

BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Optical Character Recognition (OCR) merupakan proses konversi gambar menjadi karakter ASCII yang dikenali oleh komputer. Gambar huruf yang dimaksud dapat berupa hasil scan dokumen, hasil *print-screen* halaman web, hasil foto, dan lain sebagainya [1]. OCR adalah solusi yang efektif untuk proses konversi dokumen cetak ke dokumen digital [2]. Manfaat adanya OCR yaitu kita tidak perlu mengetik ulang teks yang terkandung dalam sebuah gambar, karena sistem OCR mampu mengenalinya. Sebelum teknologi OCR tersedia, satu-satunya pilihan untuk mendigitalkan dokumen cetak adalah dengan mengetik ulang teks secara manual. Cara tersebut bukan hanya menghabiskan waktu tapi juga sangat mungkin terjadinya *typo* dalam pengetikan. Tujuan utama dari OCR sendiri yaitu untuk mengenali teks yang berada pada dokumen yang tidak terdapat salinan *softcopy*-nya, hal ini bisa dikarenakan kehilangan, kerusakan file, ataupun dokumen yang memang tidak tersedia *softcopy*. Selain itu OCR juga diperlukan agar informasi yang dihasilkan dapat diolah sesuai dengan kebutuhan. Maka dari itu, dibutuhkan sistem OCR yang dapat mengenali suatu karakter yang terdapat dalam gambar, agar dapat menghasilkan informasi yang jelas dan akurat. Tetapi beberapa masalah sering terjadi pada sistem OCR, seperti sulitnya mengenali beberapa karakter yang memiliki bentuk hampir sama seperti huruf O dengan angka 0, ataupun huruf E dengan huruf B [3].

Beberapa penelitian telah dilakukan untuk mengenali karakter pada citra diantaranya yaitu penelitian yang dilakukan oleh Mohamed Ali tentang pengenalan tulisan tangan aksara arab menggunakan *Learning Vector Quantization* menunjukkan bahwa tingkat akurasi yang dihasilkan oleh LVQ sebesar 82,2% [4]. Selanjutnya yaitu penelitian yang dilakukan oleh Muhammad Fajry Hamzah yaitu tentang pengenalan tulisan dan ekstraksi informasi pada citra abstrak skripsi menggunakan *support vector machine* dan *rules based system* yang menghasilkan tingkat akurasi 5,02% untuk case sensitive dan 5,4% untuk case insensitive,

sedangkan untuk tingkat akurasi pengenalan karakter menggunakan SVM itu sendiri mencapai 46.61% untuk case sensitive dan 54.30% untuk case insensitive [5]. Penelitian lainnya oleh Rizki Rahmat, Youllia Indrawaty dan Irma Amelia yaitu tentang Sistem Pengenalan Aksara Sunda Menggunakan Metode *Modified Direction Feature* dan *Learning Vector Quantization* yang menghasilkan tingkat akurasi 78,67 [6].

Metode *Learning Vector Quantization* sendiri sudah pernah digunakan dalam mengklasifikasikan beberapa kasus pada citra, seperti pengenalan plat nomor [7], pengenalan pola buku [8], pengenalan tulisan tangan latin bersambung [9], dengan nilai keakuratan rata-rata sebesar 80%. Berdasarkan penelitian-penelitian tersebut, dapat disimpulkan bahwa metode *Learning Vector Quantization* (LVQ) dapat digunakan untuk pengklasifikasian citra.

Berdasarkan beberapa penelitian tersebut, terdapat beberapa permasalahan yang sama, yaitu sulitnya mengklasifikasikan karakter citra yang memiliki ciri serupa, contohnya O dan angka 0 atau I, l dan angka 1, hal tersebut dapat diatasi dengan ekstraksi ciri sebelum pengklasifikasian karakter, agar didapatkan ciri-ciri unik dari setiap karakter sehingga pengklasifikasian akan lebih mudah. Dan metode LVQ akan digunakan untuk pengklasifikasian citra, oleh karna itu pada penelitian ini akan menggunakan ekstraksi ciri zoning dan metode *Learning vector Quantization* (LVQ) untuk melakukan *Optical Character Recognition* (OCR) pada dokumen karya tulis ilmiah, yang selanjutnya diharapkan sistem dapat mengenali karakter dengan baik dan tingkat akurasi yang dihasilkan tinggi.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang di atas maka diidentifikasi masalah yang didapatkan pada penelitian ini adalah rendahnya akurasi yang diperoleh dari penelitian sebelumnya tentang OCR pada citra dokumen abstrak skripsi.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk mengimplementasikan *Optical Character Recognition* (OCR) pada dokumen karya tulis ilmiah menggunakan

ekstraksi ciri zoning dan menggunakan metode Learning Vector Quantization (LVQ) untuk pengklasifikasiannya.

Adapun tujuan dari penelitian ini, yaitu mengetahui akurasi yang dihasilkan dengan penggunaan ekstraksi ciri zoning dan metode *Learning Vector Quantization* (LVQ) pada kasus pengenalan citra OCR pada dokumen karya tulis ilmiah.

1.4 Batasan Masalah

Agar penelitian yang dilakukan terarah sesuai dengan tujuan yang diinginkan, maka berikut adalah batasan masalah yang digunakan :

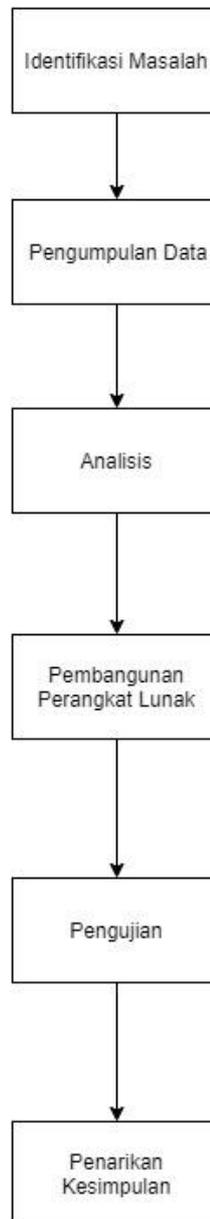
1. Data Masukan (*Input*)
 - a. Data latih yang digunakan adalah alpabet kapital (A-Z), huruf kecil (a-z), angka (0-9) dan simbol (.,:;/[]%*'()?!#+-=) dengan jenis font *Times New Roman* ukuran 15.
 - b. Data latih diambil dengan cara megubah font berformat .ttf menjadi gambar-gambar karakter yang diinginkan.
 - c. Format dari data latih yaitu .png/.jpg/.jpeg.
 - d. Data uji yang digunakan yaitu citra karya tulis ilmiah hasil *scan* atau foto berformat (.jpg/.jpeg/.png) yang diambil dari karya tulis ilmiah Teknik Informatika.
2. Proses
 - a. Metode yang digunakan pada tahapan preprocessing adalah *grayscale*, binerisasi, *resize*.
 - b. Segmentasi yang digunakan adalah *profile projection* yang terdiri dari segmentasi baris dan segmentasi karakter.
 - c. Ekstraksi fitur yang digunakan adalah *Zoning Image Centroid Zone (ICZ)*.
3. Data Keluaran (*Output*)

Keluaran yang dihasilkan yaitu berupa teks hasil pengenalan karakter.

1.5 Metode Penelitian

Dalam penelitian ini, metode yang digunakan adalah metode deskriptif [10]. Metode ini digunakan karena pembangunan sistem yang dilakukan berdasarkan dari

gambaran fenomena-fenomena yang ada, serta sistematis dan terstruktur. Berikut adalah langkah-langkah yang dilakukan dalam penelitian.



Gambar 1.1 Metode Penelitian

Berdasarkan alur penelitian pada Gambar 1.1 dapat dijelaskan sebagai berikut.

1.5.1 Identifikasi Masalah

Tahap pertama dalam penelitian adalah perumusan masalah. Berkaitan dengan kegiatan menganalisa permasalahan yang muncul mengenai kendala pada

proses pengenalan karakter pada karya tulis, serta apa saja faktor-faktor yang mempengaruhi terhadap masalah yang ada.

1.5.2 Pengumpulan Data

Tahapan kedua adalah pengumpulan data. Metode pengumpulan data yang dilakukan pada penelitian ini yaitu.

a. Studi Litelatur

Studi litelatur dilakukan dengan membaca jurnal, paper, buku dan sumber lain yang mendukung penelitian terkait topik OCR, dan juga metode LVQ. Dengan tujuan untuk peahaman dasar peneliti terhadap kasus yang diteliti.

b. Pengumpulan Dataset

Pada tahap ini diambil data latih dan juga data uji. Data latih yaitu berisi data karakter yang berjumlah 80 citra yang terdiri dari huruf A-Z,a-z angka 0-9 dan karakter tambahan berupa simbol (.,:;[]%*'()?!#+-=). Dan untuk data uji adalah hasil *scan* dokumen karya tulis ilmiah Teknik Informatika, akan dikumpulkannya data uji sebanyak 40 hasil *scan* dokumen yang akan digunakan sebagai data uji.

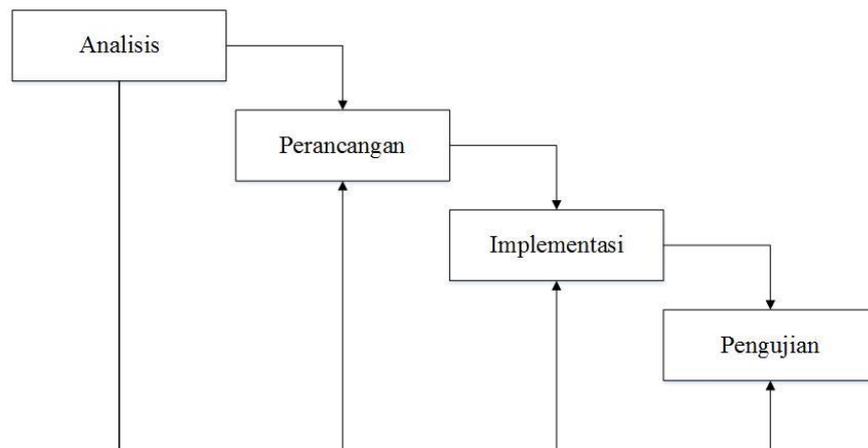
1.5.3 Analisis

Selanjutnya setelah dataset terkumpul kemudian masuk ke dalam tahap analisa, dimana tahapan ini data litelatur dan referensi yang sudah terkumpul dipelajari terkait metode yang akan digunakan yaitu LVQ, dan fitur apa saja yang dibutuhkan dan hal lainnya terkait pembangunan sistem termasuk menganalisa bagaimana sistem akan berjalan mulai dari tahapan *input* kemudian proses hingga hasil berupa *output* dari sistem berupa teks hasil pengenalan. Kemudian dataset yang telah terkumpul dibagi atau dipisahkan mana yang akan dijadikan datalatih dan juga data uji.

1.5.4 Pembangunan Perangkat Lunak

Dalam pengembangan sistem, metode yang digunakan adalah metode *waterfall* karena proses dalam membangun sistem yang dilakukan secara terurut.

Tahap perancangan tidak dapat dilakukan jika tahap analisis belum dilakukan, begitupun dengan tahap implementasi dan pengujian tidak dapat dilakukan jika tahap analisis belum dilakukan. Adapun tahapan metode *waterfall* dapat dilihat pada Gambar 1.2 [11].



Gambar 1.2 Diagram *Waterfall Sommerville* [11]

a. Analisis Kebutuhan Perangkat Lunak

Langkah ini merupakan tahapan analisis terhadap kebutuhan perangkat lunak, dan tahap dalam pengumpulan kebutuhan penelitian secara menyeluruh yaitu pengenalan teks karya tulis ilmiah. Bagaimana nantinya sistem akan dibangun dan berkerja dengan metode serta algoritma yang sudah ditentukan.

b. Perancangan

Setelah kebutuhan data dan pemrosesan telah terkumpul, maka tahap selanjutnya adalah perancangan arsitektur sistem yang akan dibuat. Mulai dari perancangan struktur tabel, perancangan struktur menu dan perancangan tampilan antarmuka

c. Implementasi

Setelah perancangan sistem secara keseluruhan, maka tahap selanjutnya adalah mengimplementasikan kedalam bahasa pemrograman algoritma yang sudah ditetapkan. Tahapan dimana algoritma *Learning Vector Quantization* (LVQ) digunakan dalam menyelesaikan permasalahan mengenai pengenalan karakter pada citra karya tulis ilmiah.

d. Pengujian

Tahap ini digunakan untuk mengintegrasikan program dan pengujian tujuan, apabila program yang dibuat berjalan sesuai dengan harapan, dan sesuai dengan pendekatan aturan maka pengenalan karakter karya tulis ilmiah siap digunakan oleh pengguna sistem.

1.5.5 Pengujian

Tahap selanjutnya adalah pengujian terait nilai performansi dari algoritma LVQ untuk pengenalan karakter pada citra karya tulis ilmiah.

1.5.6 Penarikan Kesimpulan

Tahap terakhir dalam penelitian ini adalah menyimpulkan hasil keseluruhan dari penelitian yang sudah dilakukan baik terkait sistem yang dibuat, penggunaan metode yang digunakan terhadap kasus hingga saran untuk penelitian selanjutnya.

1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran secara umum mengenai penelitian secara umum. Sistematika penulisan penelitian ini adalah sebagai berikut :

BAB 1 PENDAHULUAN

Bab ini akan membahas kerangka penelitian atau percobaan dalam penelitian, meliputi latar belakang permasalahan, perumusan masalah, menentukan maksud dan tujuan penelitian, batasan masalah, metode penyelesaian masalah serta sistematika penulisan dari penelitian implementasi *Optical Character Recognition* (OCR) pada Karya Tulis Ilmiah menggunakan metode LVQ.

BAB 2 LANDASAN TEORI

Bab ini memuat berbagai konsep dasar dan teori-teori yang berkaitan dengan penelitian yang dilakukan, diantaranya mengenai konsep *Optical Character Recognition*, penelitian-penelitian terdahulu, teknik *preprocessing*, ekstraksi fitur,

segmentasi, metode LVQ, pemodelan, bahasa pemrograman, dan perangkat lunak yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi tentang tahapan analisis dan perancangan, Analisis yang akan dibahas berupa analisis masalah analisis data masukan, analisis proses, analisis data keluaran, analisis basis data dan analisis kebutuhan fungsional dan nonfungsional, serta perancangan sistem yang mencakup perancangan antarmuka, struktur menu dan pesan.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini berisi tentang implementasi dan pengujian dari perangkat lunak yang dibangun sesuai dengan analisis dan perancangan yang telah dibuat. Mulai dari implementasi hardware dan software yang digunakan, implementasi antarmuka, serta hasil pengujian. Hasil pengujian yang dilakukan yaitu mengukur tingkat akurasi.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini akan diuraikan hasil dari penelitian yang telah dilakukan, serta akan membahas tentang saran yang dapat dijadikan masukan untuk penelitian selanjutnya.