ABSTRAK

EKSTRAKSI DATA OTOMATIS DARI KARTU TANDA MAHASISWA (KTM) MENGGUNAKAN OPTICAL CHARACTER RECOGNITION (OCR) DAN REGULAR EXPRESSION BERBASIS GUI TKINTER

Nama : Andes Putra Gunawan

NPM : 2255201038

Pembimbing : Dr. Yulia Darmi, S.Kom., M.Kom

Penelitian ini bertujuan merancang dan mengimplementasikan sistem ekstraksi data otomatis dari citra KTM menggunakan kombinasi *Optical Character Recognition* (OCR) berbasis Tesseract dan *Regular Expression*, dengan antarmuka pengguna grafis berbasis Tkinter. Sistem yang dibangun terdiri atas tahap pra-pemrosesan citra (pemangkasan zona teks, koreksi kemiringan dengan penjagaan terhadap latar belakang bertekstur, dan binerisasi Otsu), ekstraksi teks OCR menggunakan strategi multi-sumber dan multi-mode segmentasi halaman, serta ekstraksi *field* data menggunakan pola *Regular Expression* bertingkat yang dilengkapi kamus pemetaan kata kunci untuk *field* Program Studi dan Fakultas. Pengujian dilakukan terhadap 100 citra KTM Universitas Muhammadiyah Bengkulu menggunakan metrik evaluasi *confusion matrix* berupa *Precision*, *Recall*, *F1-Score*, dan *Accuracy* untuk setiap *field* data. Hasil pengujian menunjukkan sistem mencapai akurasi keseluruhan sebesar 93,8%, dengan *Precision* 94,86%, *Recall* 98,72%, dan *F1-Score* 96,75%. Akurasi per-*field* menunjukkan NPM mencapai 100%, Fakultas 98%, Program Studi 95%, TTL 94%, dan Nama 82%. Penelitian ini juga menemukan dan mengatasi permasalahan lintas platform pada pustaka OCR yang menyebabkan hilangnya spasi dan tanda baca pada hasil ekstraksi ketika dijalankan pada sistem operasi Windows. Hasil penelitian menunjukkan bahwa kombinasi OCR dan *Regular Expression* yang dirancang khusus sesuai karakteristik visual KTM mampu menghasilkan sistem ekstraksi data yang akurat dan layak digunakan untuk mendukung proses administrasi akademik.

Kata Kunci: Ekstraksi Data, Kartu Tanda Mahasiswa (KTM), Optical Character Recognition (OCR), Regular Expression, Tesseract, Confusion Matrix, Tkinter

ABSTRACT

AUTOMATED DATA EXTRACTION FROM STUDENT ID CARDS (KTM) USING OPTICAL CHARACTER RECOGNITION (OCR) AND REGULAR EXPRESSIONS BASED ON TKINTER GUI

Name : Andes Putra Gunawan

NPM : 2255201038

Advisor : Dr. Yulia Darmi, S.Kom., M.Kom

This research aims to design and implement an automatic data extraction system from student ID card (KTM) images using a combination of Tesseract-based Optical Character Recognition (OCR) and Regular Expression, with a graphical user interface built using Tkinter. The developed system consists of an image preprocessing stage (text-zone cropping, skew correction with protection against textured backgrounds, and Otsu binarization), OCR text extraction using a multi-source and multi-page-segmentation-mode strategy, and data field extraction using tiered Regular Expression patterns supported by a keyword-mapping dictionary for the Study Program and Faculty fields. Testing was conducted on 100 student ID card images from Universitas Muhammadiyah Bengkulu using confusion matrix evaluation metrics, namely Precision, Recall, F1-Score, and Accuracy for each data field. The test results show that the system achieved an overall accuracy of 93.8%, with a Precision of 94.86%, Recall of 98.72%, and F1-Score of 96.75%. Per-field accuracy results show that the Student ID Number (NPM) field reached 100%, Faculty 98%, Study Program 95%, Place/Date of Birth 94%, and Name 82%. This research also identified and resolved a cross-platform issue in the OCR library that caused the loss of spaces and punctuation marks in the extraction results when run on the Windows operating system. The findings indicate that the combination of OCR and Regular Expression, specifically designed to match the visual characteristics of student ID cards, is able to produce an accurate and reliable data extraction system suitable for supporting academic administration processes.

Keywords: Data Extraction, Student ID Card, Optical Character Recognition (OCR), Regular Expression, Tesseract, Confusion Matrix, Tkinter.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era transformasi digital yang terus berkembang pesat, institusi pendidikan tinggi dituntut untuk mampu mengelola data mahasiswa secara lebih efisien dan akurat. Salah satu dokumen identitas penting yang diterbitkan oleh setiap perguruan tinggi adalah Kartu Tanda Mahasiswa (KTM), yang memuat informasi vital seperti nama lengkap, Nomor Induk Mahasiswa (NIM), program studi, dan fakultas (Reswan *et al.*, 2024) . Proses pencatatan data dari KTM masih banyak dilakukan secara manual oleh petugas administrasi, yakni dengan menyalin informasi yang tercetak di kartu tersebut ke dalam sistem komputer satu per satu. Cara kerja seperti ini tentu memakan banyak waktu dan berpotensi menimbulkan kesalahan entri data yang tidak disadari, terutama ketika volume data yang harus diproses cukup besar. Proses pembuatan dan pengelolaan KTM yang masih dilakukan secara manual mengharuskan data mahasiswa diinput dan diverifikasi satu per satu, sehingga rawan terjadi kesalahan dan memerlukan waktu yang cukup lama. Kondisi ini menjadi salah satu penghambat utama dalam upaya meningkatkan kualitas layanan administrasi akademik di berbagai kampus, khususnya yang masih mengandalkan proses konvensional .

Masalah serupa juga ditemukan dalam konteks yang lebih luas, yaitu proses penginputan data dari berbagai dokumen identitas lainnya. Proses input data dari KTM yang dilakukan secara manual seringkali tidak efisien dan rentan terhadap kesalahan, sehingga otomatisasi proses ini menjadi solusi penting untuk

meningkatkan akurasi dan kecepatan layanan administrasi di institusi pendidikan (Rahmadani, Rozikin and Karawang, 2021). Fakta ini mengindikasikan bahwa persoalan bukan sekadar tentang kecepatan kerja manusia, melainkan juga menyangkut keandalan dan konsistensi data yang dihasilkan. Kesalahan kecil dalam pencatatan NIM atau nama mahasiswa, misalnya, dapat berdampak signifikan pada proses berikutnya seperti verifikasi kehadiran, pencetakan sertifikat, ijazah, hingga pengajuan beasiswa (Firin and Sriani, 2025).

Permasalahan inti dari penelitian ini adalah bagaimana merancang sebuah sistem yang mampu mengotomatiskan proses pembacaan dan ekstraksi data dari gambar KTM secara akurat, cepat, dan mudah digunakan. Secara khusus, persoalan teknis yang muncul meliputi bagaimana memastikan mesin dapat mengenali karakter pada citra KTM dengan presisi tinggi, bagaimana memisahkan setiap field data secara tepat dari teks mentah hasil OCR, serta bagaimana menyajikan antarmuka pengguna yang intuitif agar dapat dioperasikan oleh staf non-teknis sekalipun.

Sejumlah penelitian terdahulu telah meletakkan fondasi yang kuat bagi topik ini. Teknologi *Optical Character Recognition* (OCR) merupakan metode yang memanfaatkan algoritma pengenalan karakter untuk mengekstrak informasi dari gambar atau teks cetak, dan telah terbukti efektif dalam membaca serta mengenali karakter pada KTM secara efisien. Dalam implementasinya pada dokumen identitas, penelitian (Octaviani, Setiawan and Kelana, 2023) menunjukkan bahwa metode OCR konvensional mampu meningkatkan efisiensi proses input data e-KTP dengan akurasi sebesar 79%, meskipun performanya

masih sangat dipengaruhi oleh kualitas citra dan pencahayaan sehingga diperlukan metode pendukung untuk meningkatkan akurasi pengenalan karakter. Di sisi lain, penelitian lain yang membandingkan dua pendekatan berbeda menemukan bahwa pytesseract memberikan akurasi lebih tinggi yakni 98,33% dibandingkan template matching yang hanya mencapai 67,33%, meskipun template matching lebih stabil dalam mengenali pola karakter tertentu ketika citra berada pada kondisi kurang ideal (Bachiar Rizki, Pratomo and Wiwit Agus, 2026).

Meskipun sejumlah penelitian sebelumnya telah berhasil membuktikan efektivitas OCR untuk ekstraksi data dari dokumen identitas, terdapat beberapa celah yang belum tertangani secara menyeluruh. Pertama, sebagian besar penelitian yang ada fokus pada dokumen seperti e-KTP atau plat nomor kendaraan, bukan pada KTM yang memiliki struktur tata letak teks tersendiri dan bervariasi antar institusi (Ristanto *et al.*, 2023). Meskipun beberapa sistem telah mengintegrasikan GUI untuk memudahkan pengguna, sistem yang secara khusus menggabungkan pra-pemrosesan citra, OCR, dan Regular Expression ke dalam satu aplikasi desktop berbasis GUI yang lengkap untuk objek KTM belum banyak dikembangkan. Kedua, penelitian yang memprioritaskan kemudahan akses bagi pengguna non-teknis dengan antarmuka Python Tkinter untuk kebutuhan administrasi akademik juga masih terbatas (Novendra Sulu *et al.*, 2024).

Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem ekstraksi data otomatis dari citra KTM menggunakan kombinasi teknologi OCR dan *Regular Expression* dengan desain antarmuka grafis berbasis Tkinter. Secara lebih rinci, tujuan yang ingin dicapai yaitu membangun alur pemrosesan

citra yang mampu mempersiapkan gambar KTM sebelum dikenali oleh mesin OCR, mengimplementasikan Tesseract OCR untuk mengubah citra menjadi teks, merancang pola *Regular Expression* yang tepat untuk mengekstrak field data spesifik seperti nama, NIM, dan program studi, serta menyajikan seluruh fungsionalitas tersebut dalam aplikasi desktop berbasis GUI Tkinter yang dapat dioperasikan dengan mudah (Julian Mulyadi and Nuri David Maria Veronika, 2025).

Pendekatan yang digunakan dalam penelitian ini memadukan beberapa teknologi secara berurutan. Kebaruan utama penelitian ini terletak pada integrasi menyeluruh antara tiga komponen teknis pra-pemrosesan citra, OCR berbasis Tesseract, dan *parsing Regular Expression* ke dalam satu aplikasi desktop yang dapat langsung digunakan tanpa keahlian pemrograman (Setiawan, Guntara and Purwaamijaya, 2025). Penggunaan pola Regular Expression yang dirancang khusus untuk variasi format teks KTM (Nanda Aprillia *et al.*, 2025). Sementara implementasi Tesseract OCR untuk pengenalan karakter pada objek dokumen terstruktur telah menunjukkan tingkat akurasi rata-rata yang tinggi, penelitian ini mengembangkan lebih lanjut dengan menambahkan lapisan parsing berbasis pola dan antarmuka GUI yang belum dikombinasikan secara terpadu sebelumnya (Al Fajar, Darmayunata and Hardianto, 2025).

Hasil penelitian ini diharapkan membawa manfaat yang konkret dari dua sisi. Dari sisi praktis, aplikasi yang dihasilkan dapat langsung digunakan oleh petugas administrasi kampus untuk mempercepat proses pencatatan data mahasiswa, mengurangi risiko kesalahan manusiawi, dan menghemat waktu kerja

secara signifikan. Dari sisi akademis, penelitian ini berkontribusi pada pengembangan literatur lokal mengenai penerapan OCR dan Regular Expression untuk dokumen identitas berbahasa Indonesia, serta menjadi rujukan bagi peneliti lain yang ingin mengembangkan sistem serupa dengan cakupan yang lebih luas, misalnya untuk jenis dokumen lain atau dengan mekanisme ekspor data ke berbagai format file.

1.2 Pertanyaan Penelitian

1. Bagaimana merancang dan membangun sistem ekstraksi data otomatis dari citra KTM menggunakan Tesseract OCR dan Regular Expression berbasis GUI Tkinter?
2. Bagaimana efektivitas tahap prapemrosesan citra (grayscale, Gaussian blur, binarisasi Otsu) dalam meningkatkan kualitas hasil ekstraksi teks oleh Tesseract OCR?
3. Seberapa tinggi tingkat akurasi sistem dalam mengekstraksi field NIM, Nama, Program Studi, Fakultas, dan TTL dari citra KTM berdasarkan metrik precision, recall, F1-score, dan akurasi per karakter?

1.3 Tujuan Penelitian

1. Merancang dan membangun sistem ekstraksi data otomatis dari citra Kartu Tanda Mahasiswa (KTM) menggunakan Tesseract OCR dan Regular Expression yang disajikan dalam antarmuka GUI berbasis Tkinter.
2. Menganalisis efektivitas tahap prapemrosesan citra yang meliputi konversi grayscale, reduksi noise menggunakan Gaussian blur, dan binarisasi Otsu dalam meningkatkan kualitas hasil ekstraksi teks oleh Tesseract OCR.

3. Mengukur tingkat akurasi sistem dalam mengekstraksi field NIM, Nama, Program Studi, Fakultas, dan Tempat Tanggal Lahir (TTL) dari citra KTM menggunakan metrik precision, recall, F1-score, dan akurasi per karakter.

1.4 Kerangka Kerja Penelitian (*Research Framework*)

Framework penelitian merupakan kerangka konseptual yang menggambarkan secara sistematis hubungan antara komponen-komponen utama dalam suatu penelitian. Berikut framework penelitian secara lengkap.

Tabel 1.1 Framework Penelitian

No	Tahap Penelitian	Kegiatan Utama	Output yang Dihasilkan
1	Studi Literatur	Mengkaji teori dan penelitian terdahulu terkait OCR (Tesseract), Regular Expression, pengolahan citra digital, GUI Tkinter, serta ekstraksi data dari dokumen identitas KTM	Landasan teori dan kerangka konseptual penelitian
2	Pengumpulan Data	Mengumpulkan citra KTM dari berbagai mahasiswa dengan variasi format, kualitas cetak, dan kondisi pencahayaan yang berbeda	Dataset citra KTM (minimal 50-100 sampel) dalam format JPG/PNG
3	Pra-pemrosesan Citra	Melakukan konversi grayscale, reduksi noise (Gaussian blur), binarisasi (Otsu thresholding), dan cropping area teks	Citra KTM hasil pra-pemrosesan yang siap diolah oleh OCR
4	Ekstraksi Teks dengan OCR	Mengimplementasikan Tesseract OCR Engine untuk mengenali dan mengkonversi teks pada citra KTM menjadi string teks mentah (raw text)	Raw text hasil OCR yang berisi seluruh teks dari KTM
5	Pencarian Pola dengan Regex	Merancang dan menerapkan pola Regular Expression untuk menyaring dan memisahkan field data seperti NIM, Nama, Program Studi, Fakultas, dan TTL	Field data terpisah (NIM, Nama, Prodi, Fakultas, TTL) dalam bentuk string terstruktur
6	Pengembangan GUI Tkinter	Membangun antarmuka pengguna grafis menggunakan library Tkinter untuk	Aplikasi desktop berbasis GUI yang

No	Tahap Penelitian	Kegiatan Utama	Output yang Dihasilkan
		mengunggah citra, menampilkan hasil ekstraksi, mengedit data, dan menyimpan hasil	user-friendly
7	Pengujian dan Evaluasi	Menguji sistem dengan data uji (20-30 citra KTM) pada berbagai skenario (kualitas baik, noise, pencahayaan kurang, variasi format) dan membandingkan hasil ekstraksi dengan ground truth	Nilai akurasi per karakter, akurasi per field, dan waktu pemrosesan rata-rata
8	Analisis Hasil	Menganalisis faktor-faktor yang mempengaruhi keberhasilan dan kegagalan ekstraksi, seperti kualitas citra, font, tata letak KTM, dan kesalahan OCR	Kesimpulan mengenai kelebihan dan keterbatasan sistem, serta saran pengembangan lanjutan
9	Penyusunan Laporan	Menulis dokumentasi penelitian secara sistematis mulai dari pendahuluan, tinjauan pustaka, metode, hasil, hingga kesimpulan dan saran	Laporan skripsi lengkap

Framework penelitian ini disusun secara sistematis untuk membangun sistem ekstraksi data otomatis dari Kartu Tanda Mahasiswa (KTM) menggunakan OCR dan Regular Expression berbasis GUI Tkinter. Penelitian dimulai dengan studi literatur untuk mengkaji teori dan penelitian terdahulu tentang OCR, Regex, pengolahan citra, dan pemrograman GUI sebagai landasan akademik. Tahap berikutnya adalah pengumpulan data berupa citra KTM dari berbagai mahasiswa dengan variasi kualitas cetak, format, dan kondisi pencahayaan untuk memastikan dataset yang representatif. Setelah data terkumpul, dilakukan pra-pemrosesan citra yang mencakup konversi grayscale, reduksi noise menggunakan Gaussian blur, binarisasi dengan Otsu thresholding, dan cropping area teks untuk

meningkatkan kualitas citra sebelum dikenali oleh OCR. Selanjutnya, ekstraksi teks dengan Tesseract OCR dilakukan untuk mengubah citra menjadi teks mentah (raw text). Teks tersebut kemudian diproses pada tahap pencarian pola dengan Regular Expression untuk menyaring dan memisahkan field-field penting seperti NIM, Nama, Program Studi, Fakultas, dan Tempat Tanggal Lahir berdasarkan pola yang telah dirancang. Hasil ekstraksi kemudian ditampilkan melalui antarmuka GUI Tkinter yang memungkinkan pengguna mengunggah citra, melihat hasil, melakukan edit manual jika diperlukan, serta menyimpan data ke file TXT atau CSV. Tahap pengujian dan evaluasi dilakukan dengan membandingkan hasil ekstraksi terhadap ground truth pada berbagai skenario, seperti KTM berkualitas baik, KTM dengan noise, pencahayaan kurang, serta variasi format dari universitas berbeda. Metrik evaluasi yang digunakan meliputi akurasi per karakter, akurasi per field, dan waktu pemrosesan rata-rata. Setelah itu, dilakukan analisis hasil untuk mengidentifikasi faktor-faktor yang mempengaruhi keberhasilan atau kegagalan ekstraksi, seperti kualitas citra, jenis font, atau tata letak KTM. Seluruh proses ini didokumentasikan dalam penyusunan laporan akhir sebagai luaran penelitian. Dengan framework yang terstruktur ini, penelitian diharapkan mampu menghasilkan aplikasi desktop yang dapat membantu petugas administrasi dalam menginput data KTM secara otomatis, cepat, dan akurat, serta memberikan kontribusi ilmiah dalam pengembangan sistem ekstraksi informasi berbasis OCR dan Regex.