IMPLEMENTASI OPTICAL CHARACTER RECOGNITION (OCR) DAN PENDEKATAN THESAURUS UNTUK MENEMUKAN INFORMASI PADA SURAT MASUK DI STMIK STIKOM INDONESIA

I Made Avendias Mahawan[1)] **I Putu Agus Eka Darma Udayana**[2)]
Program Studi Teknik Informatika[1) 2)]
STMIK STIKOM Indonesia[1) 2)]
avendias@stiki-indonesia.ac.id[1)] agus.ekadarma@stiki-indonesia.ac.id [2)]

ABSTRACT

STMIK STIKOM Indonesia currently has 3,953 active students and 1,400 new students entering the academic year 2019/2020. This campus has good cooperation between educational institutions or other non-educational institutions which are indicated through an invitation to attend or participate in activities organized by other institutions. At present, incoming letters or invitations received an average of 50 letters in a month, to simplify the process of incoming letters management, this campus began to make improvements by developing a letter filing system, but the process of handling incoming mails has not been done automatically. Optical Character Recognition (OCR) is one of the technologies that can be used to recognize incoming letter details in the form of letter numbers, letter dates, subject matters, letter destinations and source of letters from the scan results of incoming letters. In this study the researchers proposed OCR technology and a thesaurus approach to be able to obtain information from the results of scans of incoming letters at STMIK STIKOM Indonesia, with 50 test data, the accuracy of recognition obtained from OCR technology will be calculated. The test results using 3 scenarios with the number of thesaurus are 10, 30 and 50, produce the highest level of recognition accuracy that is 92% when using 50 thesaurus.

Keywords*:OCR, Thesaurus, Incoming Letters,* STMIK STIKOM Indonesia.

ABSTRAK

STMIK STIKOM Indonesia saat ini memiliki mahasiswa aktif lebih dari 4.000 orang. Kampus ini memiliki kerja sama yang baik antar lembaga kependidikan ataupun lembaga lain non kependidikan yang ditunjukkan melalui undangan menghadiri ataupun mengikuti kegiatan yang diselenggarakan oleh lembaga lain. Saat ini, surat masuk atau undangan yang diterima mencapai rata-rata 50 surat dalam sebulan, untuk mempermudah proses manajemen surat masuk, kampus ini mulai melakukan perbaikan dengan mengembangkan sistem pengarsipan surat, namun proses penanganan surat masuk belum dilakukan secara otomatis. *Optical Character Recognition* (OCR) merupakan salah satu teknologi yang dapat dimanfaatkan untuk mengenali detail surat masuk berupa nomor surat, tanggal surat, perihal surat, tujuan surat serta sumber surat dari hasil scan surat masuk tersebut. Pada penelitian ini peneliti mengajukan teknologi OCR dan pendekatan *thesaurus* untuk dapat memperoleh informasi dari hasil scan surat masuk di STMIK STIKOM Indonesia, dengan 50 data uji, maka akan dihitung akurasi pengenalan yang diperoleh dari teknologi OCR. Hasil pengujian menggunakan 3 skenario dengan jumlah *thesaurus* yaitu 10, 30 dan 50 menghasilkan tingkat akurasi pengenalan tertinggi yaitu 92% saat menggunakan 50 *thesaurus*.

Kata Kunci : OCR, Thesaurus, Surat Masuk, STMIK STIKOM Indonesia

PENDAHULUAN

STMIK STIKOM Indonesia sebagai salah satu perguruan tinggi swasta di lingkungan LLDIKTI (Lembaga Layanan Pendidikan Tinggi) wilayah 8 saat ini memiliki mahasiswa aktif sebanyak 3.953 orang serta mahasiswa baru yang masuk di tahun ajaran 2019/2020 sebanyak 1.400 orang, dengan banyaknya jumlah mahasiswa tersebut membuat kampus ini mulai terkenal di kawasan Indonesia Timur khususnya di Bali. Sebagai lembaga pendidikan di bawah Kementerian Riset, Teknologi, dan Pendidikan Tinggi kampus ini memiliki kerja sama yang baik antar lembaga kependidikan ataupun lembaga lain non kependidikan di Indonesia dan khususnya di Bali. Salah satu kerjasama yang sudah terwujud adalah kerjasama dalam bidang penelitian bersama dengan Dinas Kabupaten Badung di tahun 2018-2019. Bentuk lain dari kerjasama yang baik adalah undangan-undangan untuk menghadiri ataupun mengikuti kegiatan-kegiatan yang diselenggarakan oleh lembaga atau organisasi lain. Saat ini, surat masuk atau undangan-undangan yang diterima STMIK STIKOM Indonesia mencapai rata-rata 50 surat dalam sebulan, untuk mempermudah pekerjaan dari staf *front office* (FO) dalam proses manajemen surat masuk, STMIK STIKOM Indonesia mulai melakukan perbaikan dengan mengembangkan sistem pengarsipan surat, sehingga dengan pengembangan sistem tersebut dapat mempermudah pencatatan, pencarian dan pelaporan surat setiap tahunnya. Dalam pengembangan sistem tersebut, proses penanganan surat masuk dilakukan dengan melakukan scan surat lalu melakukan *input* secara manual nomor surat, tanggal surat, perihal surat, sumber surat/pengirim dan tujuan surat, dalam hal ini peneliti melihat perlu ditambahkan suatu fitur yang dapat mempermudah staf FO untuk proses menyimpan surat masuk ke dalam sistem tanpa perlu melakukan *input* secara manual detail surat, dengan memanfaatkan teknologi *Optical Character Recognition* (OCR) diharapkan dapat mengenali detail surat berupa nomor surat, tanggal surat, perihal surat, tujuan surat serta sumber surat dari hasil *scan* surat masuk tersebut. Pada penelitian ini peneliti mengajukan teknologi OCR dan pendekatan *Thesaurus* untuk dapat memperoleh informasi dari hasil *scan* surat masuk di STMIK STIKOM Indonesia, dengan 50 data uji, maka akan dihitung akurasi pengenalan yang diperoleh dari teknologi OCR tersebut.

TINJAUAN PUSTAKA

Pengolahan Citra Digital

Pengolahan citra digital dapat diartikan sebagai proses mengolah suatu gambar (citra) pada dimensi dua dengan menggunakan komputer digital, pengolahan citra juga diartikan sebagain suatu teknik yang digunakan untuk mengolah, memanipulasi dan memodifikasi citra [1]. Citra digital merupakan suatu fungsi intensitas cahaya dua dimensi f(x,y), dimana x dan y menunjukkan koordinat spasial. Nilai f pada setiap titik (x,y) menunjukkan tingkat (nilai) warna dari citra pada titik tersebut[2].

Pengenalan Pola

Pengenalan pola (*pattern recognition*) merupakan suatu kajian ilmu yang dapat menggambarkan suatu berdasarkan ciri atau sifat utama dari suatu objek tersebut kemudian mengklasifikasikannya berdasarkan ciri atau sifat utama dari objek tersebut. Pola sendiri adalah suatu entitas yang terdefinisi dan dapat diidentifikasikan serta diberi nama. Pola bisa merupakan kumpulan hasil pengukuran atau pemantauan dan bisa dinyatakan dalam notasi vektor atau matriks [1]. Pengenalan pola dapat diimplementasikan pada kehidupan sehari-hari seperti misalkan melakukan pengenalan karakter di media cetak pada media digital menggunakan konsep OCR untuk konversi dokumen cetak ke dokumen digital [3], pengenalan pola huruf arab dengan cara melakukan pengujian terhadap kinerja OCR [4], pengenalan pola juga dapat menjadi salah satu cara untuk melestarikan budaya di Indonesia khususnya di Bali, seperti pengklasifikasian jenis ukiran Bali menggunakan *Learning Vector Quantization* (LVQ)[5].

METODE PENELITIAN

Tahapan Penelitian

Penelitian ini menggunakan 5 alur yaitu pengumpulan data, analisis dan perancangan, implementasi, pengujian serta penyusunan laporan dan publikasi. Tahap pengumpulan

data diperoleh melalui proses observasi dan studi literatur, data yang terkumpul digunakan pada tahap analisis dan perancangan, selanjutnya hasil dari perancangan tersebut diimplementasikan ke dalam program. Tahap selanjutnya adalah pengujian terhadap program yang sudah dibangun, pengujian bertujuan untuk menyesuaikan program dengan spesifikasi yang dirancang serta menghindari kesalahan-kesalahan yang terjadi pada program.

Gambar 1. Alur Penelitian

Teknik Pengumpulan Data

Teknik pengumpulan data dilakukan dengan melakukan observasi terhadap proses rekapitulasi pencatatan surat masuk melalui bagian Front Office (FO) serta melalui wawancara dilakukan bersama dengan staf Front Office (FO) STMIK STIKOM Indonesia

Analisis dan Metode

Proses yang ada di dalam sistem yang dikembangkan dimulai dari proses akuisisi data yaitu dengan menggunakan scaner sebagai alat untuk memindai berkas surat tercetak menjadi citra surat, setelah proses akuisisi maka dilanjutkan dengan preprocessing yaitu dengan melakukan proses grayscaling pada citra surat, proses grayscaling dilakukan dengan menggunakan persamaan berikut.[5]

$$I(x,y)= \alpha.R+ \beta.G+ \gamma.B$$

Keterangan :

I(x,y) = Citra Graysclale
R = Nilai red pada citra warna
G = Nilai green pada citra warna
B = Nilai blue pada citra warna

Dengan konstanta α=0,2989; β=0,5870; γ=0,1140


Kura Kura
Bali

Tanggal : 11 Maret 2019
Ref. No. : 40 / BTID 2/HRM TECH/III/2019
Lampiran : 1 lbr poster
Perihal : Undangan

Kepada Yth,

I Dewa Made Khrisna Muku, S.T., M.T.
Ketua STIKI Indonesia
Di
Jl. Tukad Pakerisan 97, Denpasar.

Dengan hormat,

Kura Kura Bali, Island of Happiness bekerja sama dengan Tsinghua Sc
dari akademisi, masyarakat, bisnis dan pemerintah untuk menciptl
memberikan dampak nyata dengan hadirnya solusi pembangunan b

Sebagai wujud realisasi, Universitas Tsinghua Southeast Asia ber
membangun **Pendidikan dan Teknologi** yang berkelanjutan melalu
Centre: Innovation for Mobility in Southeast Asia" yang difasilitasi
pembicara dari **Grab & OVO** yang akan diadakan pada:

Hari / Tanggal : Kamis, 21 Maret 2019
Pukul : 08.30 – 11.30 WITA
Tempat : Three Mountains - UID Creative Campu
Akses masuk dari Jalan Bypass Ngurah
Denpasar, Bali
Topik : "Innovation for Mobility in Southeast A

Gambar 2. Citra *Grayscale*

Proses berikutnya yang dilakukan adalah filtering untuk mengurangi noise pada citra, proses filtering dilakukan dengan menggunakan metode average filter dan median filter. Average Filter merupakan metode yang digunakan untuk mengurangi noise dengan meratakan sejumlah titik tertentu dalam citra, sehingga menghasilkan titik pixel yang merata dalam citra luaran, proses average filter dapat diterapkan pada piksel f(y, x) sebagai berikut.[6].

65	50	55
78	68	60
60	60	62

Gambar 3. Contoh Piksel f(y,x)

Pada contoh di atas, pixel bernilai 68 merupakan nilai pada f(y, x). Nilai rerata hasil average filter dari pixel pengganti untuk g(y, x) dihitung seperti berikut:

$g(y, x) = 1/9\times(65+50+55+76+68+60+60+60+62)=61,78 \cong 62$

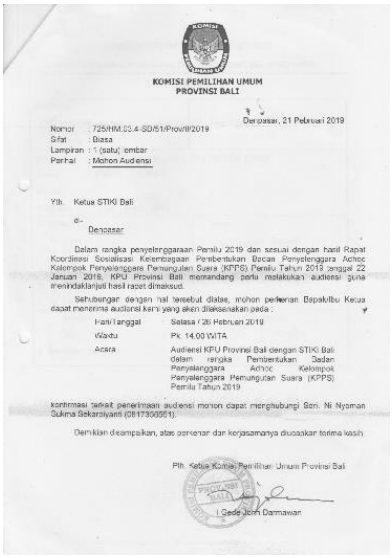
maka, nilai 68 pada f(y, x) menjadi 62 pada g(y, x) setelah melalui proses average filter. Metode median filter sebagai filter non linear yang dikembangkan oleh Tukey merupakan

sebuah metode yang digunakan untuk mengurangi noise dan menghaluskan citra. Metode ini bekerja menggunakan penapis dan dihitung dengan melakukan pengurutan nilai intensitas sekelompok pixel, yang kemudian nilai pixel tersebut diganti dengan nilai tertentu, metode ini dikatakan sebagai operasi non-linier karena tidak melakukan proses konvolusi. Penapis atau window atau mask pada median filter memuat pixel ganjil, penapis tersebut digeser titik per titik mencapai seluruh daerah citra, nilai pixel tersebut diurutkan secara ascending kemudian dihitung nilai mediannya. Hasil nilai yang diperoleh digunakan untuk menggantikan nilai yang berada pada pusat penapis [6]. Contoh penapis median filter ditunjukkan pada gambar 4.

	123	125	126	130	140	
	122	**124**	**126**	**127**	135	
	118	**120**	**150**	**125**	134	
	119	**115**	**119**	**123**	133	
	111	116	110	120	130	

Gambar 4. Contoh Piksel f(y,x)

Citra menjadi lebh halus setelah melalui proses filtering. Hasil yang diperoleh setelah proses filtering citra surat dapat dilihat pada Gambar 5.



KOMISI PEMILIHAN UMUM
PROVINSI BALI

Denpasar, 21 Pebruari 2019

Nomor : 725/HM.03.4-SD/51/Prov/II/2019
Sifat : Biasa
Lampiran : 1 (satu) lembar
Perihal : Mohon Audiensi

Yth. Ketua STIKI Bali
di-
Denpasar

Dalam rangka penyelenggaraan Pemilu 2019 dan sesuai dengan hasil Rapat Koordinasi Sosialisasi Kelembagaan Pembentukan Badan Penyelenggara Adhoc Kelompok Penyelenggara Pemungutan Suara (KPPS) Pemilu Tahun 2019 tanggal 22 Januari 2019, KPU Provinsi Bali memandang perlu melakukan audiensi guna menindaklanjuti hasil rapat dimaksud.

Sehubungan dengan hal tersebut diatas, mohon perkenan Bapak/Ibu Ketua dapat menerima audiensi kami yang akan dilaksanakan pada :

Hari/Tanggal : Selasa / 26 Februari 2019
Waktu : Pk. 14.00 WITA
Acara : Audiensi KPU Provinsi Bali dengan STIKI Bali dalam rangka Pembentukan Badan Penyelenggara Adhoc Kelompok Penyelenggara Pemungutan Suara (KPPS) Pemilu Tahun 2019

konfirmasi terkait penerimaan audiensi mohon dapat menghubungi Sdri. Ni Nyoman Sukma Sekarsiyanti (081736005051).

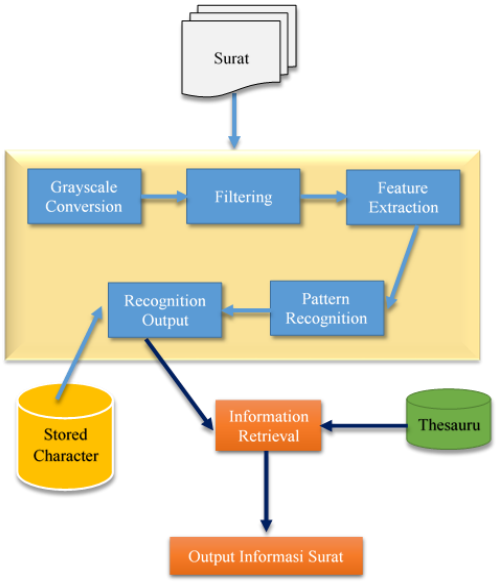
Demikian disampaikan, atas perhatian dan kerjasamanya diucapkan terima kasih.

Plh. Ketua Komisi Pemilihan Umum Provinsi Bali

I Gede Yoga Darmawan

Gambar 5. Citra Surat Hasil *Filtering*

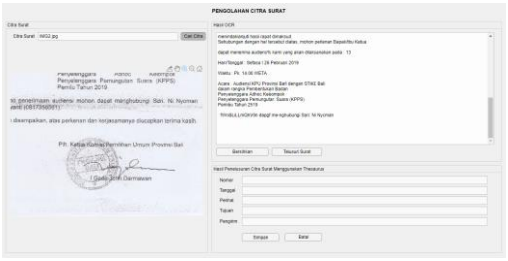
Proses selanjutnya adalah proses segmentasi citra yang menggunakan metode proyeksi vertikal dan horisontal, proses ini bertujuan untuk membagi citra menjadi region tertentu agar mempermudah proses ekstraksi ciri. Proses ekstraksi ciri dilakukan dengan menggunakan metode *template matching* yaitu piksel masing-masing gambar dicocokkan dengan *template* yang sudah disediakan, karakter yang memiliki kecocokkan tertinggi dianggap sebagai karakternya, proses terakhir yaitu *information retrieval* yang berfungsi untuk menemukan informasi yang terdapat pada surat tersebut dicocokkan berdasarkan data yang ada pada *Thesaurus Stored*. Alur penelusuran detail surat pada penelitian ini ditunjukkan pada gambar 6.



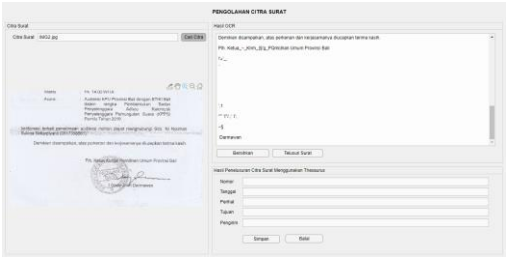
Gambar 6. Alur Penelusuran Detail Surat

Pengujian dan Hasil

Penelitian ini mengalami beberapa kendala saat melakukan OCR pada citra surat yang memiliki noise, melalui studi literatur peneliti menemukan solusi untuk mengatasi masalah tersebut. Peneliti memutuskan untuk menggabungkan dua buah filter yaitu average filter dan median filter pada pra proses sebelum dilakukan proses OCR. Hasil yang diperoleh sistem dapat mengenali kata yang lebih banyak jika dibandingkan saat sebelum menggunakan tambahan filter. Gambar 6 menunjukkan hasil yang diperoleh sebelum menggunakan tambahan filter dan sesudah.



Gambar 7a. Hasil Sebelum Menggunakan Filter

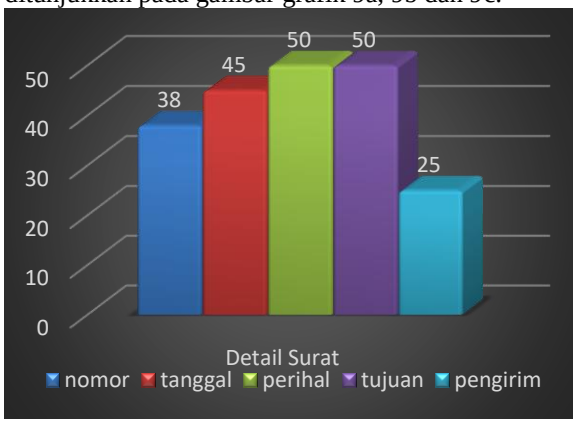


Gambar 7b. Hasil Sesudah Menggunakan Filter

Gambar 7a menunjukkan hasil OCR belum mampu untuk menemukan pengirim surat, namun setelah melalui proses filtering hasil yang diperoleh pada Gambar 7b dapat mengenali beberapa karakter yang tidak terbaca pada Gambar 7a, beberapa karakter menunjukkan pengirim surat. Hasil penelusuran detail surat menggunakan pendekatan *Thesaurus* ditunjukkan pada gambar 8.

Gambar 8. Hasil Penelusuran Detail Surat

Pengujian pada penelitian ini dilakukan dengan mencocokkan hasil penelusuran detail surat yaitu nomor, tanggal, perihal, tujuan, pengirim surat yang ditemukan oleh sistem melalui proses OCR dan pencocokan thesaurus dengan detail surat asli yang diinputkan secara manual. Skenario yang digunakan yaitu pengujian terhadp 50 data uji dengan menggunakan 10, 30 dan 50 data thesaurus. Hasil yang diperoleh pada proses pengujian ditunjukkan pada gambar grafik 9a, 9b dan 9c.



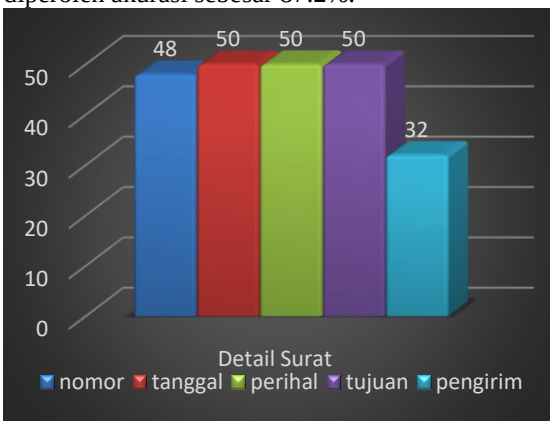
Gambar 9a. Grafik Pengujian Terhadap Data Surat menggunakan 10 Thesaurus

Grafik pada Gambar 9a menunjukkan dari 50 data uji terdapat 38 citra surat terdeteksi nomor surat yang sesuai, 45 citra surat terdeteksi tanggal surat yang sesuai, 50 citra surat terdeteksi perihal surat yang sesuai, 50 citra surat terdeteksi tanggal surat yang sesuai dan 25 citra surat terdeteksi pengirim surat yang sesuai, dari keseluruhan pengujian diperoleh akurasi sebesar 83.2%.

Gambar 9b. Grafik Pengujian Terhadap Data Surat menggunakan 30 Thesaurus

Grafik pada Gambar 9c menunjukkan dari 50 data uji terdapat 42 citra surat terdeteksi nomor surat yang sesuai, 48 citra surat terdeteksi tanggal surat yang sesuai, 50 citra surat terdeteksi perihal surat yang sesuai, 50 citra surat terdeteksi tanggal surat yang sesuai dan 28 citra surat terdeteksi pengirim surat yang sesuai, dari keseluruhan pengujian

diperoleh akurasi sebesar 87.2%.



Gambar 9c. Grafik Pengujian Terhadap Data Surat menggunakan 50 Thesaurus

Grafik pada Gambar 9b menunjukkan dari 50 data uji terdapat 48 citra surat terdeteksi nomor surat yang sesuai, 50 citra surat terdeteksi tanggal surat yang sesuai, 50 citra surat terdeteksi perihal surat yang sesuai, 50 citra surat terdeteksi tanggal surat yang sesuai dan 32 citra surat terdeteksi pengirim surat yang sesuai, dari keseluruhan pengujian diperoleh akurasi sebesar 92%.

SIMPULAN

Berdasarkan hasil pengujian maka dapat disimpulkan bahwa akurasi terbaik diperoleh dari 3 skenario yaitu sebesar 92% dengan menggunakan 50 thesaurus, dapat disimpulkan bahwa keakuratan pengenalan detail surat pada sistem ini tergantung pada proses OCR dari citra surat dan data thesaurus yang mendukung penelusuran detail surat. adanya proses reduksi noise yang baik dapat mendukung baiknya hasil OCR pada citra surat, dan dengan basis thesaurus yang lengkap dan menyesuaikan dengan kondisi data setelah hasil OCR maka dapat mendukung keakuratan pengenalan yang dihasilkan.

DAFTAR PUSTAKA

[1] D.Putra, Pengolahan Citra Digital. Yogyakarta: Andi, 2010.

[2] R. C. Gonzalez and R. E. Woods, Digital Image Processing. USA: Pearson/Prentice Hall, 2008.

[3] K. Apriyanti and T. Wahyu Widodo, "Implementasi Optical Character Recognition Berbasis Backpropagation untuk Text to Speech Perangkat Android," IJEIS (Indonesian J. Electron. Instrum. Syst., vol. 6, no. 1, p. 13, 2016.

[4] M. A. Alghamdi, I. S. Alkhazi, and W. J. Teahan, "Arabic OCR evaluation tool," Proc. - CSIT 2016 2016 7th Int. Conf. Comput. Sci. Inf. Technol., 2016.

[5] I. M. A. Mahawan and A. Harjoko, Pattern recognition of balinese carving motif using learning vector quantization (LVQ), vol. 788. 2017.

[6] A. Kadir and A. Susanto, Teori dan Aplikasi Pengolahan Citra Digital. Yogyakarta: Andi, 2013.