkah awal adalah proses binerisasi menggunakan metode Sauvola yang menghitung nilai ambang secara lokal pada setiap sub windows (*adaptive thresholding*). Ukuran sub windows adalah 31 x 31. Pengapusan lingkaran tengah dan noda yang berbentuk serupa lingkaran diproses dengan metode morfologi dengan struktur elemen *disk*. Selain itu, derau yang relatif kecil juga dihapus dengan menggunakan komponen terhubung (connected component) yang lebih kecil dari 50. Selanjutnya, citra biner yang relatif sudah bersih diberi label dengan menggunakan CCL dan dilakakukan ekstrasi berdasarkan label tersebut. Citra hasil setiap proses dapat dilihat pada gambar 5.



Gambar 5. Citra hasil setiap tahapan secara berurutan atas ke bawah : citra warna asli, citra biner Sauvola, citra setelah morfologi, citra setelah diberi label

Berdasarkan eksperimen menunjukan bahwa citra lontar dengan kondisi terdapat noda, bercak, bayangan masih dapat terdeteksi dengan baik seperti lontar DSC_0298_8a, DSC_0297_4b dan DSC_0299_9b . Noda yang melekat pada teks dan menghubungkan dua tau

ISBN: 978-602-60581-0-2

lebih elemen teks akan mengakibat tergabungnya elemen tersebut seperti lontar S9 dan S7. Jadi perlu dilakukan penangan tersendiri lagi ketika proses OCR. Lontar S10, DSC_0452_4a, dan DSC_0452_5a, yang mengandung retakan, kontras warna yang rendah dan iluminasi tidak seragam belum dapat mendeteksi dengan begitu baik bahkan menghasilkan kesalahan deteksi yang besar. Penelitian selanjutnya perlu ditingkatkan dalam hal metode perbaikan kualitas citra atau dirancang skema metode segmentasi tanpa proses binerisasi. Penjelasan lebih rinci dapat dilihat pada tabel 3.

Tabel 3. Hasil deteksi elemen teks

Citra Lontar	Jumlah target elemen teks	Jumlah elemen terdeteksi		Keterangan
		Benar	Salah	
DSC_0298_8a .tif	101	98	0	Jumlah kesalahan 0 dikarenakan kondisi citra relatif baik sekalipun ada bayangan. Namun jumlah elemen teks yang terdeteksi 98 lebih kecil dari jumlah target. Hal ini disebabkan karena ada beberapa elemen teks yang beririsan sehingga dianggap 1 elemen
DSC_0452_4a .tif	393	242	63	Jumlah kesalahan cukup besar dikarenakan adanya noda, bercak dan citra kurang kontras sehingga ada banyak elemen teks yang dianggap seelemen
DSC_0452_5a .tif	475	317	87	Jumlah kesalahan cukup besar disebabkan adanya iluminasi tidak seragam dan noda yang melekat pada elemen teks sehingga seakan-akan tergabung menjadi satu elemen
DSC_0297_4b .tif	76	64	6	Jumlah kesalahan 6 disebabkan adanya noda
DSC_0299_9b .tif	81	77	0	Jumlah kesalahan 0 dikarenakan kondisi citra relatif baik. namun jumlah elemen teks yang terdeteksi 77 lebih kecil dari jumlah target. Hal ini disebabkan karena ada beberapa elemen teks yang beririsan sehingga dianggap 1 elemen
S17.tif	8	8	1	Jumlah kesalahan 1 karena ada elemen teks yang hilang
S16.tif	6	6	0	Tidak ditemukan kesalahan karena sample yang digunakan baik, yaitu elemen teks terlihat jelas dan kontras
S9.tif	11	10	5	Jumlah kesalahan 5 disebabkan adanya noda kecil dan noda yang nempel dengan elemen teks
S7.tif	6	4	1	Jumlah kesalahan 1 disebabkan karena adanya noda yang nempel pada 2 elemen teks
S10.tif	8	3	15	Jumlah kesalahan 15 jauh dari jumlah target disebabkan kondisi citra rendah seperti adanya retakan dan derau yang begitu banyak

4. Kesimpulan

Pengujian deteksi elemen teks dengan menggunakan pendekatan connected component labelling telah dilakukan dan berhasil untuk citra lontar warna dengan mengikuti skema deteksi. Jumlah elemen teks terdeteksi hampir mendekati jumlah target untuk citra yang relatif baik walaupun terdapat noda, bayangan atau bercak. Untuk noda yang beririsan dengan dua atau

ISBN: 978-602-60581-0-2

lebih elemen teks akan berakibat tergabungnya elemen tersebut sehingga dapat mengakibatkan keambiguaan saat proses OCR. Saran penelitian selanjutnya adalah perlu dilakukan segmentasi baris sehingga proses CCL dapat dilakukan secara lokal per baris. Selain itu, pengembangan penelitian selanjutnya adalah proses segmentasi tanpa melakui proses binerisasi.

Daftar Pustaka

[1] A. Septiarini, "Segmentasi Karakter Menggunakan Profil Proyeksi," *J. Inform. Mulawarman*, vol. 7, no. 2, pp. 66–69, 2012.

[2] I. M. G. Sunarya, M. W. A. Kesiman, and I. A. P. Purnami, "Segmentasi citra tulisan tangan aksara bali berbasis proyeksi vertikal dan horisontal," J. Inform., vol. 9, no. 1, pp. 982–992, 2015.

[3] M. W. A. Kesiman, "Segmentasi Area Teks Aksara Bali pada Citra Lontar Kuno Bali Berdasarkan Peta Nilai Lacunarity," in Seminar Nasional Aplikasi Teknologi Informasi, 2013, pp. 7–12.

[4] Puneet and N. K. Garg, "Binarization Techniques used for Grey Scale Images," Int. J. Comput. Appl., vol. 71, no. 1, pp. 8–11, 2013.

[5] A. Rakhmadi, N. Z. S. Othman, A. Bade, M. S. M. Rahim, and I. M. A. Department, "Connected Component Labeling Using Components Neighbors-Scan Labeling Approach," J. Comput. Sci., vol. 6, no. 10, pp. 1099–1107, 2010.

[6] S. R. Yahya, S. N. H. S. Abdullah, K. Omar, and M. S. Zakaria, "Review on Image Enhancement Methods of Old Manuscript with Damaged Background," Int. J. Electr. Eng. Informatics, vol. 2, no. 1, pp. 1–14, 2010.

[7] J. Kaur and R. Mahajan, "A Review of Degraded Document Image Binarization Techniques," Int. J. Adv. Res. Comput. Commun. Eng., vol. 3, no. 5, pp. 6581–6586, 2014.

[8] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," IEEE Trans. Image Process., vol. 22, no. 4, pp. 1408– 1417, 2013.

[9] N. Ntogas and D. Veintzas, "A binarization algorithm for historical manuscripts," Proc. 12th WSEAS Int. Conf. Commun., pp. 41–51, 2008.

[10] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2013 document image binarization contest (DIBCO 2013)," Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, no. Dibco, pp.1471–1476, 2013.

[11] N. Ntogas and D. Ventzas, "A Binarization Algorithm For Historical Manuscripts," in 12th WSEAS International Conference on communications, 2008, pp. 41–51.

[12] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.

[13] Matlab Documentation. Matlab Graphics. The MathWork, Inc, 2016

ISBN: 978-602-60581-0-2