BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Optical Character Recognition (OCR) merupakan sebuah proses untuk mengenali karakter berupa huruf, angka, maupun simbol-simbol yang terdapat pada suatu citra atau gambar. OCR dapat dimanfaatkan untuk mengkonversi gambar yang mengandung teks hasil cetak mesin atau printer (*scan*) menjadi file dokumen berupa teks yang bisa dibaca dan disunting oleh aplikasi komputer. Sehingga informasi yang terkandung di dalam suatu gambar dapat diambil tanpa melalui proses pengetikan ulang [1]. Sebelum teknologi OCR tersedia, satu-satunya pilihan untuk mendigitalkan dokumen kertas tercetak adalah dengan mengetik ulang teks secara manual, tetapi hal ini bisa menghabiskan banyak waktu serta bisa menyebabkan ketidakakuratan dan kesalahan pengetikan. OCR diperlukan agar bisa mengenali teks yang berada pada dokumen yang tidak terdapat salinan *softcopy*-nya. Selain itu, OCR juga diperlukan agar informasi yang dihasilkan selanjutnya dapat diolah sesuai kebutuhan dan dapat menjadi informasi yang bermanfaat. Beberapa penelitian tentang OCR telah dilakukan dengan menggunakan berbagai metode, contohnya menggunakan *Artificial Neural Network* [2], *Template Matching Correlation* [3], dan menggunakan K-*Nearest Neighbor* [4], ketiga metode tersebut menghasilkan tingkat akurasi yang beragam dalam mengenali karakter huruf cetak.

Support Vector Machine (SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 di *Annual Workshop on Computational Learning Theory*, adalah suatu sistem pembelajaran yang menggunakan ruang hipotesis dari suatu fungsi linear dalam suatu ruang dimensi berfitur tinggi [5]. Metode SVM digunakan dalam penelitian ini karena sudah digunakan pada berbagai macam penelitian pengenalan karakter yang berbeda, seperti pengenalan karakter bahasa Arab [6], *font* bahasa Inggris [7], dan pengenalan aksara Jawa [8] dengan tingkat akurasi yang dihasilkan masing-masing sebesar 95,03%, 93,54%, dan 90,84%.

Penelitian lainnya tentang SVM pada citra yaitu mengenai pengenalan tulisan dan ekstraksi informasi pada citra abstrak skripsi menggunakan *support vector machine* dan *rules based system* [9] menghasilkan tingkat akurasi sebesar 5,02% untuk *case sensitive* dan 5,47% 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*. Rendahnya tingkat akurasi pengenalan pada citra abstrak dipengaruhi oleh proses segmentasi yang kurang mampu menyelesaikan masalah yang ada pada pemisahan karakter dan kurangnya metode dalam mengekstraksi ciri citra. Sementara itu, ekstraksi fitur zoning dan SVM dengan menggunakan kernel linear telah digunakan untuk mengekstraksi ciri citra aksara sunda dengan menghasilkan tingkat akurasi sebesar 99,75% [10].

Berdasarkan beberapa penelitian tersebut, ditemukan beberapa permasalahan yang sama, yaitu sulitnya mengklasifikasikan karakter yang memiliki ciri yang mirip, contohnya yaitu huruf O dan angka 0 atau angka 1 dengan huruf i dan l. Hal tersebut dapat diatasi dengan menambahkan metode ekstraksi ciri sebelum proses klasifikasi karakter, agar didapatkan ciri-ciri unik dari masing-masing karakter sehingga lebih mudah untuk diklasifikasikan. Oleh karena itu, pada penelitian ini akan menggunakan ekstraksi ciri zoning dan metode SVM untuk melakukan pengenalan karakter pada dokumen karya tulis ilmiah dengan huruf cetak, yang selanjutnya diharapkan sistem dapat mengenali karakter dengan baik dan tingkat akurasi yang dihasilkan SVM tinggi.

1.2 Identifikasi Masalah

Berdasarkan dari latar belakang yang telah diuraikan di atas, maka dapat diidentifikasi permasalahan yang dihadapi yaitu rendahnya tingkat akurasi yang diperoleh dari penelitian sebelumnya tentang OCR pada citra dokumen abstrak skripsi.

1.3 Maksud dan Tujuan

Maksud dari penelitian ini adalah untuk membangun sistem OCR pada dokumen karya tulis ilmiah menggunakan ekstraksi fitur zoning dan metode

SVM. Adapun tujuan dari penelitian ini yaitu untuk mengukur tingkat akurasi dari penggunaan ekstraksi fitur zoning dan metode SVM pada kasus pengenalan karakter dokumen karya tulis ilmiah.

1.4 Batasan Masalah

Agar penelitian yang dilakukan lebih terarah sesuai dengan tujuan yang ingin dicapai, maka batasan masalah dalam penelitian ini yaitu sebagai berikut.

1. Masukan

   Data masukan yang digunakan dibagi menjadi 2 bagian yaitu.

   a. Data latih

      1) Format dari data latih berupa .png dan .jpg.
      2) Data latih diambil dengan cara men-*scan* citra yang mengandung karakter serta dengan mengubah *font* yang berformat .ttf menjadi gambar-gambar karakter dengan ukuran 18x18 piksel.
      3) Data yang digunakan yaitu berupa citra digital yang berjumlah sebanyak 560 data dengan 80 jenis karakter, yaitu karakter yang terdiri dari huruf A-Z, a-z, dan angka 0-9, serta karakter tambahan berupa simbol .,:;/[]%*'()?!#+-=
      4) Jenis *font* yang digunakan hanya berjenis *Times New Roman*.

   b. Data uji

      1) Format dari data uji berupa .png/.jpg/.jpeg.
      2) Data uji yang digunakan yaitu gambar hasil *scan* dokumen karya tulis ilmiah skripsi yang mengandung karakter (kecuali bagian yang mengandung rumus, gambar, dan tabel).
      3) Gaya *font* yang digunakan tidak mengandung *underline*.

2. Proses

   a. Metode *preprocessing* yang digunakan untuk tahap pelatihan yaitu *grayscaling*, *thresholding*, binerisasi, serta ekstraksi fitur.

   b. Metode *preprocessing* yang digunakan untuk tahap pengujian yaitu *grayscaling*, *thresholding*, segmentasi baris, kata dan karakter, *resize*, binerisasi, serta ekstraksi fitur.

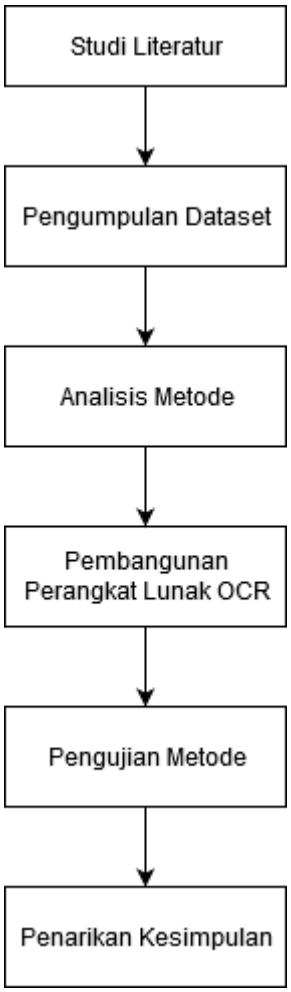
c. Ekstraksi fitur yang digunakan yaitu *Zoning Image Centroid Zone (ICZ)* untuk memisahkan ciri-ciri yang terdapat pada setiap karakter.

3. Data Keluaran

   Keluaran yang dihasilkan yaitu berupa teks hasil pengenalan karakter.

1.5 Metode Penelitian

Metode penelitian yang digunakan dalam penelitian ini adalah metode kuantitatif [11]. Metode ini digunakan karena data yang digunakan dalam penelitian ini berbentuk angka dan bersifat fakta serta bisa diukur secara akurat dengan alat yang objektif.



Gambar 1.1 Alur Penelitian

1.5.1 Studi Literatur

Studi ini dilakukan dengan cara mempelajari, meneliti dan menelaah berbagai literatur-literatur yang bersumber dari buku-buku, teks, jurnal ilmiah, situs-situs di internet, dan bacaan-bacaan yang terkait dengan topik pengenalan karakter, metode dalam pengolahan citra, metode ekstraksi fitur, dan metode klasifikasi SVM.

1.5.2 Pengumpulan Dataset

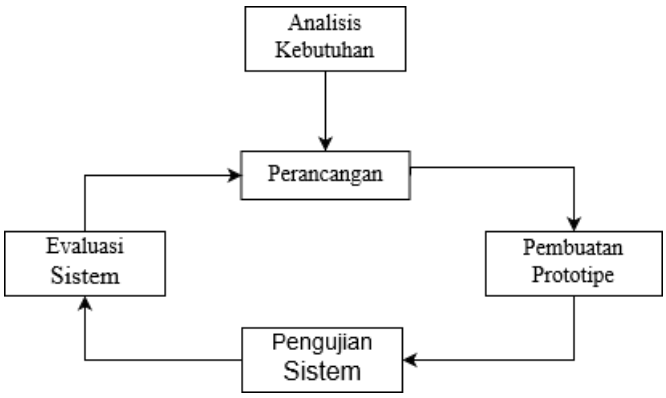
Dataset yang akan dikumpulkan terbagi menjadi dua, yaitu dataset untuk data latih dan dataset untuk data uji. Data latih yaitu berisi data citra karakter yang berjumlah sebanyak 560 gambar yang terdiri dari 80 karakter, yaitu huruf A-Z, a-z, angka 0-9, serta karakter tambahan berupa simbol .,:;/[]%*'()?!#+-= dengan ukuran 18x18 piksel. Sementara data uji adalah data hasil *scan* dokumen karya tulis ilmiah skripsi. Dokumen ini berasal dari skripsi mahasiswa Teknik Informatika Universitas Komputer Indonesia.

1.5.3 Analisis Metode

Pada tahap ini dilakukan analisis terhadap metode-metode yang akan digunakan pada penelitian ini, seperti metode untuk *preprocessing* citra, metode ekstraksi fitur zoning dan metode SVM.

1.5.4 Pembangunan Perangkat Lunak OCR

Metode pembangunan perangkat lunak yang akan digunakan pada penelitian ini adalah model prototipe [12]. Metode ini digunakan karena setiap tahap yang telah dibuat dapat dievaluasi dan dirancang kembali jika tahap yang ada tidak sesuai dengan yang diharapkan, sehingga ketika diuji didapatkan sistem yang sudah benar atau sesuai. Model pembangunan perangkat lunak prototipe bisa dilihat pada Gambar 1.2.



Gambar 1.2 Model Prototipe

Berikut adalah penjelasan dari metode prototipe.

a. Analisis kebutuhan, menganalisis segala kebutuhan yang diperlukan dalam membuat sistem OCR, seperti dataset yang diperlukan, serta kebutuhan fungsional dan non-fungsional, dan analisis pengguna.
b. Perancangan, menentukan tahapan yang akan digunakan pada pembangunan perangkat lunak, seperti perancangan struktur menu, antarmuka, perancangan pesan, dan jaringan semantik.
c. Pembuatan prototipe, mengimplementasikan perencanaan menjadi bentuk prototipe perangkat lunak untuk sistem OCR mulai dari tahap *preprocessing*, pelatihan dan pengujian.
d. Pengujian Sistem, perangkat lunak yang telah dibangun akan diuji fungsionalitasnya menggunakan *black box.*
e. Evaluasi Sistem, perangkat yang telah diuji fungsionalitasnya selanjutnya akan dievaluasi jika terdapat kekurangan atau kesalahan dalam program.

1.5.5 Pengujian Metode

Perangkat lunak yang telah dibuat dan diuji secara menyeluruh akan masuk ke tahap pengujian metode agar dapat mengetahui tingkat akurasi yang dihasilkan SVM untuk kasus pengenalan karakter pada dokumen karya tulis ilmiah. Hasil yang telah dikeluarkan oleh perangkat lunak akan dihitung akurasi ketepatan pengenalannya menggunakan metode *Classification Accuracy* sehingga bisa diketahui rata-rata tingkat akurasi pengenalan karakter yang dihasilkan.

1.5.6 Penarikan Kesimpulan

Setelah tahap pengujian selesai, selanjutnya dilanjutkan penarikan kesimpulan dari hasil penerapan metode ekstraksi fitur dan SVM untuk mengenali karakter pada citra dokumen karya tulis ilmiah.

1.6 Sistematika Penulisan

Sistematika penulisan disusun untuk memberikan gambaran secara umum mengenai bahasan penelitian pada dokumen ini. Sistematika penulisan penelitian ini adalah sebagai berikut.

BAB 1 PENDAHULUAN

Bab ini berisi uraian tentang 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 *Optical Character Recognition* (OCR) pada dokumen karya tulis ilmiah skripsi menggunakan metode *SVM*.

BAB 2 TINJAUAN PUSTAKA

Bab ini memuat berbagai konsep dasar dan teori-teori yang terkait dengan konsep OCR, pengolahan citra digital, teknik *preprocessing*, metode untuk ekstraksi fitur, pengenalan *SVM* sebagai metode klasifikasi, *Unified modeling language* (UML), bahasa pemrograman, dan perangkat lunak lain yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini berisi tentang analisis dan perancangan sistem yaitu meliputi analisis masalah, analisis data masukan, analisis proses *preprocessing*, ekstraksi fitur dan klasifikasi, analisis data keluaran, analisis kebutuhan fungsional dan *non* fungsional, perancangan antar muka, struktur menu dan pesan, serta jaringan semantik.

BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab ini berisi tentang hasil dari keseluruhan tahap analisis dan perancangan yang meliputi implementasi data masukan, implementasi perangkat keras dan perangkat lunak yang digunakan, implementasi antarmuka, serta pengujian dan hasil pengujian fungsionalitas sistem dan pengujian akurasi metode.

BAB 5 KESIMPULAN DAN SARAN

Pada bab ini akan diuraikan hasil dari penelitian yang telah dilakukan sesuai dengan tujuan yang sudah ditetapkan, disertai dengan saran untuk peneliti selanjutnya agar penelitian selanjutnya menjadi lebih baik lagi.