ANALISIS KINERJA OPTICAL CHARACTER RECOGNITION UNTUK MEMBACA DOKUMEN SECARA OTOMATIS

Susan Siti Nurhaliza[1], M. Subali[1], Lussiana ETP[2] dan Rozi[2]

[(1)] Universitas Cendekia Abditama
Jl. Islamic Raya No.1, Klp. Dua, Kec. Klp. Dua, Tangerang, Banten 15812
[(2)]STMIK Jakarta STI&K
Jl. BRI No.17, Radio Dalam, Kebayoran Baru, Jakarta Selatan 12140
susan@uca.ac.id, subali@cendekia.ac.id,
{lussiana.etp, roziborang}@gmail.com

ABSTRAK

Optical character recognition (OCR) merupakan teknologi untuk mengenali karakter pada suatu citra, termasuk text atau dokumen. Salah satu menfaat implementasi metode OCR adalah untuk pengenalan dokumen pada bagian customs clearance. Berdasarkan fakta penggunaan huruf suatu dokumen sangat beragam, tidak hanya jenis huruf Calibri. Berdasarkan hal tersebut tujuan penelitian ini adalah melakukan pengkajian kinerja OCR dalam mengenali karakter dokumen dengan menggunakan jenis huruf Arial, Bahnschrift Condensed, Georgia, Lucida Sans Unicode, Roman, Segoe UI Semibold, dan Times New Roman. Adapun tahapan-tahapan penelitian antara lain adalah preprocessing yang terdiri dari proses grayscaling, binerisasi, cropping selanjutnya adalah tahap segmentasi, ekstraksi fitur dan untuk proses terakhir adalah proses metode pencocokkan karakter berdasarkan pada template matching. Berdasarkan hasil pengujian metode OCR mampu mengenali dengan akurasi 100% untuk karakter jenis huruf Georgia, Lucida Sans Unicode, Roman dan Segoe UI Semibold sedangkan akurasi terendah 71.21% pada karakter jenis huruf Bahnschrift Condensed. Dengan demikian dapat disimpulkan bahwa metode OCR secara umum mampu mengenali karakter, namun demikian masih terbuka untuk melakukan pengembangan penelitian untuk meningkatkan akurasi jenis huruf lain. Ditinjau dari waktu proses metode OCR relatif singkat, yaitu rata-rata 1.33 detik.

Kata Kunci : *Pengenalan karakter, OCR, Template Matching*

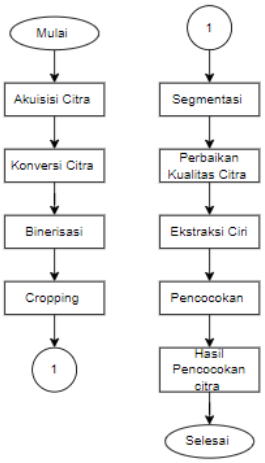
PENDAHULUAN

Pengolahan citra termasuk pengenalan pola, saat ini sudah banyak diimplementasikan untuk menyelesaikan berbagai masalah, antara lain pengenalan wajah untuk keamanan, identifikasi originalitas tandatangan seseorang, verifikasi identitas nasabah bank dan pengenalan karakter dokumen untuk kepentingan tertentu suatu perusahaan. Salah satu implementasi pengenalan karakter dokumen telah dilakukan oleh [1] pada bagian customs clearance yang digunakan untuk proses identifikasi jenis dokumen yang berkaitan dengan pengeluaran barang dari beacukai. Pengenalan pola memiliki pengertian proses pengelompokkan data berupa numerik dan simbolik (termasuk citra) yang dilakukan oleh komputer secara otomatis, untuk mengenali objek dalam citra. [2] [3] Berkaitan dengan pengenalan karakter, telah banyak metode yang dikembangkan antara lain [4]menggunakan metode JST untuk pengenalan karakter huruf A, B, C dan D, namun hanya mampu mengenali huruf, tidak dapat mengenali tanda baca. Pengenalan pola huruf tulisan tangan menggunakan ekstraksi ciri geometri dilakukan oleh [5] tetapi metode ini hanya mampu mengenali karakter yang telah dikondisikan. Metode OCR (optical character recognition) merupakan suatu prangkat lunak untuk mengenali karakter yang telah dikembangkan pada tahun 2010 untuk pengenalan text dan memiliki tingkat akurasi sebesar 96.67%. Proses pengenalan pada metode OCR dilakukan dengan menggunakan teknik Template Matching. Berdasarkan pada hasil uji implementasi OCR yang dilakukan oleh [1] pada dokumen ijin alat kesehatan berhasil mengenali sebesar 98,7% untuk font Calibri. Pada kenyataannya terdapat banyak dokumen yang menggunakan font berbeda seperti Times New Roman, Arial dan Lucida Sans

Unicode. Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk mengkaji kemampuan kinerja metode OCR dalam mengenali karakter dokumen dengan menggunakan font Arial, Bahnschrift Condensed, Georgia, Lucida Sans Unicode, Roman, Segoe UI Semibold, dan Times New Roman.

METODE PENELITIAN

Tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. *Tahapan penelitian*

Berdasarkan Gambar 1 terdapat tahapan penelitian yang dilakukan: Akuisisi Citra Merupakan tahap pemindaian dokumen asli dan disimpan dengan format JPG. Tahap ini bertujuan mendapatkan dokumen digital. Contoh hasil akuisisi citra seperti pada Gambar 2.

Gambar 2. *Citra hasil pemindaian*

Gambar 2 adalah citra dokumen hasil pemindaian yang disimpan dalam format JPG, dan merupakan citra RGB. Preprocessing Tahap ini dibutuhkan untuk kelancaran dan kemudahan pemrosesan selanjutnya, terdiri atas tahap konversi citra ke dalam citra keabuan, binerisasi citra dan cropping untuk mendapatkan citra tertentu yang diperlukan. Contoh-contoh hasil preprocessing terdapat pada Gambar 3, Gambar 4, dan Gambar 5.

Gambar 3. *Citra Keabuan*

Gambar 3 merupakan citra hasil konversicitra RGB (Gambar 2) ke dalam citrakeabuan. [6] Tampak warna pada logo berubah dari warna hijau dan kuning menjadi keabuan. (Gambar 3).

Selanjutnyacitra hasil binerisasi disajikan pada Gambar4.

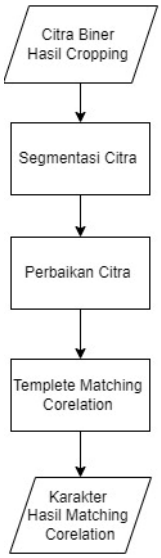
Gambar 4. *Citra hasil binerisasi*

Untuk mendapatkan citra yang lebih jelas dan kontras yang tajam sehingga memudahkan proses lanjutan, dilakukan binerisasi dengan menggunakan thresholding yang hasilnya seperti pada Gambar 4. Tampak citra merupakan citra biner, hanya memiliki warna hitam dan putih. Tahap cropping diperlukan untuk mendapatkan area dokumen tertentu yang dibutuhkan dalam pemrosesan. Gambar 5 adalah contoh citra hasil cropping.

Gambar 5. *Citra hasil cropping*

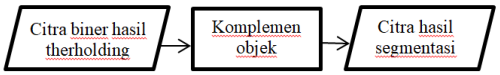
Gambar 5 merupakan hasil cropping citra pada Gambar 4, yang merupakan bagian atau area tertentu dari dokumen yang dibutuhkan.

Pengenalan Karakter



Gambar 6. *Tahapan Pengenalan karakter*

Tahap pengenalan karakter merupakan inti dari metode OCR yaitu proses segmentasi, proses perbaikan kualitas citra dan proses pencocokkan citra. Proses Segmentasi Segmentasi merupakan proses untuk mendapatkan objek yang diinginkan dengan cara memisahkan batas satu objek dengan objek lainnya berdasarkan pengelompokan ketetanggaan piksel [7] [8]. Pada dasarnya proses segmentasi pada penelitian ini adalah proses lanjutan dari binerisasi yang bertujuan mendapatkan karakter-karakter dokumen yang harus dikenali. Secara umum proses segmentasi pada penelitian ini diilustrasikan pada Gambar 7:



Gambar 7. *Tahap Segmentasi*

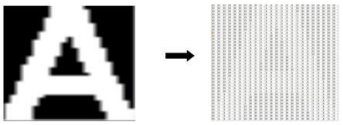
Sedangkan Gambar 8 merupakan contoh hasil segmentasi dari karakter K, E dan S dari kata KESEHATAN.

Gambar 8. *Contoh karakter hasil Segmentasi*

Karakter hasil tahap segmentasi seringkali mengalami penurunan kualitas akibat proses sebelumnya, sehingga perlu tahap lanjutan berupa peningkatan kualitas citra. Proses Perbaikan Kualitas Citra Proses ini dilakukan untuk mendapatkan kualitas citra yang lebih baik, [1] hal ini disebabkan citra hasil pengolahan sebelumnya dapat terjadi penurunan kualitas, sepertinya munculnya noise, hilangnya sebagian citra atau sebaliknya terdapat penambahan citra di bagian lain [7]. Proses perbaikan citra ini dengan menggunakan metode morfologi, yang terdiri antara lain adalah filling merupakan pengisian piksel dan thinning dilakukan agar tepi objek atau karakter memiliki tebal 1 piksel; dilasi dan normalisasi.

Gambar 9. *Ilustrasi proses Filling , Thining dan Dilasi*

Gambar 9 adalah contoh karakter A dengan penurunan kualitas, hal ini ditunjukkan dengan adanya piksel yang kosong (kiri atas) dengan proses filling menjadi terisi, kemudian gambar tengah adalah citra hasil thining, tampak pada gambar sebelah kiri terdapat sisi citra yang tebal berubah menjadi selebar 1 piksel, selanjutnya gambar bagian kanan adalah citra hasil proses dilasi. Tujuan dilasi adalah membuat objek menjadi lebih lebar dan jelas. Proses Ekstraksi Ciri Ekstraksi ciri merupakan proses dalam pengolahan citra untuk mengekstrak ciri atau informasi suatu objek, untuk mendapatkan suatu nilai [1] sehingga dapat membedakan antara satu objek dengan objek lainnya. Contoh citra hasil proses ekstraksi ciri seperti pada Gambar 10.



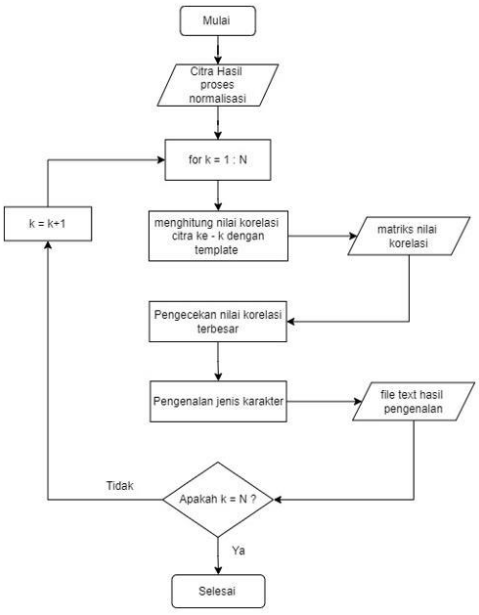
Gambar 10. *Contoh citra hasil Proses Ekstraksi Ciri [7]*

Hasil ekstraksi ciri pada Gambar 10, berupa nilai-nilai yang sesuai dengan ciri tiap piksel dan direpresentasikan dalam bentuk matrik. Selanjutnya matrik hasil ekstraksi ciri digunakan untuk tahap pencocokkan karakter. Tahap Pencocokkan Proses pencocokkan pada OCR dilakukan berdasarkan pada teknik Template Matching dengan menghitung nilai korelasi tiap objek terhadap template hasil pelatihan [9]. Selanjutnya kategori cocok dilakukan pengukuran kemiripan dengan menghitung jarak setiap piksel terhadap piksel citra template [8]. Diagram alir (flowchart) tahap pencocokkan karakter ditampilkan pada Gambar 11. Proses penghitungan nilai korelasi dilakukan dengan menggunakan persamaan:

$$r = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i).(x_{jk} - \bar{x}_j)}{\sqrt{\left[\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 . \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2\right]}}$$

Keterangan:

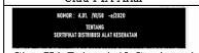
r	: korelasi antara dua buah matriks
Xik	: nilai pixel ke-k dalam matriks i
Xjk	: nilai pixel ke-k dalam matriks j
$\bar{xi}$	: rata-rata nilai pixel matriks i
$\bar{xj}$	: rata-rata nilai pixel matriks j
n	: jumlah pixel dalam satu matriks



Gambar 11. *Flowchart Tahap Pencocokkan*

Karakter pada metode OCR HASIL DAN PEMBAHASAN Tabel 1 merupakan hasil implementasi pengenalan karakter dokumen dengan menggunakan berbagai jenis huruf, antara lain Arial, Georgia, Lucida Sans Unicode, Roman, Segoe UI Semibold, Times New Roman dan Bahnschrift Condensed.

Tabel 1. *Hasil pengenalan karakter 6 jenis huruf*

CitraHasil Cropping Ijin Alkes	Output	Akurasi
 Citra 4 IA Georgia	NOMOR: o1.o1/VI/50 -e/2020 TENTAN G SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	100 %
 Citra 3 IA Lucida Sans Unicode	NOMOR: K.01. /VI/50 —e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	100 %
 Citra 5 IA Segoe UI Semibold	NOMOR: 0101/VI/50 -e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	100 %
 Citra 5 IA Times New Roman	NOMOR : 01.01/V/4 1—e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	100 %
 Citra 4 IA Roman	NOMOR : F.K.01. /VI/50 -e/2020 TENTANG SERTIFIKATUISTRIBUSI ALAT KESEHATAN	98,48 %
 Citra 1 IA Arial	NOMOR: K.01.IVI/50 -e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	95,45%
 Citra 7IA Bahnschrift Condensed	NIJMIJR: K.l]1. IVIIEIJ -eIZl]Zl] TENTANG SER11F|KAT|]|STR|E|JS|ALATKESEIIAT	71.2%

Dari Tabel 1 tampak bahwa implementasi metode OCR sebagian besar berhasil mengenali semua karakter yang diujikan, hal ini ditunjukkan dengan tingkat akurasi 100% (Georgia, Lucida Sans Unicode, Segoe UI Semibold dan Times New Roman), namun beberapa penggunaan jenis huruf lain tidak semua berhasil dikenali. Beberapa karakter penggunaan jenis huruf yang tidak berhasil dikenali antara lain: Arial, terdapat karakter "/" (garis miring) yang berdekatan dengan karakter "V" dapat dikenali sebagai karakter "N". Roman, terdapat kesalahan mengenali karakter "D" sebagai "IJ". Bahnschrift Condensed, untuk jenis huruf ini terdapat banyak kesalahan pengenalan, seperti: Pada tulisan " NOMOR : K.01 /VI/50 e/2020" dikenali sebagai "NIJMIJR: K.l]1. IVIIEIJ -eIZl]Zl]", sehingga dapat dinyatakan karakter "0" dikenali sebagai "IJ", karakter tanda baca "/" dikenali sebagai karakter "I" karakter "V," dikenali sebagai "IE" karakter "2" dikenali sebagai "Z" selanjutnya tulisan "SERTIFIKAT DISTRIBUSI ALAT KESEHATAN" dikenali sebagai "SER11F|KAT|]|STR|E|JS|ALATKESEIIATAN" karakter yang tidak tepat dikenali antara lain karakter "TI" dikenali sebagai "II"; karakter "DI" dikenali sebagai "l]l"; karakter "IB" dikenali sebagai "lE"; karakter "BU" dikenali sebagai "lJ" dan karakter "H" dikenali sebagai "II". Penghitungan nilai akurasi dilakukan dengan menggunakan persamaan:

$$\text{Akurasi} = \frac{\textit{Jumlah karakter benar}}{\textit{Jumlah karakter keseluruhan}} \text{ x } 100\%$$

Nilai akurasi untuk jenis huruf BahnschriftCondensed, diketahui jumlah karakter yang diujikan sebanyak 66 karakter, dan yang berhasil dikenali sejumlah 47 karakter, sehingga akurasi dari pengenalan ini:

$$\text{akurasi } (\%) = \frac{47}{66} \text{ x } 100\% = 71{,}21\%$$

Dengan cara yang sama diperoleh nilai akurasi untuk jenis huruf Arial 95,45% (jumlah karakter yang dikenali 63), untuk jenis huruf Roman 98,48% (jumlah karakter yang dikenali 65) Selain melakukan uji coba berbagai jenis huruf, juga dilakukan pengukuran waktu proses pengenalan karakter untuk setiap penggunaan jenis huruf yang diujikan. Hasil pengukuran waktu proses terhadap jenis huruf yang diujikan tercantum pada Tabel 2.

Tabel 2. *Waktu proses Pengenalan Dokumen*

CitraHasil Cropping Ijin Alkes	Output	Waktu Proses
 Citra 4 IA Georgia	NOMOR: o1.o1/VI/50 -e/2020 TENTAN G SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	1,293 s
 Citra 3 IA Lucida Sans Unicode	NOMOR: K.01. /VI/50 —e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	1,022 s
 Citra 5 IA Segoe UI Semibold	NOMOR: 0101/VI/50 -e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	1,223 s
 Citra 5 IA Times New Roman	NOMOR : 01.01/V/4 1—e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	1,255 s
 Citra 4 IA Roman	NOMOR : F.K.01. /VI/50 -e/2020 TENTANG SERTIFIKATUISTRIBUSI ALAT KESEHATAN	1,706 s
 Citra 1 IA Arial	NOMOR: K.01.IVI/50 -e/2020 TENTANG SERTIFIKAT DISTRIBUSI ALAT KESEHATAN	1,04 s
 Citra 7IA Bahnschrift Condensed	NIJMIJR: K.l]1. IVIIEIJ -eIZl]Zl] TENTANG SER11F|KAT|]|STR|E|JS|ALATKESEIIATAN	1.775 s

Dari Tabel 2 terlihat bahwa rata-rata waktu proses yang dibutuhkan untuk

mengenali karakter yang diujikan sebesar1,33 detik dan dapat dinyatakan waktuterbesar yang dibutuhkan pada pengenalan jenis huruf Bahnschrift Condensed dengan hasil akurasi terkecil.

PENUTUP

Berdasarkan pada hasil pengujian karakter terhadap 7 jenis huruf yang berbeda, metode OCR mampu mengenali100% untuk jenis font Georgia, LucidaSans Unicode, Segoe UI Semibold, Times New Roman sedangkan untuk jenis font Roman tingkat akurasi sebesar 98.48% dan pada jenis font Bahnschrift Condensed 71.21 % dengan rata-rata waktu proses sebesar 1.33 detik. Dengan demikian dapat disimpulkan bahwa metode OCR secara umum mampu mengenali karakter karakter dengan waktu relatif singkat namun demikian masih terbuka kesempatan pengembangan penelitian untuk menigkatkan akurasi yang masih di bawah 95%.

DAFTAR PUSTAKA

[1] S. Susan and E. Lussiana, "SISTEM PENGENALAN DOKUMEN OTOMATIS MENGGUNAKAN OPTICAL CHARACTER RECOGNITION," PETIR, vol. 15, p. 1, 2022.

[2] C. Sandy, P. Rexy and A. Derry, "Penerapan Algoritma Template Matching Dengan Fitur Ekstraksi PCA Untuk Pengenalan Karakter Pada Citra Surat Izin Mengemudi," 2015.

[3] H. Suryo, A. Sugiharto and Sukmawati, "Optical Character Recognition Menggunakan Algoritma template matching correlation," Journal of Informatics and Tecnology, 2021.

[4] A. Dewi and U. Utan, "Aplikasi jaringan syaraf tiruan untuk mengenali tulisan tangan huruf A, B, C, dan D pada jawaban pilihan saol ganda," Jurnal matematika sain dan teknologi, 2011.

[5] M. Herviana, Ilhamsyah and R. Ikhwan, "Aplikasi Pengenalan pola pada huruf tulisan tangan dengan menggunakan jaringan saraf tiruan dengan metode ekstraksi fitur geometri," Jurnal Coding sistem komputer, 2018.

[6] A. Fahmi, M. Asvial and D. Gunawan, "Uplink Resource Allocation Algorithms with Fractional Power Control as Power Constraintsfor OFDMA System," in TELCON, Depok, Indonesia, 2011.

[7] S. Tiwari, S. Mishra, P. Bhatia and P. K. Yadav, "Optical Character Recognition using MATLAB," International Journal of Advanced Research in Electronics and Communication Engineering, Vols. Volume 2,, no. 5, 2013.

[8] Manik and N. Im, "Perancangan program aplikasi pengenalan text menggunakan fuzzy logic," Seminar Nasional Informatika, 2010.

[9] Angaribuan Yohanes, "Menerapkan Jaringan Saraf Tiruan untuk Mengenali Pola Huruf Menggunakan Metode Perceptron," Jurnal Teknik Informatika Unika St.Thomas, Vols. Vol 02 No 02, , 2017.

[10] F. Mohammad, J. Anarase, Shingote, M and P. Ghanwat, "Optical Character Recognition Implementation Using Pattern Matching," International Journal of Computer Science and Information Technologies, 2014.