Optical Character Recognation Menggunakan Partisi Citra

Ardi Sanjaya
Teknik Informatika, Fakultas Teknik, UNP Kediri
Jl.K.H. Ahmad Dahlan No.76 Kediri
dersky@gmail.com

Abstrak **- Penelitian ini mencoba memberikan alternatif baru untuk proses pengenalan karakter pada data citra. Proses yang digunakan yaitu memisahkan paragraf dengan teknik *cropping*. Kemudian memisahkan baris pada paragraf dengan cara membaca area kosong secara vertikal. Dilanjutkan dengan memisahkan karakter pada masing-masing baris dengan cara membaca area ksosong secara horizontal. Setelah karakter berhasil dipisah, masing-masing karakter dipartisi menjadi 64 bagian dan kemudian diambil nilai luas piksel tiap-tiap partisinya dan disimpan sebagai data *training*. Untuk tahap pengenalan karakter, menggunakan pencocokan data testing terhadap data training dan dicari selisih jarak terpendek dengan euclidean distance. Font yang digunakan sebagai data taraining dan testing yaitu arial, calibri dan times new romans ukuran 8pt, 10pt, 12pt dan 14pt. Didapati hasil bahwa font times new romans memiliki akurasi rendah 23,44% pada ukuran 10pt dan akurasi tertinggi pada font arial ukuran 14pt.**

Kata **Kunci—OCR, *Optical Character Recognation*, Segemntasi Citra.**

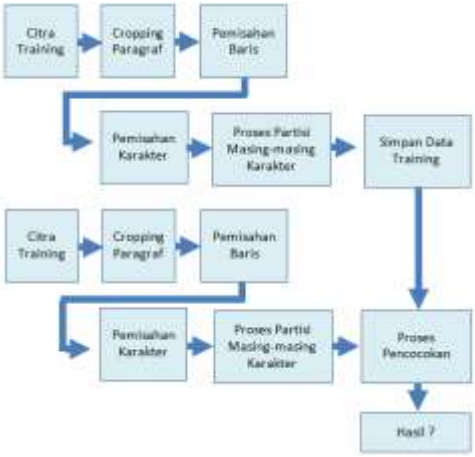
I. PENDAHULUAN

Ilmu pengetahuan berkembang pesat seiring dengan perkembangan teknologi seperti saat ini. Salah satunya adalah ilmu pengetahuan yang berkaitan dengan pengenalan pola. Pengenalan pola (*pattern recognation*) merupakan salah satu ilmu untuk mengklasifikasikan atau menggambarkan sesuatu berdasarkan pengukuran kuantitatif fitur (ciri) atau sifat utama suatu subyek.

Optical Character Recognation (OCR) merupakan salah satu area studi dalam bidang pengenalan pola yang menarik untuk diteliti. OCR adalah proses konversi karakter pada data citra digital menjadi bentuk teks [1]. Secara umum terdapat dua hal utama yang mempengaruhi proses OCR yaitu mekanisme ekstraksi ciri dan mekanisme pengenalan. Mekanisme ekstraksi ciri dilakukan untuk mendapatkan ciri atau identitas dari suatu karakter atau huruf. Proses pengenalan dilakukan setelah mekanisme ekstraksi ciri. Proses pengenalan bertujuan untuk mencocokkan pola huruf yang berasal dari inputan dengan pola yang ada dalam basis pengetahuan.

Raden Sofian dan Irfan Maliki dalam penelitiannya Perbandingan Algoritma Template Matching dan Feature Extraction Pada Optical Character Recognation [2], mengemukakan bahwa algoritma feature extraction lebih unggul dalam hal akurasi, pengembangan dan waktu dibandingkan template matching pada OCR. Sedangkan Suryo Hartanto dkk dalam penelitiannya Optical Character Recognation menggunakan Algoritma Template Matching Correlation [3] meneliti bahwa algoritma template matching correlation cukup efektif untuk pengenalan karakter huruf cetak dengan rata-rata tingkat keberhasilan 92,90%.

Pada penelitian ini mencoba membuat alternatif baru pada proses ekstraksi ciri yaitu dengan memisahkan baris dan karakter. Kemudian masing-masing karakter dilakukan proses partisi dan diambil ciri berupa data jumlah piksel dari masing-masing partisi tiap karakternya dan disimpan sebagai data training. Untuk tahap pengenalan karakter, menggunakan pencocokan data testing terhadap data training dan dicari selisih jarak terpendek dengan menghitung akar dari kuadrat perbedaan 2 vector atau lebih dikenal dengan *euclidean distance*.



Gambar 1. Alur penelitian

Berdasarkan uraian diatas maka di dapat rumuskan permasalahan bagaimana membuat alternatif baru untuk mengenali karakter (OCR) pada data citra.

Adapun batasan-batasan dalam penelitian ini adalah :

1. Data training dan data testing berupa data teks paragraf tunggal dalam bentuk citra dan berformat hitam putih bitmap.
2. Font yang digunakan adalah Arial, Calibri dan Times New Romans dengan ukuran tinggi font 8, 10, 12 dan 14.
3. Partisi citra karakter yang digunakan adalah ukuran 8x8.
4. Data training tersimpan dalam database flat file (menggunakan file .ini)
5. Aplikasi yang digunakan untuk melakukan pengolahan dan perhitungan adalah Delphi 7.0.

Tujuan dari penelitian ini secara khusus adalah memberikan alternatif baru dengan menggunakan partisi citra dan euclidean distance untuk mengenali karakter. Sedangkan tujuan secara umum adalah memperkaya konten pendidikan dan ilmu pengetahuan.

II. LANDASAN TEORI

A. Pengolahan Citra

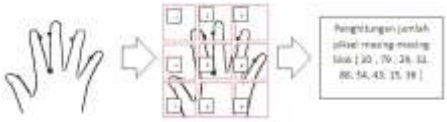
Meskipun sebuah citra kaya informasi, namun seringkali citra yang kita miliki mengalami penurunan mutu (degradasi), misalnya mengandung cacat atau derau (*noise*), warnanya terlalu kontras, kurang tajam, kabur (*blurring*), dan sebagainya [4]. Tentu saja citra semacam ini menjadi lebih sulit diinterpretasi karena informasi yang disampaikan oleh citra tersebut menjadi berkurang. Agar citra yang mengalami gangguan mudah diinterpretasi (baik oleh manusia maupun mesin), maka citra tersebut perlu dimanipulasi menjadi citra lain yang kualitasnya lebih baik. Bidang studi yang menyangkut hal ini adalah pengolahan citra (*image processing*).

Umumnya, operasi-operasi pada pengolahan citra diterapkan pada citra bila :

1. Perbaikan atau memodifikasi citra perlu dilakukan untuk meningkatkan kualitas penampakan atau untuk menonjolkan beberapa aspek informasi yang terkandung di dalam citra.
2. Elemen di dalam citra perlu dikelompokkan, dicocokkan, atau diukur.
3. Sebagian citra perlu digabung dengan bagian citra yang lain.

B. Partisi Citra

Partisi citra adalah membagi citra menjadi beberapa blok dimana masing-masing blok atau bagian memiliki ukuran yang sama besar [5]. Tujuan dilakukan partisi terhadap citra adalah menghitung jumlah piksel masing-masing blok pada data *training* dan data *testing*. Kemudian masing-masing blok pada data *training* dan *testing* dicari selisih yang terpendek menggunakan *Euclidean*.



Gambar 2. Contoh partisi citra

C. Pengenalan Pola

Pengenalan Pola mengelompokkan data numerik dan simbolik (termasuk citra) secara otomatis oleh mesin (dalam hal ini komputer) [6]. Tujuan pengelompokan adalah untuk mengenali suatu objek di dalam citra. Manusia bisa mengenali objek yang dilihatnya karena otak manusia telah belajar mengklasifikasi objek-objek di alam sehingga mampu membedakan suatu objek dengan objek lainnya. Kemampuan sistem visual manusia inilah yang dicoba ditiru oleh mesin. Komputer menerima masukan berupa citra objek yang akan diidentifikasi, memproses

citra tersebut, dan memberikan keluaran berupa deskripsi objek di dalam citra.

D. Euclidean Distance

Jarak *Euclidean* (*Euclidean distance*) yaitu metrika yang digunakan untuk menghitung kesamaan 2 vector. Jarak *Euclidean* menghitung akar dari kuadrat perbedaan 2 vector (*root of squere difference betweens 2 vectors*) [7].

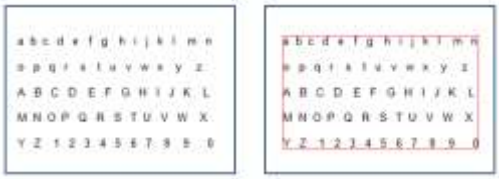
$$d_e = \sqrt{\sum_{k=1}^{m} (fd_{i,k} - k_j)^2}$$

Keterangan :
de : jarak euclidean
fdi : bobot citra pelatihan
kj : data bobot citra test
m : jumlah data pelatihan

III. HASIL DAN PEMBAHASAN

A. Cropping Paragraf

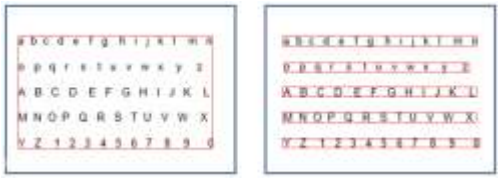
Proses awal dari sistem yang diteliti adalah *cropping* paragraf bertujuan untuk mendapatkan mengenali batas terluar dari obyek teks dalam citra secara keseluruhan.



Gambar 3. *Cropping* paragraf

B. Pemisahan Baris

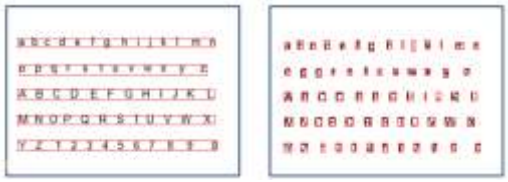
Tahap selanjutnya adalah pemisahan baris. Metode yang digunakan adalah dengan membaca area kosong (piksel putih) dengan arah vertikal atau yang dikenal dengan jarak antar baris.



Gambar 4. Pemisahan baris

C. Pemisahan Karakter

Setelah baris dapat dipisahkan, selanjutnya adalah tahap memisahkan karakter dengan cara membaca area kosong (warna putih) dengan arah horizontal pada masing-masing baris.



Gambar 5. Pemisahan karakter

D. Partisi Karakter

Setelah karakter dapat dipisahkan, proses selanjutnya adalah mempartisi karakter menjadi 64 bagian. Kemudian pada masing-masing partisi dihitung luas piksel dan disimpan sebagai data *training*. Karakter yang digunakan ada masing-masing data adalah a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Tabel 1. Data Training

No	Nama File	Font	Ukuran
1	Arial_8.bmp	Arial	8
2	Arial_10.bmp	Arial	10
3	Arial_12.bmp	Arial	12
4	Arial_14.bmp	Arial	14
5	Calibri_8.bmp	Calibri	8
6	Calibri_10.bmp	Calibri	10
7	Calibri_12.bmp	Calibri	12
8	Calibri_14.bmp	Calibri	14
9	Times_8.bmp	Times New Roman	8
10	Times_10.bmp	Times New Roman	10
11	Times_12.bmp	Times New Roman	12
12	Times_14.bmp	Times New Roman	14

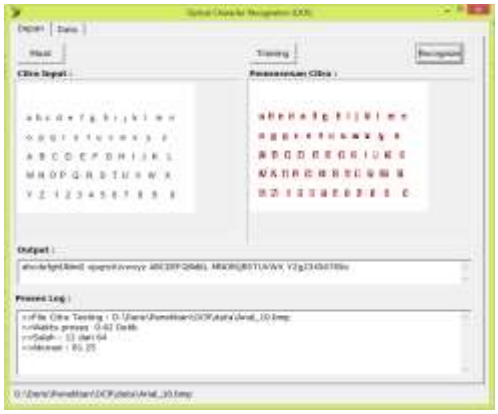
E. Testing

Proses pengujian atau testing dilakukan dengan cara menguji masing-masing data citra terhadap seluruh data acuan. Didapat hasil sebagai berikut :

Tabel 2. Hasil Percobaan

No	Citra Testing	Waktu (s)	Salah	Akurasi
1	Arial_8.bmp	0.61	11	82.81
2	Arial_10.bmp	0.63	12	81.25
3	Arial_12.bmp	0.66	6	90.63
4	Arial_14.bmp	0.66	3	95.31
5	Calibri_8.bmp	0.61	14	78.13
6	Calibri_10.bmp	0.64	8	87.50
7	Calibri_12.bmp	0.64	33	48.44
8	Calibri_14.bmp	0.64	29	54.69

9	Times_8.bmp	0.61	46	28.13
10	Times_10.bmp	0.61	49	23.44
11	Times_12.bmp	0.64	45	29.69
12	Times_14.bmp	0.64	45	29.69



Gambar 6. Tampilan aplikasi

IV. KESIMPULAN DAN SARAN

Berdasarkan dari hasil percobaan dapat disimpulkan bahwa :

1. Metode ini memiliki kekurangan tidak bisa mengenali 2 karakter yang tergabung menjadi 1.
2. Metode partisi citra menghasilkan akurasi yang rendah pada jenis font Times New Roman
3. Penggunaan partisi 8x8 berdampak proses pencocokan menjadi lebih sensitif dan memiliki toleransi jarak *euclidean* yang kecil.

Kedepannya, harus ada penelitian lanjutan terkait pengaruh nilai jumlah partisi yang digunakan.

DAFTAR PUSTAKA

[1] Sukhpreet Singh, "Optical Charakter Recognation Techniques : A Survey", Journal of emerging Trends in Computing and information Sciences Vol 04 No 6 June 2013, ISSN 2079-8407

[2] Sofian, R, Maliki, I, "Perbandingan Algoritma Template Matching dan Feature Extraction Pada Cotical Character Recognation", Jurnal Komputer dan Informatika, Edisi I Vol 1 Maret 2012

[3] Hartanto, S, dkk, "Optical Character Recognation Menggunakan Algoritma Template Mathcing Correlation" , Journal of Informatics and Technlogy, Vol 1 tahun 2012, p 11-12

[4] Munir, R, "Pengolahan Citra Digital", Informatika Bandung, 2004

[5] Sanjaya, Ardi, "Identifikasi Personal Berdasarkan Bentuk Tangan", Prosiding Seminar Nasional Teknologi Informasi dan Multimedia 2014, ISSN 2302-3805

[6] Gonydjaja, R, "Pengantar Pengolahan Citra", Tersedia : http://aqwam.staff.jak-stik.ac.id/files/30.-pengolahan-citra%5B3%5D.pdf, diakses 20 Desember 2014

[7] Darma Putra, I., *Sistem Biometrika*, Andi Publishing, Yogyakarta, 2009