Implementasi Teknik Identifikasi Teks Pada Image

Amirah[1)], Salman[2)]
[1)] Teknik Informatika STMIK Dipanegara Makassar
Email : amirah01.am@gmail.com[1)]
[2)]Sistem Informasi STMIK Dipanegara Makassar
Email : salmanhannake@gmail.com[2)]

Abstrak

Keinginan manusia yang membutuhkan suatu sistem aplikasi yang dapat bekerja untuk membantu meringankan tugas-tugas mereka sehari-hari menjadi salah satu faktor yang mendorong perkembangan teknologi yang sangat pesat. Seiring semakin pesatnya kemajuan teknologi dan semakin meningkatnya kebutuhan hidup manusia. Dari banyak aplikasi yang berkembang salah satunya adalah metode pengenalan karakter dari sebuah citra atau yang lebih sering disebut OCR (Optical Character Recognitions). Dengan memanfaatkan metode OCR (Optical Character Recognitions) dalam mendeteksi sebuah citra teks untuk mengkonversi ke teks asli ini dapat mempermudah dalam mengelolah dokumen teks analog ke teks digital tanpa mengetik ulang dokumen teks tersebut ke Komputer atau Laptop dengan memanfaatkan kamera Laptop untuk mengcapture atau menginput gambar teks langsung dari penyimpanan internal dan memperoses hasil capture teks tersebut untuk di konversi menjadi teks digital.

Kata Kunci : OCR (Optical Character Recogition) , Identifikasi, Teks, Image

Abstract

of technology. Along the rapid advances in technology and the increasing needs of human life. Of the many applications that developed one of which is a method of character recognition of an image or the more commonly called OCR (Optical CharacterRecognitions). By utilizing Human desire which require an application system that can work to help the ease the tasks of their day-to-day to be one of the factors that encourage the rapid development the methods of OCR (Optical Character Recognitions) in detecting an image of the text to convert to the original text can be easier to manage text documents analog to digital text without retyping text documents into a computer or laptop by using the camera Laptop to capture or input text image directly from internal storage and to process the captured text to be converted into digital text.

***Keywords** : OCR (Optical Character Recognition), text, identification, image*

1. Pendahuluan

Teknologi yang terus berkembang membuat sistem komputerisasi bergerak dengan cepat, namun hal ini tidak seimbang dengan kemampuan manusia memindahkan data secara manual ke dalam komputer untuk dapat diolah lebih lanjut. Keinginan manusia yang membutuhkan suatu sistem aplikasi yang dapat bekerja untukmembantu meringankan tugas-tugas mereka sehari-hari menjadi salah satu faktor yang mendorong perkembangan teknologi yang sangat pesat. Seiring semakin pesatnya kemajuan teknologi dan semakin meningkatnya kebutuhan hidup manusia. Dari banyak aplikasi yang berkembang salah satunya adalah metode pengenalan karakter dari sebuah citra atau yang lebih sering disebut *OCR (Optical Character Recognitions)*. Suatu sistem dikembangkan untuk menjawab permasalahan tersebut. OCR merupakan aplikasi dari teknologi pengenalan teks, yaitu suatu teknologi yang mampu mengenali teks pada citra digital dan mengalihkannya pada dokumen digital.

Dengan memanfaatkan metode *OCR (Optical Character Recognitions)* dalam mendeteksi sebuah citra teks untuk mengkonversi ke teks asli ini dapat mempermudah dalam mengelolah dokumen teks analog ke teks digital tanpa mengetik ulang dokumen teks tersebut ke Komputer atau Laptop dengan memanfaatkan kamera Laptop untuk mengambil gambar dan menginput gambar teks langsung dari

penyimpanan internal dan memperoses hasil tersebut untuk di konversi menjadi teks digital. Dengan adanya aplikasi ini diharapkan bisa mempermudah user untuk melakukan proses konversi dari gambar yang berupa dokumen teks menjadi dokumen teks yang bisa dilakukan pembaharuan pada dokumen tersebut.

a. Teks

Kata teks dalam bahasa Inggris ditulis dengan text. Pada dasamya kata tersebut berarti *something woven* atau Sesuatu yang ditenun atau dirangkai. Dalam bahasa Latin *texere*, *texrum* berarti *to weave* yailu menenun atau merangkai. Dari hal tersebut dapat kita lihat kaitannya antara teks, tekstil ("dapat dirangkai") dan tekstur, semuanya dapat dirangkai sehingga membentuk sebuah pola. Teks dapat terdiri dari beberapa kata, namun dapat pula terdiri dari milyaran kata yang tertulis dalam sebuah naskah berisi cerita yang panjang [5].

Sebuah leks terkadang diartikan sebagai kalimat super, yang merupakan unit gramatikal yang lebih panjang dari sebuah kalimat dan saling berhubungan salu dengan yang lain.

b. *Image* (Citra)

Image adalah suatu representasi (gambaran), kemiripan, atau imitasi dari suatu objek. *Image* sebagai keluaran suatu sistem perekaman data dapat bersifat optik berupa foto, bersifat analog berupa sinyal-sinyal video seperti gambar pada monitor televisi, atau bersifat digital yang dapat langsung disimpan pada suatu media penyimpan. [6]. Menurut arti secara harfiah, *Image* (Citra) adalah gambar pada bidang dua dimensi. Ditinjau dari sudut pandang matematis, *Image* merupakan fungsi menerus (*continue*) dari intensitas cahaya pada bidang dua dimensi. Sumber cahaya menerangi objek, objek memantulkan kembali sebagian dari berkasi cahaya. Pantulan cahaya ini ditangkap oleh alat-alat optik, seperti mata pada manusia, kamera, pemindai (scanner), dan lain-lain sehingga bayangan objek dalam bentuk *Image* dapat terekam. *Image* sebagai output dari suatu sistem perekaman data dapat bersifat:

1. Optik, berupa foto,
2. Analog berupa sinyal video, seperti gambar pada monitor televisi,
3. Digital yang dapat langsung disimpan pada suatu pita magnetic.

Image dapat dikelompokkan menjadi dua bagian yaitu *Image* diam dan *Image* bergerak *Image* diam adalah *Image* tunggal yang tidak bergerak. Sedangkan *Image* bergerak adalah rangkaian *Image* diam yang ditampilkan secaraberuntun (sekuensial) sehingga memberi kesan pada mata sebagai gambar yang bergerak. Setiap *Image* didalam rangkaian itu disebut frame. Gambar-gambar yang tampak pada film layar lebar atau televisi pada hakikatnya terdiri dari ratusan sampai ribuan frame.

c. *OCR* (*Optical Character Recognition*s)

Teknologi pengenalan teks merupakan teknologi yang mampu mengenali teks pada citra digital dan mengalihkannya pada dokumen digital. Aplikasi dari teknologi pengenalan teks ini dikenal dengan nama *Optical Character Recognition* (OCR). OCR adalah sebuah sistem komputer yang dapat membaca huruf, baik yang berasal dari sebuah pencetak (printer atau mesin ketik) maupun yang berasal dari tulisan tangan[2]. Adanya sistem pengenal huruf ini akan meningkatkan fleksibilitas ataupun kemampuan dan kecerdasan sistem komputer. Dengan adanya sistem OCR maka user dapat lebih leluasa memasukkan data karena user tidak harus memakai papan ketik tetapi bisa menggunakan pena elektronik untuk menulis sebagaimana user menulis di kertas. Adanya OCR juga akan memudahkan penanganan pekerjaan yang memakai input tulisan seperti penyortiran surat di kantor pos, pemasukan data buku di perpustakaan, dll. Adanya sistem pengenal huruf yang cerdas akan sangat membantu usaha besar-besaran yang saat ini dilakukan banyak pihak yakni usaha digitalisasi informasi dan pengetahuan, misalnya dalam pembuatan koleksi pustaka digital, koleksi sastra kuno digital, dll. OCR dapat dipandang sebagai bagian dari pengenal otomatis yang lebih luas yakni pengenal pola otomatis (automatic pattern recognition).

OCR (*Optical Character Recognition*) merupakan suatu proses mengkonversi *scanned image* ke dalam suatu *editable text*. *Scanned image* yang dimaksudkan di sini yaitu gambar/ *Image* yang dimasukkan ke dalam komputer melalui sebuah alat *scanner* maupun hasil pemotretan melalui kamera. Gambar / *Image* ini berisi karakter-karakter, teks, atau, simbol yang akan diproses oleh komputer, dikenali dan kemudian dikonversikan ke kode-kode karakter seperti *ASCII* maupun *unicode* lainnya seperti *UTF-8*. Setelah pengkonversian ini, karakter-karakter dalam *image* tersebut bukan lagi berbentuk *image* yang tidak dapat di-*edit*, namun sebaliknya telah menjadi teks yang dapat di-*edit*, disalin dan digunakan untuk keperluan

apapun, salah satunya yang untuk diterjemahkan ke bahasa lain.

Tingkat keberhasilan dari perangkat lunak aplikasi pengenalan teks ini sangat bergantung dari sejumlah faktor berikut.[4]

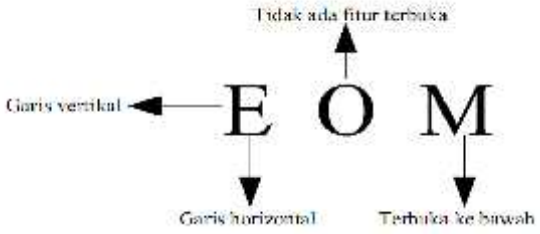
1. Kualitas gambar teks yang ada pada dokumen yang dibaca serta tingkat kompleksitasnya (ukuran, format teks, warna, latar belakang).
2. Kualitas alat optik yang dipakai (scanner).
3. Kualitas perangkat lunak aplikasi pengenalan teks itu sendiri.

Gambar 1 menjelaskan proses umum yang dilakukan oleh OCR. [4]

Gambar 1 *OCR* (*Optical Character Recognition)*

Dalam sistem OCR, input yang dimasukkan adalah gambar yang berisi karakter yang ingin kenali. Sebelum karakter pada gambar dapat dikenali, gambar tersebut harus terlebih dahulu dilakukan proses ini ialah menghilangkan bagian-bagian dari gambar yang tidak di inginkan (bukan bagian dari karakter yang ingin dikenali) dan untuk memperbaiki kualitas gambar sehingga objek pada gambar lebih mudah untuk dikenali. *Preprocessing* pada gambar biasanya meliputi *grayscalin*, *noise removal*, dan *thresholdin. Grayscalling* merupakan tahap awal dalam *image preprocessin*, yaitu mengubah gambar berwarna menjadi gambar yang hanya memiliki derajat keabuan saja. Selanjutnya dilakukan proses *noise filtering*, yaitu proses meredukti atau megurangi noise. Proses akhir dari *image processing* adalah *thresholding*, yaitu suatu proses untuk memisahkan background dengan objek yang ingin diamati dengan mengubah gambar menjadi hitam putih.

Setelah preprocessing selesai dilakukan, maka tahap selanjutnya ialah segmentasi. Proses ini digunakan untuk memisahkan area-area pengamatan dari setiap karakter yang ingin dikenali, seperti pemisahan kalimat ke dalam kata-kata dan pemisahan kata kedalam kata ke dalam karakter-karakter. Tahap selanjutnya ialah normalisasi karakter hasil segmentasi, yaitu proses untuk mengubah dimensi *region* dari setiap karakter, seperti ketebalan karakter. Hal yang biasa dilakukan untuk proses normalisasi pada OCR adalah *scaling* dan *thinning*. Setelah normalisasi dilakukan, kemudian akan dilakukan ektraksi fitur untuk mendapatkan karakteristik dari masing-masing karakter dilakukan ektraksi fitur untuk mendapatkan karakteristik dari masing-masing karakter yang membedakannya dengan karakter lain. Tahap akhir dari proses OCR adalah pengenalan atau recognising, dalam tahap ini algoritma akan membandikngkan ciri-ciri fitur yang ingin kenali dengan data yang tersimpan sebelumnya. Hasil pengenalan berupa teks dengan kemiripan paling besar antara fitur karakter yang ingin dikenali dengan informasi yang tersimpan. *OCR* (*Optical Character Recognition*) menggunaan metode *algoritma feature extraction* . *Algoritma* feature extraction merupakan salah satu cara untuk mengenali suatu objek dengan melihat ciri-ciri khusus yang dimiliki objek tersebut. Tujuan dari feature extraction adalah melakukan perhitungan dan perbandingan yang bisa digunakan untuk mengklasifikasikan ciri-ciri yang dimiliki oleh suatu citra.



Gambar 2. Ilustrasi feature extraction

Gambar 2 merupakan ilustrasi citra karakter dengan ciri-cirinya. Ciri-ciri dari masing-masing citra template akan di simpan. Citra masukan yang akan dibandingkan akan dianalisis berdasarkan ciri-ciri citra. Ciri-ciri yang dimiliki citra masukan akan diklasifikasikan terhadap ciri-ciri citra template.

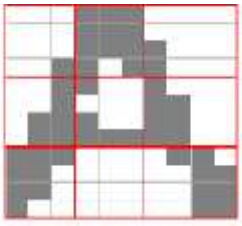
Pemetaan ciri-ciri khusus yang dilakukan terhadap citra karakter adalah keterbukaan citra, jumlah garis vertikal dan horizontal, jumlah perpotongan pixel hitam di bagian tengah citra secara vertikal dan horizontal, rasio citra, dan histogram pixel hitam pada sembilan bagian citra yang dibandingkan dengan resolusi citra karakter.

Algoritma feature extraction adalah sebagai berikut:

1. Pemetaan fitur citra karakter. Pemetaan ini dilakukan satu kali saat citra karakter akan dikenali.
2. Perhitungan jarak dengan menggunakan persamaan. Perhitungan yang dilakukan sebanyak citra template dinotasikan dengan n.
3. Pencarian jarak minimum dilakukan pada seluruh citra template. Jumlah pencarian yang dilakukan terhadap seluruh citra template dinotasikan dengan n.
4. Tabel dibawah ini adalah perhitungan jumlah operasi dasar yang dilakukan pada algoritma feature extraction.
5. Berdasarkan hasil perhitungan jumlah operasi dasar feature extraction pada tabel diatas maka kompleksitas algoritma feature extraction adalah 2n+1.

Dari lima proses pemetaan fitur pada algoritma feature extraction, masih ada kemungkinan penambahan fitur-fitur khusus citra karakter. Salah satu contohnya adalah fitur jumlah stroke (jumlah garis yang membentuk karakter). Berdasarkan hasil pengenalan, algoritma feature extraction dapat mengenali citra karakter lebih baik dan algoritma feature extraction membutuhkan waktu yang lebih singkat. Sebelum mengambil ciri dari suatu citra karakter, maka citra karakter tersebut harus dicari batas kanan, kiri, atas, dan bawahnya terlebih dahulu. Ciri-ciri yang diambil dari citra karakter template maupun citra karakter masukan adalah jumlah garis horizontal dan garis vertikal, perpotongan pixel hitam terhadap garis vertikal dan garis horizontal di tengah citra. Citra karakter dianalisis apakah terbuka ke kanan, kiri, atas, atau bawah. Karakter citra akan dibagi menjadi sembilan bidang simetris kemudian pixel yang berwarna hitam pada masing-masing bidang akan dibagi dengan jumlah pixel keseluruhan.

Perbandingan Pixel Hitam Di Setiap Blok Citra Citra karakter dibagi menjadi sembilan bagian simetris dan dihitung perbandingan jumlah pixel hitam pada masing-masing bagian tersebut dengan jumlah seluruh pixel.



Gambar 3. Citra karakter yang dibagi menjadi sembilan blok.

Bagian yang pertama dihitung adalah blok pada bagian kiri atas citra karakter. Jika ditemukan pixel berwarna hitam, maka pixel tersebut akan diwakili dengan angka 1. Sedangkan jika ditemukan pixel berwarna putih, maka pixel tersebut akan diwakili dengan angka 0. Jumlahkan angka-angka pixel yang terdapat di blok pertama dan dibagi dengan resolusi citra karakter. Hal yang sama dilakukan pada seluruh blok. Blok kedua terletak di bawah blok pertama. Blok ketiga terletak di bawah bagian blok kedua. Blok

keempat terletak di sebelah kanan blok pertama dan begitu seterusnya hingga mencapai blok terakhir yaitu blok kesembilan yang berada di bagian kanan bawah citra.

d. Pencarian Teks-Line Dan Kata

Algoritma *line finding* dirancang supaya halaman yang miring dapat dikenali tanpa harus *de-skew (*proses untuk mengubah halaman yang mirip menjadi tegak lurus) sehingga tidak menurunkan kualtas gambar Kunci proses dari algoritma ini adalha *blob filtering* dan *line construction* (konstruksi baris).

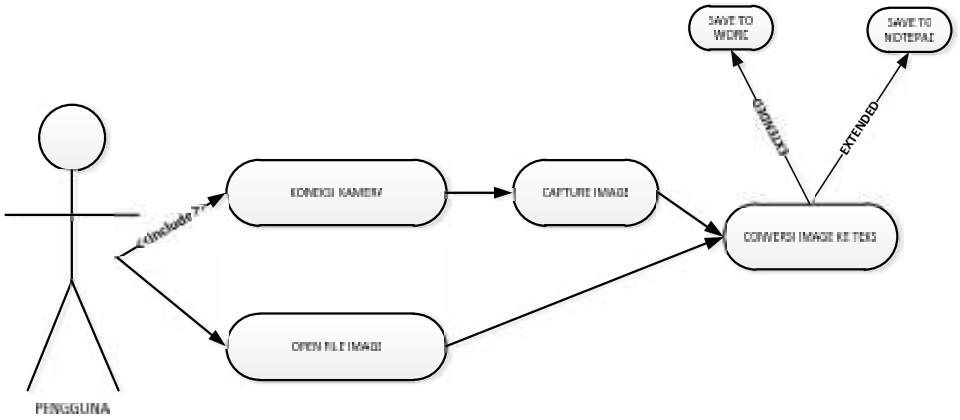
Algoritma pencarian text-line bekerja secara independen untuk setiap *region* teks dari hasil analisis *layout* dan dimulai dengan mencari ketetanggan dari CC kecil (relatif terhadap perkiraan ukuran huruf) unutuk menemukan *body-text-sized* CC terdekat. Jika tidak ada yag dekat dengan *body-text-sized* CC, maka CC kecil ini akan dianggap sebagai noise dan dbuang (pengecualian harus dilakukan untuk titik-titik/garis putus-putus yang biasanya ditemukan daftar isi. Jika tidak, sebuah *bounding box* yang berisi CC kecil dan *neighbor* yang lebih besar dibangun dan digunakan di *bounding box* CC kecil pada prjeksi selanjutnya.

Sebuah profil projeksi horizontal dibangun, paralel terhadap kemiringan horizontal yang diperkirakan, dari *bounding box CC* menggunakan *boxes* yang sudah dimodifikasi unutk CC kecil. Algoritma *dynamic progrmming* kemudian memilih *set* terbaik dari titik-titik segmentasi pada profil projeksi. Setelah garis-garis potong telah ditentukan, Sebuah CC ditempatkan pada *text-line* untuk CC yang saling tumpag tindih secara vertikal. Setelah baris teks diekstrak, *blob* pada garis disusun menjadi unit-unit pengenalan.

2. Metode Penelitian

2.1 Analisis dan Perancangan Sistem

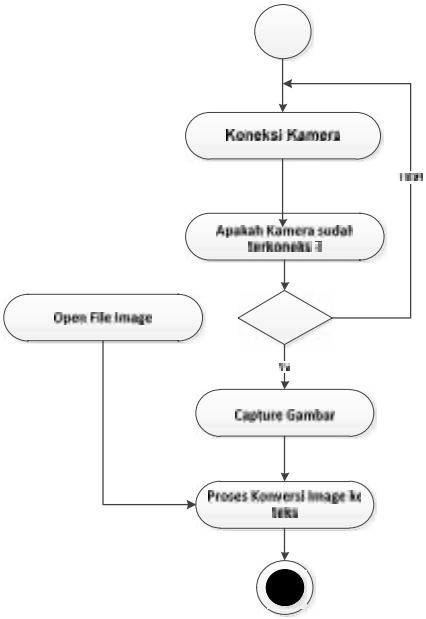
Metode yang digunakan adalah metode perancangan berbasis object melalui tahapan Pembuatan UML (Unified Modeling Language). Adapun gambar sistem yang diusulkan sebagai berikut :



Gambar 4. Usecase yang diusulkan

Activity Diagram

Pada aktivitas berikut ini menjelaskan tentang aliran proses user ketika ingin melakukan identifikasi teks pada image.

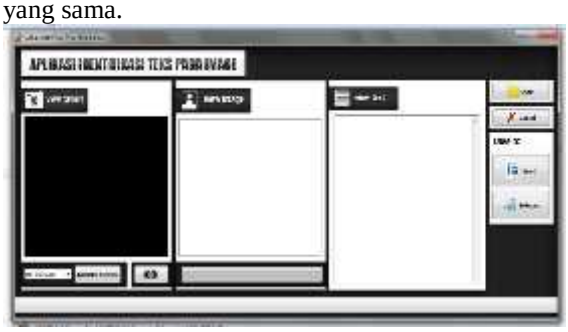


Gambar 5. Activity Diagram

3. Hasil dan Pembahasan

3.1 Rancangan *Input* dan *Output*

Tampilan from ini merupakan tampilan awal form *input* dan *output*, semua form input dan output diproses pada form yang sama.


Gambar 6. *Interface* aplikasi

3.1.1 Rancangan Input

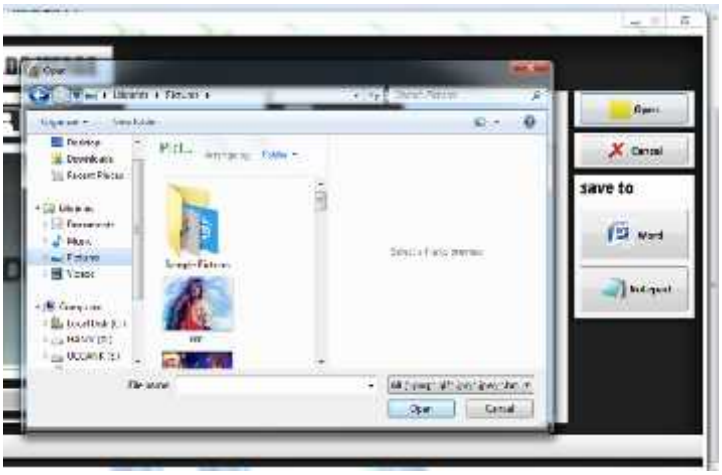
a. Tampilan View Camera

ini adalah *view camera* menampilkan hasil gambar dari camera serta proses mengambil gambar / capture gambar.

Gambar 7. Rancangan input view camera

b. Tampilan Input / Open File Image

Ini adalah proses memasukkan file gambar untuk di proses menjadi teks.



Gambar 8. Rancangan input file image

3.1.2 Rancangan Output

a. Tampilan Interface Output

Tampilan ini adalah menampilkan hasil identifikasi teks.

Gambar 9 Rancangan interface output

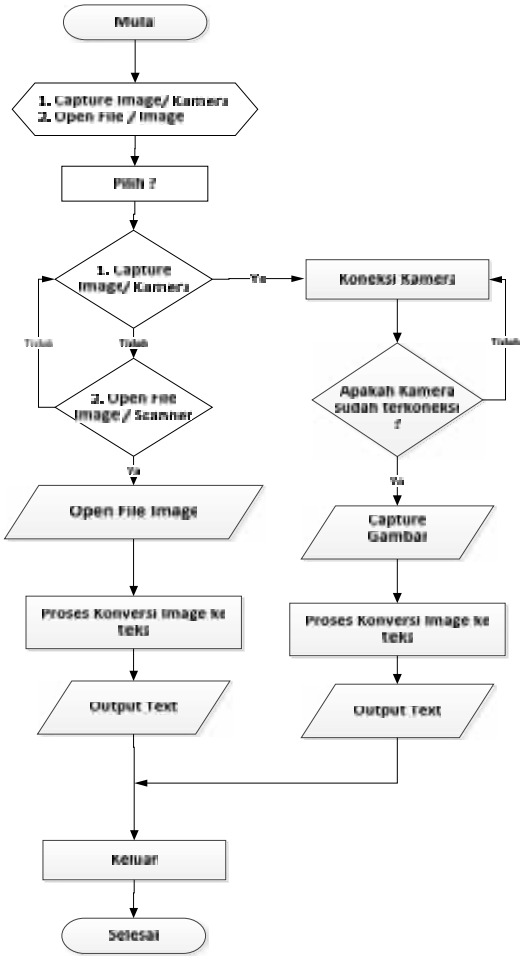
b. Tampilan Save File Word dan Notepad

Tampilan ini adalah proses penyimpanan file berupa format txt dan docx.

Gambar 10. Tampilan save file word dan notepad

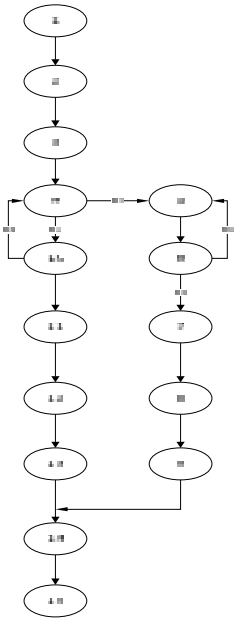
3.2 Pengujian White-Box

a. Flowchart Identifikasi Teks Pada Image



Gambar 11. Flowchart identifikasi teks pada image

b. Flowgraph Identifikasi Teks Pada Image



Gambar 12. Flowgraph identifikasi teks pada image

Dari gambar 12 di atas dapat dilakukan proses perhitungan sebagai berikut :

1) Untuk menghitung *cyclomatic complexcity* V(G) = E – N + 2

E (edge) = 17
N (note) = 15
V(G) = 17 – 15 + 2
= 2 + 2
= 4

2) *Path-path* yang ada di Aplikasi Identifikasi Teks Pada Image, yaitu :
Path 1 = 1,2,3,4,10,4
Path 2 = 1,2,3,4,5,6,3
Path 3 = 1,2,3,4,5,6,7,8,9,14,15
Path 4 = 1,2,3,4,10,11,12,13,14,15

c. Pengujian Identifikasi Teks Menggunakan Scanner

Sample Menggunakan Scanner Image Teks Hitam Background Putih

Gambar 13. Sample menggunakan scanner image teks hitam background putih

Sample Menggunakan Scanner Image Teks Modif

Gambar 14 Sample menggunakan scanner image teks modif

Sample Menggunakan Scanner Image Teks Modif Dengan Background Warna

Gambar 15 Sample menggunakan scanner image teks modif background warna

4. Kesimpulan

Dari hasil pengujian yang dilakukan dengan menggunakan teknik pengujian white-box diperoleh diperoleh jumlah region = 4, jumlah jalur independent = 4 dan cyclomatic complexity = 4, karna nilai ketiganya sama maka dapat disimpulkan bahwa sistem yang diuji telah bebas dari kesalahan logika .

Dari hasil pengujian sample akurasi proses identifikasi teks pada image menggunakan kamera, didapatkan bahwa image teks yang berwarna hitam dengan background putih didapatkan akurasi yang baik dalam identifikasi teks tersebut, dan juga didapatkan warna yang tidak mempengaruhi teks terebut dapat diidentifikasi dengan baik oleh aplikasi. Serta kualitas pixel dan pencahayaan dari kamera

mempengaruhi kualitas dari Identifkasi image ke teks tersebut sedang teks-teks pada image yang mempunyai banyak kombinasi warna akan mempengaruhi kualitas indentifikasi image ke teks.

Dari hasil pengujian sample akurasi proses identifikasi teks pada image menggunakan scanner, didapatkan bahwa image teks yang dapat di identifikasi dengan baik, karena pixel dari scanner tersebut mempunyai kualitas yang baik sedangkan teks pada image yang mempunyai banyak kombinasi warna akan mempengaruhi kualitas indentifikasi image ke teks.

DAFTAR PUSTAKA

[1] Adi Nugroho.Rekayasa Perangkat Lunak Berorientasi Objek dengan Metode USDP. Yogyakarta:Andi Ofset. 2010

[2] Hartanto, Suryo. "Optical Character Recognition menggunakan Algoritma Template Matching Correlation". Dalam Journal of Informatics and Technology, 2015; Vol 1, No 1, p 11-20.

[3] Kristanto, Andri. Rekayasa Perangkat Lunak (Konsep Dasar). Yogyakarta: Gava Media. 2004

[4] Mori S., Nishida H., Yamada H., "Optical Character Recognition", Willey Interscience. 1999

[5] Sudardi, Bani. Dasar-Dasar Teori Filologi :Surakarta: Badan Penerbit Sastra Indonesia Fakultas Sastra Universitas Sebelas Maret.2001

[6] Sutoyo, T, dkk. 2009. Teori Pengolahan Citra Digital. Penerbit Andi: Yogyakarta

[7] Usman Ahmad. Pengolahan Citra Digital Dan Teknik Pemrogramannya. Yogyakarta: Gava Media.2005

[8] Wahana Komputer. *Step by step Delphi 2010*. Semarang: Andi Offset. 2010

[9] Zand, M., Nilchi, AN., Monadjemi, SA. Recognition-based Segmentation in Persian Character Recognition. International Journal of Computer and Information Science and Engineering. 2008; Vol 2(1), pp14-18.