

ABSTRAK

Denny Setiawan:

Tugas Akhir

Optical Character Recognition dengan Metode Gabungan *Template Matching* dan *Feature Extraction*.

Saat ini banyak orang telah menggunakan *scanner* untuk melakukan *scan* pada dokumen. Akan tetapi dokumen hasil *scan* ini dalam bentuk format gambar sehingga tidak dapat dilakukan proses penyuntingan pada dokumen tersebut. Untuk mengubah dokumen tersebut menjadi format teks maka dilakukan proses pengenalan karakter. Metode yang digunakan untuk pengenalan karakter antara lain metode *template matching* dan *feature extraction*.

Tujuan dari Tugas Akhir ini adalah membuat program pengenalan karakter dengan metode gabungan *template matching* dan *feature extraction*. Proses pengenalan karakter ini didukung dengan proses *noise removal* dan *skew angle detection* sehingga proses pengenalan karakter dapat menjadi lebih akurat.

Hasil pengujian menunjukkan bahwa pengenalan karakter dengan menggunakan dokumen hasil *scan* dengan *color mode text enhanced* pada jenis *font* times new roman sebesar 96,05%, arial sebesar 98,42% dan verdana sebesar 99,34%. Untuk hasil pengujian dengan menggunakan dokumen hasil *scan* dengan *color mode grayscale* pada jenis *font* times new roman sebesar 97,33%, arial sebesar 99,25% dan verdana sebesar 99,41%.

Kata kunci:

Pengenalan karakter, *Feature Extraction*, *Template Matching*.

ABSTRACT

Denny Setiawan

Final Assignment

Optical Character Recognition with Combination of Template Matching and Feature Extraction Methods.

Today many people use scanner to scan documents. But the scanned document is in image format so the document can not be edited. To change the scanned documents to text format, character recognition needs to perform. Methods that used to recognize characters are template matching and feature extraction.

The main goal of this final assignment is to create a software that can be used to recognize characters using template matching and feature extraction methods. The character recognition process is supported by noise removal process and skew angle detection process, so the character recognition process can be more accurate.

The result of testing shows that character recognition process using scanned documents with color mode text enhanced for font *times new roman* was 96.05%, *arial* was 98.42% and *verdana* was 99.34%. The result of testing using scanned documents with color mode grayscale for font *times new roman* was 97.33%, *arial* was 99.25%, and *verdana* was 99.41%.

Keywords:

Character Recognition, Feature Extraction, Template Matching.

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN	ii
LEMBAR PENGALIHAN HAK ATAS KEKAYAAN INTELEKTUAL	iii
KATA PENGANTAR	iv
ABSTRAK	v
ABSTRACT	vi
DAFTAR ISI	vii
DAFTAR GAMBAR	xi
DAFTAR TABEL	xv
DAFTAR LAMPIRAN	xvii
1. PENDAHULUAN	1
1.1. Latar Belakang dan Permasalahan	1
1.2. Perumusan Masalah	1
1.3. Ruang Lingkup	2
1.4. Tujuan Tugas Akhir	2
1.5. Metodologi Penelitian	2
1.5.1. Pengumpulan Data	2
1.5.2. Studi Literatur	3
1.5.3. Pembuatan Program	3
1.5.4. Pengujian Program	3
1.6. Sistematika Penulisan	3
2. TEORI PENUNJANG	5
2.1. <i>Image Preprocessing</i>	5
2.1.1. <i>Thresholding</i>	5
2.1.2. <i>Skew Angle Detection</i>	6
2.1.3. Rotasi	7
2.1.4. <i>Pepper Noise Removal</i>	8
2.1.5. <i>Salt Noise Removal</i>	9
2.2. Segmentasi Karakter	9
2.3. <i>Character Recognition</i>	11
2.3.1. <i>Feature Extraction</i>	11
2.3.2. <i>Template Matching</i>	11
2.4. Algoritma Garis Bresenham	11

3.	DESAIN SISTEM.....	14
3.1.	<i>Image Preprocessing</i>	15
3.1.1.	<i>Thresholding</i>	15
3.1.2.	<i>Noise Removal 1</i>	16
3.1.3.	<i>Skew Angle Detection</i>	17
3.1.3.1.	Menentukan Koordinat Daerah Teks	18
3.1.3.2.	Menghitung Jumlah Baris Pada Dokumen	18
3.1.3.3.	Menentukan Koordinat <i>Pixel</i> Hitam Bagian Bawah Karakter	19
3.1.3.4.	Menghitung Derajat Kemiringan Satu Baris Teks	19
3.1.3.5.	Menghitung Rata-Rata Derajat Kemiringan Baris Teks.....	19
3.1.4.	Rotasi	20
3.1.4.1.	Menentukan Koordinat Titik Acuan Rotasi	20
3.1.4.2.	Mengalikan Seluruh Koordinat <i>Pixel</i> Hitam Dengan Matriks Transformasi Rotasi	21
3.1.5.	<i>Noise Removal 2</i>	21
3.1.5.1.	<i>Pepper Noise Removal</i>	21
3.1.5.2.	<i>Salt Noise Removal</i>	22
3.2.	<i>Image Document Analysis</i>	22
3.2.1.	Segmentasi Baris.....	23
3.2.2.	Segmentasi Karakter	23
3.2.2.1.	Melakukan <i>Scan</i> Secara Vertikal	24
3.2.2.2.	<i>Connected Component Labeling</i>	24
3.2.2.3.	Memotong Pada Kolom Yang Memiliki Jumlah <i>Pixel</i> Hitam Terkecil	24
3.3.	<i>Character Recognition</i>	25
3.3.1.	<i>Feature Extraction</i>	26
3.3.1.1.	<i>Feature Hook</i>	26
3.3.1.2.	<i>Feature Open</i>	26
3.3.1.3.	<i>Feature Lubang</i>	27
3.3.1.4.	<i>Feature Garis Vertikal</i>	28
3.3.1.5.	<i>Feature Garis Horizontal</i>	29
3.3.1.6.	<i>Feature Vertical Intersect</i>	30
3.3.1.7.	<i>Feature Horizontal Intersect</i>	30
3.3.1.8.	<i>Feature Blok</i>	30
3.3.1.9.	<i>Feature Garis</i>	31
3.3.1.10.	Lebar dan Tinggi Karakter	32
3.3.2.	Klasifikasi	32
3.3.3.	<i>Template Matching</i>	32
3.3.3.1.	Melakukan Perbandingan Setiap <i>Pixel</i> Antara Karakter <i>Template</i> Dengan Karakter Yang Akan Dikenali.....	33
3.3.3.2.	Melakukan Perbandingan <i>Feature</i> Antara Karakter <i>Template</i> Dengan Karakter Yang Akan Dikenali	33
4.	IMPLEMENTASI SISTEM	34
4.1.	<i>Image Preprocessing</i>	34

4.1.1.	<i>Thresholding</i>	34
4.1.2.	<i>Noise Removal 1</i>	34
4.1.3.	<i>Skew Angle Detection</i>	36
4.1.4.	Rotasi	37
4.1.5.	<i>Noise Removal 2</i>	38
4.1.5.1.	<i>Pepper Noise Removal</i>	38
4.1.5.2.	<i>Salt Noise Removal</i>	39
4.2.	<i>Image Document Analysis</i>	40
4.2.1.	Segmentasi Baris.....	40
4.2.2.	Segmentasi Karakter	41
4.3.	<i>Character Recognition</i>	45
4.3.1.	<i>Feature Extraction</i>	45
4.3.1.1.	<i>Feature Hook</i>	46
4.3.1.2.	<i>Feature Open</i>	47
4.3.1.3.	<i>Feature Lubang</i>	48
4.3.1.4.	<i>Feature Garis Vertikal</i>	49
4.3.1.5.	<i>Feature Garis Horizontal</i>	50
4.3.1.6.	<i>Feature Vertical Intersect</i>	51
4.3.1.7.	<i>Feature Horizontal Intersect</i>	52
4.3.1.8.	<i>Feature Blok</i>	52
4.3.1.9.	<i>Feature Garis</i>	53
4.3.2.	Klasifikasi	54
4.3.3.	<i>Template Matching</i>	55
4.4.	Penggunaan Program	56
4.4.1.	Membuka <i>Document Image</i> Pada Program	56
4.4.2.	Menjalankan Proses <i>Character Recognition</i>	59
4.4.3.	Melakukan Transfer Hasil <i>Character Recognition</i> Dari Program ke Microsoft Word	59
5.	PENGUJIAN PROGRAM.....	62
5.1.	Pengujian Dokumen Hasil <i>Scan</i> Dengan <i>Color Mode Grayscale</i>	62
5.1.1.	Pengujian Dokumen Pertama Dengan Jenis <i>Font Times New Roman</i>	62
5.1.2.	Pengujian Dokumen Kedua Dengan Jenis <i>Font Times New Roman</i>	64
5.1.3.	Pengujian Dokumen Pertama Dengan Jenis <i>Font Arial</i>	66
5.1.4.	Pengujian Dokumen Kedua Dengan Jenis <i>Font Arial</i>	67
5.1.5.	Pengujian Dokumen Pertama Dengan Jenis <i>Font Verdana</i>	70
5.1.6.	Pengujian Dokumen Kedua Dengan Jenis <i>Font Verdana</i>	72
5.2.	Pengujian Dokumen Hasil <i>Scan</i> Dengan <i>Color Text Enhanced</i>	74
5.2.1.	Pengujian Dokumen Pertama Dengan Jenis <i>Font Times New Roman</i>	74
5.2.2.	Pengujian Dokumen Kedua Dengan Jenis <i>Font Times New Roman</i>	76
5.2.3.	Pengujian Dokumen Ketiga Dengan Jenis <i>Font Times New Roman</i>	78
5.2.4.	Pengujian Dokumen Pertama Dengan Jenis <i>Font Arial</i>	80
5.2.5.	Pengujian Dokumen Kedua Dengan Jenis <i>Font Arial</i>	82

5.2.6.	Pengujian Dokumen Ketiga Dengan Jenis <i>Font</i> Arial.....	84
5.2.7.	Pengujian Dokumen Pertama Dengan Jenis <i>Font</i> Verdana	86
5.2.8.	Pengujian Dokumen Kedua Dengan Jenis <i>Font</i> Verdana.....	88
5.2.9.	Pengujian Dokumen Ketiga Dengan Jenis <i>Font</i> Verdana	90
6.	PENUTUP	93
6.1.	Kesimpulan	93
6.2.	Saran.....	93
	DAFTAR REFERENSI	94
	LAMPIRAN.....	95

DAFTAR TABEL

5.1. Hasil Pengujian Dokumen Pertama Times New Roman.....	64
5.2. Hasil Pengenalan Karakter Pada Dokumen Pertama Times New Roman Dengan OmniPage SE	64
5.3. Hasil Pengujian Dokumen Kedua Times New Roman.....	66
5.4. Hasil Pengenalan Karakter Pada Dokumen Kedua Times New Roman Dengan OmniPage SE	66
5.5. Hasil Pengujian Dokumen Pertama Arial	68
5.6. Hasil Pengenalan Karakter Pada Dokumen Pertama Arial Dengan OmniPage SE.....	68
5.7. Hasil Pengujian Dokumen Kedua Arial.....	70
5.8. Hasil Pengenalan Karakter Pada Dokumen Kedua Arial Dengan OmniPage SE.....	70
5.9. Hasil Pengujian Dokumen Pertama Verdana.....	72
5.10. Hasil Pengenalan Karakter Pada Dokumen Pertama Verdana Dengan Omni Page SE	72
5.11. Hasil Pengujian Dokumen Kedua Verdana	74
5.12. Hasil Pengenalan Karakter Dokumen Kedua Verdana Dengan OmniPage SE.....	74
5.13. Hasil Pengujian Dokumen Pertama Times New Roman	76
5.14. Hasil Pengenalan Karakter Pada Dokumen Pertama Times New Roman Dengan OmniPage SE.....	76
5.15. Hasil Pengujian Dokumen Kedua Times New Roman.....	78
5.16. Hasil Pengenalan Karakter Pada Dokumen Kedua Times New Roman Dengan OmniPage SE.....	78
5.17. Hasil Pengujian Dokumen Ketiga Times New Roman.....	80
5.18. Hasil Pengenalan Karakter Pada Dokumen Ketiga Times New Roman Dengan OmniPage SE	80

5.19. Hasil Pengujian Dokumen Pertama Arial	82
5.20. Hasil Pengenalan Karakter Pada Dokumen Pertama Arial Dengan OmniPage SE.....	82
5.21. Hasil Pengujian Dokumen Kedua Arial.....	84
5.22. Hasil Pengenalan Karakter Pada Dokumen Kedua Arial Dengan OmniPage SE.....	84
5.23. Hasil Pengujian Dokumen Ketiga Arial.....	86
5.24. Hasil Pengenalan Karakter Pada Dokumen Ketiga Arial Dengan OmniPage SE.....	86
5.25. Hasil Pengujian Dokumen Pertama Verdana.....	88
5.26. Hasil Pengenalan Karakter Pada Dokumen Pertama Verdana Dengan OmniPage SE.....	88
5.27. Hasil Pengujian Dokumen Kedua Verdana	90
5.28. Hasil Pengenalan Karakter Pada Dokumen Kedua Verdana Dengan OmniPage SE.....	90
5.29. Hasil Pengujian Dokumen Ketiga Verdana	92
5.30. Hasil Pengenalan Karakter Pada Dokumen Ketiga Verdana Dengan OmniPage SE.....	92

DAFTAR GAMBAR

2.1. Menetapkan <i>Threshold</i> Untuk Segmentasi Baris.....	7
2.2. Kotak <i>Structuring Element</i> 3x3 Untuk <i>Pepper Noise Removal</i>	8
2.3. Kotak <i>Structuring Element</i> 3x3 Untuk <i>Salt Noise Removal</i>	8
2.4. Segmentasi Karakter	9
2.5. <i>Feature Extraction</i>	11
3.1. Garis Besar Perencanaan Program.....	14
3.2. <i>Flowchart Image Preprocessing</i>	15
3.3. <i>Noise</i> Berupa Garis Pada Tepi Atas Dokumen	16
3.4. <i>Flowchart Skew Angle Detection</i>	17
3.5. Teks Miring.....	19
3.6. <i>Pixel</i> Hitam Bagian Bawah Karakter.....	19
3.7. <i>Flowchart</i> Rotasi.....	20
3.8. Titik Acuan Rotasi	20
3.9. Kotak <i>Structuring Element</i> 3x3 Untuk <i>Pepper Noise Removal</i>	21
3.10. Kotak <i>Structuring Element</i> 3x3 Untuk <i>Salt Noise Removal</i>	22
3.11. <i>Flowchart Image Document Analysis</i>	22
3.12. Segmentasi Baris.....	23
3.13. Segmentasi Karakter	23
3.14. Karakter Yang Saling Bertumpukan.....	24
3.15. Karakter Yang Saling Terhubung	24
3.16. <i>Flowchart Character Recognition</i>	25
3.17. Karakter Yang Memiliki <i>Hook</i>	26
3.18. Karakter Yang Memiliki <i>Feature Open</i>	26

3.19. Karakter Dengan Jumlah Lubang Satu	27
3.20. Batas Awal Lubang dan Batas Akhir Lubang	27
3.21. Karakter Dengan Jumlah Garis Vertikal Satu.....	28
3.22. Karakter Dengan Jumlah Garis <i>Horizontal</i> Satu	29
3.23. <i>Vertical Intersect</i> Pada Kolom 50% Dari Lebar Karakter	30
3.24. <i>Horizontal Intersect</i> Pada Baris 50% Dari Tinggi Karakter.....	30
3.25. Karakter Dengan Jumlah Blok Dua	30
3.26. <i>Feature</i> Garis Menggunakan Jumlah <i>Pixel</i> Hitam	31
3.27. <i>Feature</i> Garis Menggunakan Persentase Jumlah <i>Pixel</i> Hitam.....	31
3.28. Karakter Yang Memiliki Tinggi Lebih Besar Dari Lebar	32
3.29. Karakter Yang Memiliki Tinggi Lebih Besar Empat Kali Dari Lebar Karakter	32
4.1. Sebelum Proses Rotasi.....	37
4.2. Hasil Proses Rotasi	37
4.3. Hasil Proses <i>Salt Noise Removal</i>	39
4.4. Segmentasi Baris.....	40
4.5. Segmentasi Karakter Dengan Cara Melakukan <i>Scan</i> Secara Vertikal.....	41
4.6. Segmentasi Karakter Dengan Cara <i>Connected Component Labeling</i>	43
4.7. Memotong Pada Kolom yang Memiliki Jumlah <i>Pixel</i> Hitam Terkecil	44
4.8. Karakter <i>Template</i>	55
4.9. Tombol <i>Load File</i>	57
4.10. Kotak <i>Open Dialog</i> Untuk Membuka <i>File</i> Gambar	57
4.11. <i>Document Image</i> yang Dipilih	58
4.12. <i>Setting Threshold</i>	58
4.13. Tombol <i>Perform OCR</i>	59

4.14. Hasil <i>Character Recognition</i>	59
4.15. Tombol <i>Transfer to Word</i>	60
4.16. Kotak <i>Save Dialog</i>	60
4.17. Teks Hasil <i>Character Recognition</i> Pada Microsoft Word	61
5.1. Dokumen Pertama Times New Roman (tms11_kn_AG.bmp)	63
5.2. Potongan Dokumen Pertama Times New Roman (tms11_kn_AG.bmp)	63
5.3. Dokumen Kedua Times New Roman (tms11_kn_BG.bmp)	65
5.4. Potongan Dokumen Kedua Times New Roman (tms11_kn_BG.bmp)	65
5.5. Dokumen Pertama Arial (arl11_kn_AG.bmp)	67
5.6. Potongan Dokumen Pertama Arial (arl11_kn_AG.bmp)	67
5.7. Dokumen Kedua Arial (arl11_kr_BG.bmp)	69
5.8. Potongan Dokumen Kedua Arial (arl11_kr_BG.bmp)	69
5.9. Dokumen Pertama Verdana (vdn11_kn_AG.bmp)	71
5.10. Potongan Dokumen Pertama Verdana (vdn11_kn_AG.bmp)	71
5.11. Dokumen Kedua Verdana (vdn11_kr_BG.bmp)	73
5.12. Potongan Dokumen Kedua Verdana (vdn11_kr_BG.bmp)	73
5.13. Dokumen Pertama Times New Roman (tms11_kn_A.bmp)	75
5.14. Potongan Dokumen Pertama Times New Roman (tms11_kn_A.bmp)	75
5.15. Dokumen Kedua Times New Roman (tms11_kr_B.bmp)	77
5.16. Potongan Dokumen Kedua Times New Roman (tms11_kr_B.bmp)	77
5.17. Dokumen Ketiga Times New Roman (tms11_kn_C.bmp)	79
5.18. Potongan Dokumen Ketiga Times New Roman (tms11_kn_C.bmp)	79
5.19. Dokumen Pertama Arial (arl11_kn_A.bmp)	81
5.20. Potongan Dokumen Pertama Arial (arl11_kn_A.bmp)	81

5.21. Dokumen Kedua Arial (arl11_kr_B.bmp)	83
5.22. Potongan Dokumen Kedua Arial (arl11_kr_B.bmp)	83
5.23. Dokumen Ketiga Arial (arl11_kn_C.bmp)	85
5.24. Potongan Dokumen Ketiga Arial (arl11_kn_C.bmp)	85
5.25. Dokumen Pertama Verdana (vdn11_kn_A.bmp)	87
5.26. Potongan Dokumen Pertama Verdana (vdn11_kn_A.bmp)	87
5.27. Dokumen Kedua Verdana (vdn11_kr_B.bmp)	89
5.28. Potongan Dokumen Kedua Verdana (vdn11_kr_B.bmp)	89
5.29. Dokumen Ketiga Verdana (vdn11_kn_C.bmp)	91
5.30. Potongan Dokumen Ketiga Verdana (vdn11_kn_C.bmp)	91

DAFTAR LAMPIRAN

1. Klasifikasi 1	95
2. Klasifikasi 2	96
3. Klasifikasi 3	97
4. Klasifikasi 4	98