PERANCANGAN APLIKASI SISTEM MANAJEMEN DOKUMEN DAN PENCARIAN TEKS DENGAN MENGGUNAKAN OPTICAL CHARACTER RECOGNITION(OCR)

Maulana Sandy Dermawan [1] **Bagus Mulyawan** [2] **Manatap Dolok Lauro** [3]
[1), 2), 3)] Teknik Informatika Universitas Tarumanagara
Jl. Letjen S. Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
[1)] dermawanmaulana@gmail.com, [2)] bagus@fti.untar.ac.id, [3)] manataps@fti.untar.ac.id

ABSTRACT

This application aims to make the mail manager more practical and safe because the management system and the letter using the web-based OCR. This application is made with the System Development Life Cycle (SDLC) waterfall model through stages, namely the planning stage, the analysis phase, the design stage, the programming stage, and the testing phase. With this application can help archive letters in digital form (image file) by selecting the desired digital letter, then the application will convert the image form into character with the Optical Character Recognition method with the help of tesseract library so that all the text in the image can be converted well into the form of characters, then can be stored into the database and can do a search with certain keywords then the application will search the database for these keywords and then bring up the search results easily & quickly. At the end of this research the author concluded that this application can help speed up & make practical the search process for digital archive letters.

Key words

Archives, optical character recognition, search letters, system development life cycles, tesseract

1. Pendahuluan

Seperti organisasi pada umumnya, kegiatan di Perusahaan PT. Ditra Manunggal Jaya tidak terlepas dari proses administrasi *document entry* dan *template*. Proses ini dilakukan oleh bagian administrasi dokumen. Setiap harinya terdapat ± 20-40 dokumen yang masuk dan keluar. Kearsipan di dunia perdagangan (*trading*) sangat dibutuhkan baik untuk mengasipkan surat penawaran, surat pesanan, surat perjanjian penjualan, daftar pelanggan dan daftar harga serta arsip keuangan misalnya laporan keuangan, bukti pembayaran, daftar gaji, bukti pembelian, dan dan surat perintah bayar

Setiap dokumen yang masuk, dicatat pada buku penerimaan dokumen. Pada buku ini dicatat nomer *document entry*, tanggal *document entry*, nama pengirim dan nama tujuan dokumen. Pada buku ini juga disediakan kolom tanda tangan sebagai bukti dokumen telah diserahkan oleh orang yang dituju. Begitu juga untuk document template, bagian administrasi mencatat nomer dokumen, tanggal *document template*, nama pengirim dan tujuan pengiriman. Karena proses masih dilakukan secara manual beresiko terjadinya kehilangan data karena hilangnya buku catatan atau rusak serta pencarian dokumen yang kurang efektif dan efisien karena masih memakai lemari-lemari dokumen yang jumlah nya selalu bertambah serta arsip yang telah bertahun-tahun menumpuk. Begitu juga rentan terjadinya lupa tidak mencatat detail dokumen yang masuk atau keluar karena faktor manusia.

Begitu juga untuk kegiatan pengarsipan, bagian administrasi dapat melakukan foto surat dokumen dan melakukan *upload* foto surat dokumen dengan bantuan OCR *tesseract* agar surat teks tersebut dapat dijadikan acuan untuk pencarian teks pada sistem. Data digital dokumen akan tersimpan pada database sistem dan dapat dilakukan pencarian dengan mudah dan cepat berdasarkan nomor dokumen dan cari secara spesifik dengan bantuan OCR *tesseract*, jika sewaktu-waktu dibutuhkan kembali. Dengan adanya sistem ini, juga dapat mengurangi beban kerja bagian administrasi untuk melayani karyawan yang membutuhkan format baku *document template* perusahaan. Karena karyawan yang membutuhkan dapat langsung *download* format *document template* pada sistem. Sebagai keamanan penggunaan *format document template*, bagian administrasi dapat melakukan *setting* hak akses pengguna dokumen. Sehingga tidak semua pengguna sistem dapat *download format document template* yang tersedia. Diharapkan sistem yang diusulkan dapat memberikan solusi atas persoalan yang terjadi dan memberikan efisiensi penggunaan kertas dan tempat

untuk pengarsipan dokumen di PT. Ditra Manunggal Jaya.

2. Dasar Teori

2.1 Dokumentasi

Dokumentasi adalah suatu cara yang dilakukan untuk menyediakan dokumen dengan menggunakan bukti yang akurat dari pencatatan sumber informasi khusus dari sebuah karangan atau tulisan, wasiat, buku, undang-undang dan lain sebagainya. Atau dengan kata lain, pengertian dokumentasi secara umum adalah suatu pencarian, penyelidikan, pengumpulan, pengawetan, penguasaan, pemakaian dan penyediaan dokumen.[1]

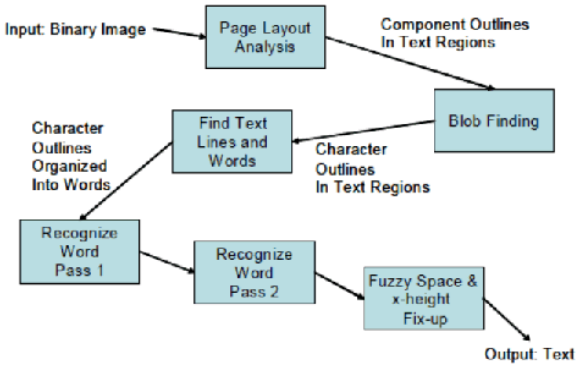
2.2 *Optical Character Recognition(OCR)*

Character Recognition bertugas untuk mengenali tulisan di dalam mengenali karakter tulisan dalam gambar dan merubahnya kedalam *American Standad Code for Information Interchange (ASCII)* atau bahasa mesin lainnya yang setara dan dapat diedit. [2]

2.3 *Tesseract*

Tesseract adalah mesin *OCR open-source* yang dikembangkan di *HP (Hewlett-Packard)* antara tahun 1984 dan 1994. *Tesseract* muncul sebagai projek penelitian *PhD* di *HP Labs, Bristol* yang tersedia di *http://code.google.com/p/tesseract-ocr* [3]. Tahapan proses *Optical Character Recognition* oleh *Tesseract* adalah:

1. *Page Layout Analysis*
2. *Blob Finding*
3. *Find Text Line and Words*
4. *Recognition Word Pass 1*
5. *Recognition Word Pass 2*
6. *Fuzzy Space and x-height Fix-up*



Gambar 1 Arsitektur Tesseract

3. Hasil Percobaan

Pengujian terhadap data bertujuan untuk mengetahui apakah program tersebut sudah dapat berjalan sesuai dengan konsep. Pengujian dilakukan dengan kuesioner *User Acceptance Test* seperti:

1. *User Acceptance Test*
   Pengujian *UAT* dilakukan oleh karyawan perusahaan. Berikut hasil dari pengujiannya:

Tabel 1 *UAT Form Document Entry*

Diuji Oleh			
No	Peran	Nama	Menguji
1	Staff Admin	Delia Puspa	Form *Document Entry* dan *Document Template*

Tabel 1 (Lanjutan)

Deskripsi	Prosedur Pengujian	Masukan	Keluaran yang Diharapkan	Hasil yang Didapatkan
Pengujian Form Document Entry	Verifikasi Add Document	Input Document Type, Document Name, dan Role	Menampilkan List Document Entry	Menampilkan List Document Entry
Pengujian Form Document Entry	Verifikasi Upload Document	Browse Image dan Document	Menampilkan List Document Entry yang sudah terupdate dikolom Document dan hasil OCR dikolom Word	Menampilkan List Document Entry yang sudah terupdate dikolom Document dan hasil OCR dikolom Word
Pengujian Form Document Entry	Verifikasi Delete Document	Klik button Delete	Tampilan Document yang dipilih telah dihapus	Tampilan Document yang dipilih telah dihapus
Pengujian Form Document Entry	Verifikasi Add Document	Input Document Type, Document Name, dan Role	Menampilkan List Document Entry	Menampilkan List Document Entry
Pengujian Form Document Entry	Verifikasi Upload Document	Browse Image dan Document	Menampilkan List Document Entry yang sudah terupdate dikolom Document dan hasil OCR dikolom Word	Menampilkan List Document Entry yang sudah terupdate dikolom Document dan hasil OCR dikolom Word
Pengujian Form Document Entry	Verifikasi Delete Document	Klik button Delete	Tampilan Document yang dipilih telah dihapus	Tampilan Document yang dipilih telah dihapus

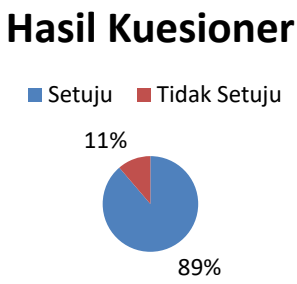
Tabel 3 *User Acceptance Test* Alternatif Jawaban

Alternatif Jawaban	Jumlah	Presentase
Sangat Setuju	13	65%
Setuju	5	25%
Tidak Setuju	2	10%
Sangat Tidak Setuju	-	-
Total	**20 Orang**	**100%**

Perhitungan atas hasil kuesioner diatas dapat dilakukan secara manual. Berikut cara menghitung hasil pengamatan secara manual menggunakan penskoran Skala *LIKERT* :

1. Jumlah skor untuk 2 orang yang menjawab
   Sangat Setuju (4) : 13 x 4 = 52
2. Jumlah skor untuk 4 orang yang menjawab
   Setuju (3) : 5 x 3 = 15
3. Jumlah skor untuk 7 orang yang menjawab
   Tidak Setuju (2) : 2 x 2 = 4
4. Jumlah skor untuk 0 orang yang menjawab
   Sangat Tidak Setuju (1) : 0 x 1 = 0
   J U M L A H : = 71

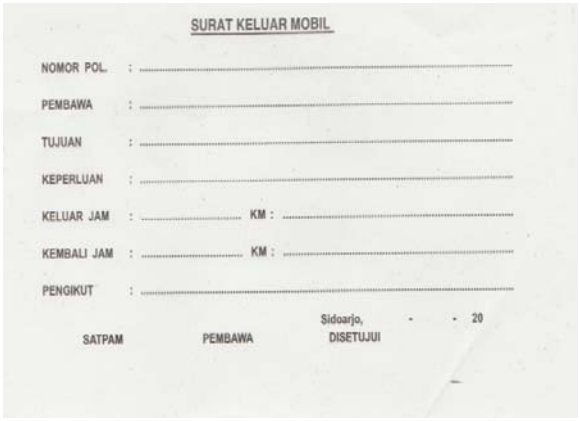
Jumlah skor ideal untuk pertanyaan yang diajukan kepada responden :
1. Skor tertinggi : 4 x 20= 80 (Sangat Setuju)
2. Skor terendah : 1 x 20= 20 (Sangat Tidak Setuju)
Interpretasi skor hasil pengamatan :
(71/80) x 100% = 88,75%



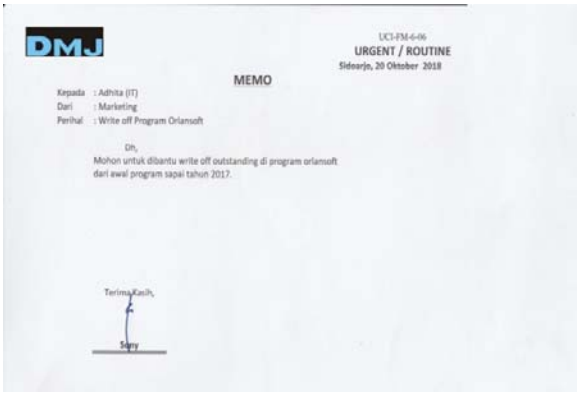
Gambar 2 Hasil Kuesioner

2. *OCR*

Berikut contoh surat-surat perusahaan dan perhitungan akurasi OCR berdasarkan pemisahan karakter terhubung dan *chopping* atau pemotongan karakter.



Gambar 3 Surat Keluar Mobil



Gambar 4 Surat Memo

Tabel 4 Perbandingan Data Asli dan Hasil OCR Surat Keluar Mobil

No	Data Asli Pada Surat Keluar Mobil (*Font Arial*)	Hasil Pengujian *OCR*
1.	SURAT KELUAR MOBIL NOMOR POL: PEMBAWA: TUJUAN: KEPERLUAN: KELUAR JAM: KM: KEMBALI JAM: KM: PENGIKUT: SATPAM PEMBAWA Sidoarjo, - - 20 DISETUJUI	NOMOR POL. PEMBAWA TUJUAN KEPERLUAN KELUAR JAM KEMBALI JAM PENGIKUT SATPAM SURAT KELUAR MOBIL PEMBAWA Sidoarjo, DISETUJUI 20

Tabel 5 Perbandingan Data Asli dan Hasil OCR Surat Memo

No	Data Asli Pada Surat Memo (*Font Arial)*	Hasil Pengujian *OCR*
1.	UCI-FM-6-06 **URGENT / ROUTINE** Sidoarjo, 20 Oktober 2018 **MEMO** Kepada _: Adhita (IT) Dari : Marketing Perihal : Write off Program Orlansoft Dh, Mohon untuk dibantu write off outstanding di program orlansoft dari awal program sapai tahun 2017. Terima,Kasih, Sony	UCI-FM-6-06 URGENT / ROUTINE Sidoarjo, 20 Oktober 2018 Kepada _: Adhita (IT) Dari : Marketing Perihal : Write off Program Orlansoft Dh, Mohon untuk dibantu write off outstanding di program orlansoft dari awal program sapai tahun 2017. Terima,Kasih,

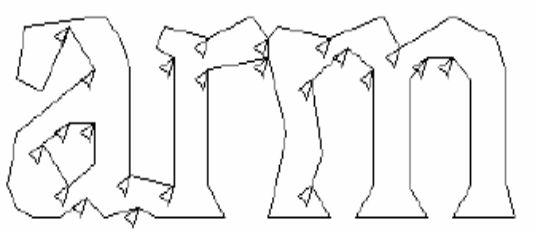
Dari tabel-tabel pengujian perbandingan data asli dengan data hasil pengujian OCR memiliki permasalahan yang berbeda. Berikut tabel keterangan surat-surat perusahaan:

Tabel 6 Keterangan Hasil Uji

No.	Nama Surat	Keterangan Uji *OCR*	Keterangan
1.	Surat memo	Berhasil	• Surat hasil *scan* sedikit miring • teks distabilo tidak terbaca • tanda tangan mengenai teks tidak terbaca
2.	Surat mobil keluar	Berhasil	• Surat hasil scan sedikit miring • *Font* terlalu kecil

3. Pemisahan Karakter Terhubung

Apabila hasil dari pengenalan kata tidak memuaskan, *tesseract* berusaha untuk memperbaiki hasil dengan memisahkan *blob* dengan keyakinan terburuk dari pengklasifikasian (*classifier*) karakter. Kandidat untuk titik-titik pemisahan ditemukan dari simpul cekung dari pendekatan poligonal *outline* dan mungkin saja terdapat titik cekung berlawanan lainnya atau segmen garis.



Gambar 5 Kandidat Titik Potong

Tabel 7 Perbandingan analisis pemisahan karakter terhubung *font arial* pada surat jalan

No.	Asli	Hasil
1.	No	Ho
2.	jml	jmi
3.	Kg	ko
4.	bukti resmi	buldiresmi
5.	bukan	buken
6.	bukti	bulti
7.	penjualan	Penjualan
8.	dilengkapi	Cllengkapl
9.	/	J

Contoh pada surat jalan pada huruf ‘No’ dengan ‘Ho’ memiliki tingkat kemiripan satu set calon titik potong (*chop points*) dengan tanda panah dengan potongn terpilih sebagai sebuah garis melintasi kerangka dimana huruf ‘N’ melintas vertikal dan sedikit menyerong lalu ‘H’ melintas *horizontal*.

4. *Chopping* atau Pemotongan Karakter

Tesseract menguji garis teks *(text line)* untuk menentukan apakah mereka merupakan *fixed pitch*. Bila ditemukan *fixed pitch text*, *tesseract* memotong kata-kata menjadi karakter-karakter

Gambar 6 Pemotongan karakter

Tabel 8 Perbandingan analisis *Chopping font arial* surat keluar mobil

No.	Asli	Hasil
1.	Surat Keluar Mobil	Tidak terbaca
2.	.	:
3.	:	Tidak terbaca
4.	-	Tidak terbaca

Contoh *chopping* pada surat jalan yaitu dari kata '(Tanda' menjadi 'ands' karena garis teks (text line) berhimpitan antara '(' dengan 'T' maka tesseract memotong kata-kata tersebut. Hanya surat kwitansi yang tidak memilik permasalahan *chopping*.

4. Perhitungan dan analisis hasil OCR

Dari total 10 surat perusahaan yang diuji dengan *OCR* dengan *tesseract library* adalah 5 surat yang memiliki kualitas baik yang dapat dikonversi menjadi teks. 5 hasil yang gagal memiliki kriteria berupa kualitas hasil *scan* tidak baik, *over brightness*, *brightness* tidak merata, *font* berwarna dan terlalu kecil, kualitas surat hasil fotokopi, surat miring, dan *crop* scan tidak baik. Oleh karena itu diperlukan kehati-hatian dalam melakukan proses *scan* dan pemilihan surat dengan kriteria yang disebutkan diatas harus terpenuhi.

5 hasil yang lolos akan diambil nilai akurasinya. Nilai akurasi merupakan persentase data yang benar berbanding dengan keseluruhan data sehingga rumus menghitung akurasi seperti pada rumus 4.1.

$$akurasi = \frac{data\ yang\ benar}{keseluruhan\ data}\ x\ 100\% \quad (4.1)$$

1. Total dalam surat jalan memiliki 476 teks, 24 diantaranya tidak akurat. 478 – 24 = 452 teks sesuai dengan surat aslinya.

$$akurasi = \frac{452}{476}\ x\ 100\%$$
$$= 94{,}95\ \%$$

2. Surat kwitansi memiliki akurasi 100%
3. Total dalam surat memo memiliki 258 teks, 9 diantaranya tidak akurat. 258 – 9 = 249 teks sesuai dengan surat aslinya.

$$akurasi = \frac{249}{258}\ x\ 100\%$$
$$= 96{,}51\ \%$$

4. Total dalam surat keluar mobil memiliki 145 teks, 20 diantaranya tidak akurat. 145 – 20 = 125 teks sesuai dengan surat aslinya.

$$akurasi = \frac{125}{145}\ x\ 100\%$$
$$= 86{,}20\ \%$$

5. Total dalam surat memo internal memiliki 440 teks, 32 diantaranya tidak akurat. 440 – 32 = 408 teks sesuai dengan surat aslinya.

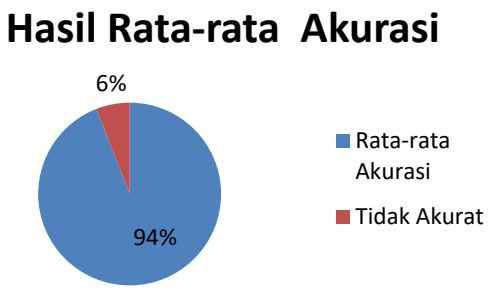
$$akurasi = \frac{408}{440}\ x\ 100\%$$
$$= 92{,}72\ \%$$

Dari hasil perhitungan diatas akan menghasilkan rata-rata tingkat akurasi yang dicapai oleh setiap surat yang melalui proses *OCR* dengan *tesseract libarary*. Nilai rata-rata dari suatu kelompok data adalah jumlah nilai dibagi dengan banyaknya data. sehingga rumus menghitung rata-rata akurasi seperti pada rumus 4.1.

$$nilai\ rata - rata\ akurasi = \frac{jumlah\ nilai}{banyaknya\ data}\ x\ 100\% (4.1)$$

$$rata - rata\ akurasi = \frac{94{,}95 + 100 + 96{,}51 + 86{,}20 + 92{,}72}{4}\ x\ 100\%$$

$$rata - rata\ akurasi = \frac{470{,}38}{5}\ x\ 100\%$$
$$= 94{,}08\ \%$$



Gambar 6 Hasil Rata-rata Akurasi

4. Kesimpulan

Kesimpulan yang dapat diperoleh berdasarkan pembuatan dan pengujian dari aplikasi ini adalah sebagai berikut:

1. Pengujian terhadap modul-modul web yang terdapat pada aplikasi dapat berjalan dengan baik tanpa adanya *error* atau bug untuk bagian administrator mengelola dokumen pengarsipan dokumen.
2. Pengujian terhadap pengelolaan dokumen dapat mengurangi resiko terjadinya kehilangan data karena hilangnya buku catatan atau rusak serta efisiensi waktu pencarian dokumen. Karena proses sudah menggunakan aplikasi sistem manajemen dokumen dan pencarian teks menggunakan *Optical Character Recognition (OCR)*.
3. Pengujian terhadap 10 surat perusahaan hanya ada 5 surat yang memiliki kualitas baik yang dapat dikonversi menjadi teks oleh OCR dengan *tesseract* library yaitu surat jalan dengan 94,95%, surat kwitansi dengan 96,51%, surat 86,20%, dan surat memo internal dengan 92,72%. Maka hasil rata-rata

akurasi dari jumlah akurasi diatas dibagi dengan banyaknya data mencapai nilai 94,08%. Mencari dan menelusur informasi terhadap arsip surat yang sudah disimpan dengan OCR didapatkan nilai akurasi 92,72% hasil teks OCR.

4. Pengujian pada bagian administrator untuk melayani karyawan yang membutuhkan format baku document template perusahaan sudah bisa *download* di *website* dengan *role* yang sudah disediakan masing-masing.

Saran untuk yang ingin mengembangkan aplikasi adalah sebagai berikut:

1. Untuk menggunakan aplikasi sistem manajemen dokumen dan pencarian teks dengan menggunakan *Optical Character Recognition (OCR*) memiliki kriteria yaitu kualitas hasil scan yang baik(atur *brightness* secara baik), *font* berwarna dan terlalu kecil sulit dibaca, kualitas surat hasil fotokopi yang baik, pastikan surat miring, dan *crop scan* dengan rapih.
2. Dari hasil *User Acceptance Testing* karyawan yang telah terbantu dengan adanya aplikasi ini mencapai 88,75% dan 11,25% ada beberapa karyawan bagian logistik belum pernah mengakses *website* sehingga menurut mereka sulit digunakan. Maka aplikasi kedepannya dibuat dalam bentuk *mobile* agar lebih mudah diakses dan fleksibel.

REFERENSI

[1] Si Manis, “Pengertian Dokumentasi, Fungsi, Tujuan, Peranan dan Kegiatan Dokumentasi Menurut Para Ahli Terlengkap, http://www.pelajaran.co.id/2017/28/pengertian-dokumentasi-menurut-para-ahli-fungsi-tujuan-peranan-kegiatan-dokumentasi.html , 8 Maret 2018.

[2] Rao V, Sasrty A, Chakracarthy A, & Kalyanchakravarthi P., “Optical Character Recognition Technique Algorithms”, Journal of Theoretical and Applied Information Technology, Vol 83, Nomor 2, (January, 2016), h. 275-282.

[3] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc 9th Int.Conf.on Document Analysis and Recognition, 2007.

Maulana Sandy Dermawan, mahasiswa tingkat akhir Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta

Bagus Mulyawan, S.Kom.,MM, memperoleh gelar S.Kom. dari Universitas Gunadarma pada tahun 1992. Kemudian memperoleh gelar M.M. dari Universitas Budi Luhur. Saat ini sebagai dosen Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.

Manatap Dolok Lauro, S.Kom., MMSI, memperoleh gelar S.Kom dari Universitas Tarumanagara pada 2006. Kemudian memperoleh gelari MMSI dari Universitas Binas Nusantara pada tahun 2010. Saat ini sebagai dosen Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara, Jakarta.