PENGENALAN TEKS CETAK PADA CITRA TEKS BINER

Iwan Donal Paska Manurung
Achmad Hidayatno
Budi Setiyono

***Abstrak* : Salah satu topik khusus pengolahan citra digital dibidang analisa citra adalah pengenalan teks. Pengenalan teks adalah metoda yang dikembangkan untuk membuat sistem yang mampu memberi pengertian atau deskripsi mengenai objek teks pada citra. Metode dasar yang digunakan pada pengenalan teks adalah inisialisasi awal citra, segmentasi, ekstraksi fitur, klassifikasi dan pengenalan. Metoda tersebut dilakukan secara berurutan dari pemrosesan citra awal hingga menghasilkan keluaran. Keluaran aplikasi pengenalan teks adalah string (deskripsi), bukan citra baru. Pada tugas akhir ini, dibuat suatu aplikasi pengenalan teks tercetak pada citra teks biner. Aplikasi dibuat dengan piranti lunak Matlab 6.5 dari Mathworks, Inc. Ada beberapa pengujian dilakukan terhadap program aplikasi. Pada pengujian dengan menggunakan citra uji dengan variasi gaya tulis Arial, Georgia, Lucida Console, Times New Roman, Tahoma dan verdana dan variasi ukuran 11 hingga 20, ditemukan efisiensi pengenalan adalah 95,8%. Pada pengujian dengan citra uji yang memiliki kecacatan bentuk menghasilkan efisiensi 64,7%. Pada pengujian penggunaan jumlah ciri yang diekstrak terhadap efisiensi pengenalan, bisa diambil kesimpulan bahwa semakin banyak ciri yang diekstrak, semakin baik efisiensi pengenalan program.**

Kata-kunci: citra digital, *optical character recognition (OCR).*

Perkembangan teknologi sekarang ini memungkinkan kita untuk melakukan penelitian terhadap objek citra yang sangat kecil, rumit atau yang tidak beraturan. Dengan dikembangkannya teknologi pengolahan citra digital, hal yang semula tidak mungkin sedikit demi sedikit mulai dapat dimungkinkan. Sebagai contoh, perkembangan teknologi chip sekarang yang sudah mencapai tahap nanochip, telah menghasilkan chip dengan ukuran yang sangat kecil, sehingga mata manusia sulit atau bahkan tidak mungkin untuk melihat tanpa menggunakan alat bantu. Disinilah salah satu peran pengolahan citra secara digital, membantu mata manusia untuk melihat objek yang terdeteksi mata manusia normal.

Selain dibidang teknologi chip, pengolahan citra juga dimamfaatkan sebagai pengenal objek. Hal ini dimungkinkan dengan membuat sejumlah ciri dari sebuah objek yang akan dikenali sehingga diharapkan objek tersebut spesifik dengan objek lain, dalam arti objek tersebut akan terdeteksi berbeda dengan objek lain yang memang berbeda.

Dalam tugas akhir ini akan dibahas mengenai aplikasi pengolahan citra digital dalam bidang pengenalan teks, termasuk ciri-ciri apa saja yang akan dicari dari objek teks sehingga memungkinkan ciri dari 2 objek atau lebih yang sama akan sama dan ciri dari 2 objek atau lebih yang berbeda akan beda, sehingga setiap objek akan spesifik. Jika cara tersebut berhasil maka dapat dibuat aplikasi sebagai bentuk implementasi dari pencirian.

Penelitian ini bertujuan untuk membuat suatu aplikasi yang mampu mengenali teks pada citra teks dan juga memberikan usulan pencirian apa saja yang diharapkan mampu untuk mecirikan bentuk karakter.

Pengenalan Teks Cetak Pada Citra Teks Biner

Pengolahan citra digital adalah suatu penerapan ilmu matematik dibidang perekayasaan 2-dimensi atau lebih sering disebut sebagai perekayasaan matrik. Pengolahan citra digital mempunyai beberapa teknik dasar yang digunakan dalam merekayasa objek citra digital, salah satunya adalah teknik analisa citra. Pengenalan teks merupakan bagian dari teknik analisa citra. Elemen dasar citra yang dianalisa pada teknik pengolahan teks adalah elemen bentuk.

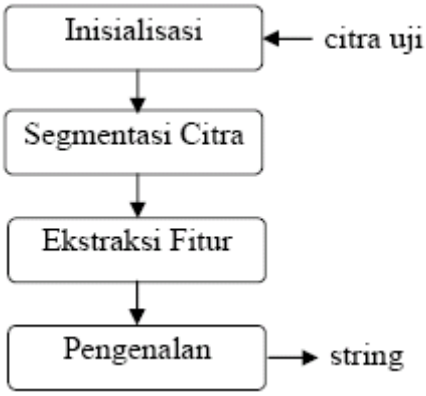
Proses pengenalan teks atau pendeteksian teks pada media gambar berarti sistem dapat mengenali/membaca suatu objek teks pada media gambar dan menulisnya ke bentuk string. String merupakan defenisi karakter berdasarkan kode ASCII. String sering kita temui pada piranti lunak pengolah kata atau pada program penyunting teks. Teks pada media citra adalah bukan merupakan string, karena teks tersebut bukan merupakan perwakilan dari kode – kode ASCII, tapi merupakan objek yang terbentuk dari susunan piksel.

Achmad Hidayatno dan Budi Setiyono adalah dosen di jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro Semarang. Jl. Prof. Sudharto, S.H. Tembalang, Semarang 50275
Iwan D P M adalah mahasiswa di jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro

Metode dasar yang digunakan untuk mengenali objek tersebut sebagai teks adalah dengan mencirikan bentuk dari karakter tersebut masing – masing. Dengan membuat ciri yang unik untuk setiap karakter maka dapat dibuat algoritma yang sesuai untuk mengenali karakter dengan baik.

Perancangan Perangkat Lunak

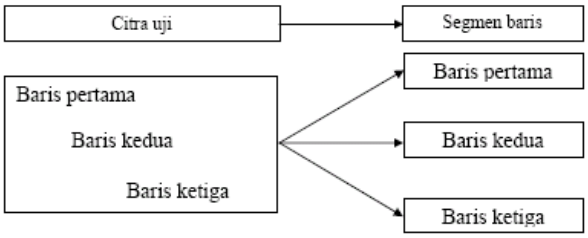
Aplikasi perangkat lunak pengenalan teks pada citra teks biner ini terdiri dari 4 proses umum, yaitu proses inisialisasi, segmentasi, ekstraksi fitur, dan pengenalan. Keluaran dari aplikasi ini adalah teks string.



Gambar 1. Diagram alir sistem pengenalan teks

Proses **inisialisasi** bertanggung jawab pada hal – hal yang perlu dilakukan untuk menampilkan citra uji dan mempersiapkan citra uji untuk diproses segmentasi. Proses inisialisasi akan menetapkan parameter awal dan memastikan kelayakan citra uji untuk proses segmentasi.

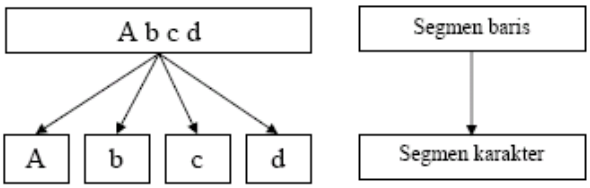
Proses **segmentasi citra** bertanggung jawab atas semua proses yang berhubungan dengan pembagian, pemotongan atau pemisahan citra uji menjadi segmen – segmen yang lebih sederhana. Proses segmentasi dilakukan tanpa mengurangi esensi informasi yang ada pada citra uji, meskipun citra uji mengalami pemotongan atau pembagian. Proses segmentasi akan berakhir jika citra awal telah dibagi menjadi segmen – segmen kecil yang terdiri dari 1 objek karakter per segmen kecil.



Gambar 2. Ilustrasi proses segmentasi baris

Segmentasi citra yang dilakukan pada penelitian ini dibagi dalam 2 bagian, yaitu segmentasi baris dan segmentasi karakter. Segmentasi baris bertujuan untuk membagi citra berdasarkan baris teks pada citra, sehingga citra awal akan dibagi menjadi beberapa segmen baris teks sejumlah barisan teks yang ada pada citra awal. Citra baru hasil segmentasi baris disebut sebagai segmen baris.

Segmentasi karakter bertujuan untuk membagi tiap – tiap segmen baris menjadi segmen yang lebih kecil berdasarkan jumlah objek karakter yang ada pada segmen baris. Segmen hasil segmentasi karakter disebut dengan segmen karakter. Jumlah segmen karakter yang dihasilkan dari satu segmen baris harus sama dengan jumlah karakter yang ada pada segmen baris tersebut.



Gambar 3. Ilustrasi proses segmentasi karakter

Proses **ekstraksi fitur** merupakan proses untuk mencari, menandai, dan menyimpan semua fitur dari segmen karakter. Fitur adalah semua informasi atau karakteristik bentuk dari segmen yang bisa dijadikan sebagai tanda pengenal dari bentuk segmen karakter tersebut, oleh karena itu fitur harus unik untuk tiap bentuk karakter.

Fitur yang akan diekstrak dari segmen luasan karakter ada 2 bagian, yaitu pola segmen dan ciri segmen. Fitur ini akan mewakili objek yang akan dikenali. Fitur yang akan diambil dari tiap segmen akan ada beberapa, karena satu fitur belum tentu mewakili satu karakter dengan unik. Tiap segmen akan dicari beberapa fitur dan dikombinasikan menjadi satu kesatuan fitur yang akan mewakili satu jenis karakter.

Pertama, Pola segmen deretan nilai yang mewakili suatu runtun bentuk. Pola ini dibutuhkan untuk mencari kemiripan antara pola objek yang akan dikenali dengan pola pembandingnya. Pola pembanding adalah pola yang disimpan pada program, digunakan sebagai bahan perbandingan untuk objek yang akan dikenali.

Pola yang akan diekstrak dari segmen ada 2 jenis pola, yaitu pola jumlah perpotongan pada tengah dan pola jarak tepi. Kedua pola tersebut akan digabungkan menjadi satu kesatuan pola yang disebut sebagai fitur pola segmen. Pola inilah yang akan dibandingkan dengan pola objek lain untuk mencari kemiripan antar kedua objek.

Kedua, Ciri segmen adalah pengambilan informasi atau tanda khusus/unik dari segmen. Pemilihan ciri ini dilakukan dengan sistem coba – coba, karena tidak selalu sebuah ciri mampu membuat bentuk karakter spesifik.

Sebaiknya, semakin banyak ciri maka akan semakin baik pula pengenalan yang dapat dilakukan sistem terhadap objek yang akan dikenali. Dalam tugas akhir ini, diusulkan 9 ciri yang akan dicari dari tiap segmen. Dibawah ini disajikan tabel pencirian yang akan diekstrak untuk tiap segmen karakter.

Tabel 1. Tabel pencirian untuk tiap segmen karakter

no	Nama Ciri
1	ciri jumlah komponen terhubung objek
2	ciri euler
3	ciri objek terbuka ke atas
4	ciri objek terbuka ke bawah
5	ciri objek terbuka ke kanan
6	ciri sudut kiri atas
7	ciri sudut kanan atas
8	ciri sudut kiri bawah
9	ciri sudut kanan bawah

Proses **pengenalan** merupakan proses dimana semua fitur yang diekstraksi dari segmen uji dibandingkan dengan segmen uji yang tesimpan pada program, kemudian diambil keputusan fitur ekstraksi tersebut merupakan perwakilan dari karakter apa. Pada proses pengenalan, segmen karakter uji tidak dibutuhkan lagi, karena yang diperlukan hanya fitur hasil ekstraksi.

Hasil Pengujian Dan Pembahasan

Perangkat lunak yang dibuat pada penelitian ini akan dilakukan pengujian untuk mengetahui seberapa jauh perangkat lunak mampu mengenali sampel uji yang diberikan. Dari pengujian juga dapat diambil kesimpulan dari penelitian ini, juga hal yang perlu dilakukan kedepan untuk perbaikan kualitas maupun perluasan bidang penelitian.

Pada penelitian ini, dilakukan 4 pengujian pada perangkat lunak, yaitu pengujian terhadap objek karakter tanpa kecacatan, pengujian pengaruh penggunaan jumlah ciri, pengujian terhadap objek karakter ukuran kecil dan pengujian terhadap objek karakter dengan kecacatan.

Pengujian pertama, pengujian perangkat lunak menggunakan objek karakter tanpa kecacatan bertujuan untuk mengetahui total efisiensi perangkat lunak.

Citra uji dibentuk menggunakan program MsPaint pada Microsoft Windows. Karakter uji diketik pada program tersebut, kemudian berkas disimpan dengan ekstensi *.bmp dengan 2 level keabuan (biner).

Citra uji divariasi berdasarkan jenis gaya penulisan dan ukuran. Jenis gaya tulis yang digunakan pada pengujian ini adalah Arial, Times New Roman, Tahoma dan Verdana, ukuran objek teks juga divariasi antara skala 11, 12, 14 dan 16.

Total sampel yang diuji ada 806 sampel uji, dari semua variasi gaya tulis dan ukuran berikut semua huruf besar dan kecil.

Dibawah ini disajikan tabel hasil pengujian :

Tabel 2. Tabel hasil pengujian tanpa kecacatan

	I	II	III	IV	V	VI	/size
11	100	90,4	86.5	100	98.0	98.0	95.5
12	100	90,4	90.4	100	98.0	98.0	95.8
14	100	88,5	88.5	100	98.0	98.0	96.1
16	100	94,2	90.3	100	98.0	98.0	96.7
18	100	86,5	92.2	100	98.0	98.0	95.8
20	100	86,5	86.5	100	98.0	98.0	95.2
/font	100	89.4	89.4	100	98.0	98.0	

I=arial; II=georgia, III=lucida console, IV=Times, V=Tahoma, VI=verdana.

Jika dari tabel akan dihitung total efisiensi untuk semua jenis gaya tulis berikut variasi ukurannya, dihasilkan total efisiensi adalah 95.8% dari total sampel uji 1872 sampel.

Pengujian kedua, pengujian ini dilakukan untuk mengetahui pengaruh penggunaan banyaknya ciri yang diekstrak dari segmen uji terhadap efisiensi pengenalan. Pengujian dilakukan dengan variasi gaya tulis Arial dan times New Roman dengan ukuran skala 14. Pembentukan sampel dilakukan persis dengan pembentukan citra sampel pada pengujian tanpa kecacatan.

Banyaknya ciri yang diekstrak dari segmen adalah 9 ciri, melalui pengujian ini akan diketahui pengaruh banyak ciri terhadapa efisiensi program. Pengujian dilakukan dengan variasi jumlah ciri, yaitu penggunaan banyak ciri 1, 3, 5, 7 dan 9. Jumlah sampel total akan menjadi 104 sampel.

Tabel 3. Tabel hasil pengujian pengaruh banyak ciri

Banyak ciri	Benar	Efisiensi %	Ciri yang Digunakan*
1	5	4,8	1
3	7	6,7	1,2,3
5	21	20,2	1,2,3,4
7	47	45,2	1,2,3,4,5,6,7
9	66	63,5	1,2,3,4,5,6,7,8,9

* disesuaikan dengan tabel 1

Dari tabel diatas dapat dilihat bahwa terjadi peningkatan hasil efisiensi pengujian jika banyak ciri yang diekstrak dari segmen diperbanyak. Hal ini menunjukkan bahwa semakin banyak ciri yang diekstrak maka semakin baik efisiensi program pengenalan teks.

Pengujian ketiga, Pengujian terhadap citra uji dengan objek karakter ukuran kecil dilakukan untuk mengetahui apakah ada perbedaan efisiensi pengenalan jika objek yang dikenali dibuat dengan ukuran kecil, lebih kecil dari variasi ukuran pada pengujian tanpa kecacatan. Sampel pada pengujian ini dibentuk dengan cara yang sama dengan pengujian tanpa kecacatan, hanya saja variasi ukuran yang digunakan adalah skala 10. Variasi gaya tulis yang digunakan, yaitu Arial, Georgia, Lucida Console, Times New Roman, Tahoma dan Verdana. Total sampel adalah 312 sampel uji.

Pada pengujian ditemukan bahwa, ada 28 sampel yang gagal dikenali dari 312 semesta sampel, sehingga efisiensi menjadi 91%.

Pengujian keempat, Pengujian terhadap objek teks dengan kecacatan acak yang dimaksud disini adalah objek teks yang diuji memiliki kecacatan yang tidak teratur atau tidak bisa diketahui pola kecacatan objek tersebut. Pengujian ini dilakukan untuk mengetahui sejauh mana program ini mampu mengenali objek dengan kecacatan.

Sampel dengan objek yang memiliki kecacatan dapat dibentuk cara mengetik karakter sampel yang akan diuji pada program pengolah kata, kemudian mengubah berkas tersebut menjadi berkas *.pdf, kemudian buka berkas *.pdf dengan perbesaran 100%, salin (*copy*) objek karakter dengan *snapshot tool* dan ditulis ke program penyunting gambar MsPaint, simpan berkas dengan format *.bmp, biner.

Dengan cara tersebut diatas, objek akan mengalami degradasi bentuk karena proses salin tulis objek karakter dari berkas *.pdf ke penyunting gambar dan proses pengubahan ke format biner.

Pada pengujian ini dilakukan dengan variasi gaya tulis Arial, Georgia, Lucinda console, Times New Roman, Tahoma dan Verdana dengan variasi ukuran 10 (mewakili objek kecil), 12 (ukuran sedang) dan 16 (ukuran besar). Total sampel adalah 936 sampel uji.

Tabel 4. Tabel hasil pengujian objek cacat

	I	II	III	IV	V	VI	
10	67.3	53.8	55.7	44.2	67.3	78.8	61.1
12	82.6	36.5	51.9	61.5	76.9	61.5	61.8
16	94.2	34.6	73.0	44.2	90.3	92.3	71.4
	81.3	41.6	60.2	49.9	78.1	77.53	%

I,II,III,IV,V,VI lihat di tabel 2

Dari tabel diatas jika inigin dihitung efisiensi total dari semua sampel yang diuji, maka total efisiensi untuk pengujian terhadap objek cacat adalah 64.7% dari total sampel 936 sampel uji.

Hasil pengujian yang ditemukan pada pengujian terhadap citra uji tanpa kecacatan dan pengujian terhadap citra dengan kecacatan berbeda signifikan, dimana perbandingan perbedaan efisiensi terpaut hingga 31.1%. Sehingga bisa disimpulkan bahwa program ini bekerja efektif untuk citra dengan objek teks tanpa kecacatan. Untuk penggunaan pada citra dengan kecacatan objek teks tidak disarankan karena masih memiliki banyak kekurangan, masih perlu dilakukan pembenahan.

Penutup

Dari hasil penelitian dan pengujian terhadap perangkat lunak, dapat diambil beberapa kesimpulan sebagai berikut : *pertama*, Aplikasi menghasilkan tingkat keberhasilan pengenalan 95,8% untuk variasi tipe huruf Arial, Georgia, Lucida Console, Times New Roman, Tahoma dan Verdana dengan variasi skala ukuran dari 11 hingga 20.

Kedua, Berdasarkan pengujian terhadap tipe huruf, tipe huruf Arial dan Times New Roman dapat dikenali dengan baik, 100%.

Ketiga, tingkat keberhasilan pengenalan paling rendah yaitu pada tipe huruf Georgia dan Lucida Console, 89,4%.

Keempat, pada pengujian dengan menggunakan 1 ciri, yaitu ciri objek terpisah, menghasilkan tingkat keberhasilan 4,8%.

Kelima, pada pengujian dengan menggunakan keseluruhan ciri menghasilkan tingkat keberhasilan pengenalan 63,5%.

Keenam, program memiliki tingkat keberhasilan 64,7% untuk citra uji dengan objek karakter yang memiliki degradasi bentuk.

Beberapa hal yang perlu diperhatikan untuk penelitian ke depan adalah sebagai berikut : *pertama*, perlu dilakukan penelitian terhadap citra teks dengan kasus lain, seperti keberadaan teks yang miring atau adanya beberapa latar berbeda dalam citra teks.

Kedua, perlu dilakukan penelitian lebih lanjut untuk meningkatkan kualitas pengenalan pada citra dengan kecacatan bentuk objek teks maupun adanya derau.

Ketiga, perlu dilakukan penelitian terhadap pencirian yang lebih komprehensif dan kompleks, sehingga hasil pendeteksian dapat mencakup variasi gaya penulisan yang lebih variatif

DAFTAR PUSTAKA

Jain, A.K., *Fundamentals of Digital Image Processing*, Prentice Hall, 1989.

Munir, R., *Pengolahan Citra Digital dengan Pendekatan Algoritmik*, Informatika, Bandung, 1992.

Alata, Mohanad. Mohammad Al-Shabi. *Text Detection and Character Recognition Using Fuzzy Image Processing*. Journal of ELECTRICAL ENGINEERING 57, NO. 4, 2006.

Datta, Soumita. K B Roy Choudhuri. Ashish Gangguli. *Text Extraction System*.

Hanselman, Duane dan Bruce Littlefield, *Matlab Bahasa Komputasi Teknis*, Andi, Yogyakarta, 2000.

Hunt, Brian R. Ronald L Lipsman. Jonathan M Rosenberg., *A Guide to MATLAB for Beginners and Experienced Users*, Cambrigde University Press, Cambridge, 2001.

Cheriet, Mohamed, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen, *Character Recognition System : A Guide for Students and Practioners*, John Wiley & Sons, Inc. New Jersey, 2007.

Gonzalez, Rafael C., Richard E. Woods, *Digital Image Processing second edition*, Prentice Hall. New Jersey, 2001.

Jahne, Bernd. *Digital Image Processing*. Springer-Verlag. Berlin, 2005

Russ, John C.*The Image Processing Handbook*. Taylor & Francis Group, 2006

Matlab Image Processing Toolbox, http://www.mathworks.com.

Iwan Donal Paska Manurung (L2F 003 510)Lahir di Porsea, 8 April 1985. Saat ini masih menjadi Mahasiswa S1 di Jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro Semarang dengan konsentrasi Teknik Telekomunikasi.

Mengetahui dan Mengesahkan :

Pembimbing I

Achmad Hidayatno, S.T., M.T.
NIP. 132 137 933
Tanggal :

Pembimbing II

Budi Setyono, S.T, M.T.
NIP. 132 283 184
Tanggal :