JKTi Jurnal Keilmuan Teknologi informasi

Volume 01, No. 1, Juli 2025, Hal. 11 – 14
e-ISSN : XXXX–XXXX

Implementasi OCR berbasis Tesseract untuk Ekstraksi data kartu mahasiswa UMKLA

Muhammad Nashiruddin[1], Fiusyam Dhaza Noor Praditya[2], Agiel Faiz Mufazzal[3], Ardiansyah[4]

[1] Program Studi Teknologi Informasi, Fakultas Kesehatan dan Teknologi, Universitas Muhammadiyah Klaten, Klaten

Email: [1]Anasruddin050@gmail.com, [2]Fiusyamdnp@gmail.com, [3]Mufazzalagiel@gmail.com, [4]Ardiansyah@umkla.ac.id

ABSTRACT — Manual data entry from student ID cards (KTM) is often inefficient and prone to errors. Therefore, automating this process is a crucial solution for educational institutions to improve accuracy and the speed of administrative services. This research aims to design and implement an Optical Character Recognition (OCR) system to automatically extract information from student ID card images of Universitas Muhammadiyah Klaten (UMKLA). The methodology involves image pre-processing using the OpenCV library to enhance image quality through grayscale conversion and Otsu's binarization. Subsequently, the Tesseract OCR Engine is used to convert the image into raw text, which is then parsed using Regular Expressions (Regex) to separate data fields such as Name, Student ID Number (NIM), and Program of Study. Test results indicate that the system can extract information with a good success rate, although accuracy is heavily influenced by image quality factors like lighting and text clarity. Fields with standard printed formats were found to have higher accuracy. In conclusion, this Tesseract-based system successfully demonstrates its feasibility for local automation of student ID card data. However, further development in the post-processing stage is required to handle more complex OCR output variations.

KEYWORDS — OCR; Student ID Card; Tesseract; Data Extraction;

INTISARI — Proses input data dari Kartu Tanda Mahasiswa (KTM) yang dilakukan secara manual seringkali tidak efisien dan rentan terhadap kesalahan. Oleh karena itu, otomatisasi proses ini menjadi solusi penting untuk meningkatkan akurasi dan kecepatan layanan administrasi di institusi pendidikan. Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem *Optical Character Recognition* (OCR) untuk mengekstraksi informasi dari gambar KTM Universitas Muhammadiyah Klaten (UMKLA) secara otomatis. Metodologi yang digunakan meliputi pra-pemrosesan citra menggunakan pustaka OpenCV untuk meningkatkan kualitas gambar melalui konversi *grayscale* dan *binarisasi Otsu*. Selanjutnya, *Tesseract* OCR *Engine* digunakan untuk mengubah citra menjadi teks mentah, yang kemudian diurai (*parsing*) menggunakan *Regular Expressions* (Regex) untuk memisahkan *field* data seperti Nama, NIM, dan Program Studi. Hasil pengujian menunjukkan sistem mampu mengekstrak informasi dengan tingkat keberhasilan yang baik, meskipun akurasi sangat dipengaruhi oleh kualitas gambar, seperti pencahayaan dan kejelasan teks. *Field* dengan format cetak standar terbukti memiliki akurasi lebih tinggi. Kesimpulannya, sistem berbasis Tesseract ini berhasil membuktikan kelayakannya untuk otomatisasi data KTM secara lokal, namun diperlukan pengembangan lebih lanjut pada tahap pasca-pemrosesan untuk menangani variasi hasil OCR yang lebih kompleks.

KATA KUNCI — OCR; Kartu Mahasiswa; Tesseract; Ekstraksi Data;

I. PENDAHULUAN

Pemanfaatan teknologi *Optical Character Recognition (OCR)* telah menjadi solusi esensial dalam digitalisasi dokumen dan otomatisasi proses bisnis di berbagai sektor. Di era digital saat ini, kebutuhan akan efisiensi dan akurasi dalam pengelolaan informasi semakin meningkat, mendorong adopsi teknologi yang mampu mengubah data analog menjadi digital secara otomatis.

Dalam konteks institusi pendidikan, pengelolaan data mahasiswa, termasuk identifikasi dan verifikasi melalui kartu mahasiswa, seringkali masih melibatkan proses manual yang rentan terhadap kesalahan, membutuhkan waktu yang lama, dan memakan sumber daya. Kartu mahasiswa memuat informasi krusial seperti Nama, Nomor Induk Mahasiswa (NIM), Program Studi, dan Fakultas. Dengan meningkatnya jumlah mahasiswa dan tuntutan akan efisiensi operasional, pendekatan otomatis untuk mengekstrak data ini menjadi sangat relevan.

Penelitian ini bertujuan untuk mengimplementasikan dan mengevaluasi sistem OCR menggunakan **Tesseract OCR Engine** untuk secara otomatis mengekstrak informasi penting dari gambar kartu mahasiswa Universitas Muhammadiyah Klaten (UMKLA). Berbeda dengan pendekatan berbasis *cloud*, proyek ini berfokus pada solusi lokal yang dapat dijalankan melalui *command prompt* atau skrip Python. Diharapkan sistem ini dapat mengurangi beban kerja manual, mempercepat proses *data entry*, dan meningkatkan akurasi data mahasiswa, meskipun dengan ketergantungan pada kualitas gambar masukan

II. TINJAUAN PUSTAKA

Optical Character Recognition (OCR) adalah teknologi yang mengubah berbagai jenis dokumen, seperti gambar yang dipindai atau dokumen cetak, menjadi data teks yang dapat dibaca oleh mesin. Secara umum, sistem OCR bekerja melalui beberapa tahapan inti, yaitu akuisisi citra, pra-pemrosesan (*pre-processing*), segmentasi, ekstraksi fitur, pengenalan, dan pasca-

pemrosesan (*post-processing*) [15]. Pra-pemrosesan menjadi langkah krusial karena kualitas gambar input sangat memengaruhi akurasi hasil akhir [19].

Tesseract OCR Engine, yang dikembangkan oleh Hewlett-Packard dan kini dikelola oleh Google, merupakan salah satu mesin OCR *open-source* yang paling populer [17]. Sejak versi 4.0, Tesseract telah mengadopsi arsitektur jaringan saraf tiruan *Long Short-Term Memory* (LSTM), yang secara signifikan meningkatkan kemampuannya dalam mengenali baris teks secara kontekstual, tidak hanya per karakter [4], [15]. Kemampuan ini, ditambah dengan dukungannya terhadap lebih dari 100 bahasa, membuat Tesseract menjadi pilihan yang kuat untuk berbagai aplikasi, mulai dari deteksi plat nomor kendaraan [16] hingga ekstraksi data pada dokumen seperti KTP [5] dan kartu vaksin [3].

Untuk mencapai akurasi yang tinggi, pra-pemrosesan citra memegang peranan vital. Penelitian oleh Rozi, dkk. [19] menunjukkan bahwa teknik seperti konversi *grayscale*, binarisasi, dan *noise reduction* dapat meningkatkan kinerja OCR secara signifikan pada gambar berkualitas rendah. Studi lain juga membandingkan kinerja Tesseract dengan metode lain seperti *Template Matching* dan EasyOCR. Octaviani, dkk. [9] menemukan bahwa Pytesseract (implementasi Tesseract di Python) memiliki akurasi (98,33%) dan kecepatan yang jauh lebih unggul dibandingkan *Template Matching* (67,33%) untuk ekstraksi data KTP. Sementara itu, Darpito, dkk. [8] menyimpulkan bahwa EasyOCR cenderung lebih akurat dalam mengenali kata (*Word Error Rate* lebih rendah) pada dokumen kompleks, meskipun Tesseract unggul dalam kecepatan pemrosesan.

Setelah teks mentah diekstraksi oleh OCR, diperlukan tahap pasca-pemrosesan untuk memilah informasi spesifik. Metode yang umum digunakan adalah *Regular Expressions* (Regex) untuk menemukan pola teks tertentu. Namun, metode ini memiliki keterbatasan jika hasil OCR tidak sempurna. Sebagai alternatif, penelitian telah mengeksplorasi metode yang lebih cerdas seperti pencocokan kata menggunakan *Hamming Distance* untuk koreksi teks [14] atau *Named Entity Recognition* (NER) untuk mengidentifikasi dan mengklasifikasikan entitas seperti nama orang, lokasi, dan nomor identitas secara kontekstual [6]. Evaluasi kinerja sistem OCR sendiri idealnya menggunakan metrik standar seperti *Character Error Rate* (CER) dan *Word Error Rate* (WER) untuk memberikan ukuran akurasi yang objektif [20].

III. METODOLOGI

Metodologi penelitian ini mencakup langkah-langkah dalam pengembangan sistem OCR untuk kartu mahasiswa UMKLA menggunakan Tesseract OCR Engine dan Python.

A. PERSIAPAN ENVIRONMENT

Lingkungan pengembangan disiapkan menggunakan Python. Pustaka-pustaka yang diinstal meliputi opencv-python (untuk cv2), pytesseract (sebagai wrapper untuk Tesseract), Pillow, dan re (untuk Regular Expressions). Tesseract OCR Engine versi 5.3.0 diinstal secara lokal pada sistem, dan path ke executable Tesseract diatur dalam skrip Python untuk memastikan pytesseract dapat berfungsi dengan benar.

B. PENGUMPULAN DAN PRA-PEMROSESAN DATA

Dataset terdiri dari 2 gambar kartu mahasiswa UMKLA yang diambil menggunakan kamera ponsel dengan format JPG. Untuk meningkatkan akurasi OCR, setiap gambar melalui tahap pra-pemrosesan menggunakan pustaka OpenCV. Langkah-langkah pra-pemrosesan meliputi:

- **Konversi ke Grayscale**: Gambar berwarna diubah menjadi skala abu-abu (cv2.cvtColor()) untuk menyederhanakan analisis kontras teks.
- **Binarisasi (Thresholding)**: Gambar *grayscale* diubah menjadi hitam-putih murni menggunakan metode Otsu (cv2.threshold() dengan flag cv2.THRESH_OTSU). Langkah ini efektif memisahkan teks dari latar belakang.

C. IMPLEMENTASI OCR DAN TESSERACT

Gambar yang telah dipra-proses kemudian dimasukkan ke Tesseract OCR Engine melalui pytesseract. Konfigurasi spesifik yang digunakan adalah --psm 6 (Page Segmentation Mode) yang mengasumsikan gambar sebagai satu blok teks seragam. Bahasa yang digunakan adalah bahasa Indonesia (ind). Fungsi pytesseract.image_to_string() digunakan untuk mengekstrak seluruh teks dari gambar.

D. EKSTRAKSI INFORMASI SPESIFIK DENGAN REGEX

Setelah teks lengkap berhasil diekstrak, informasi spesifik seperti Nama, NIM, Tempat Tanggal Lahir, dan Program Studi dipisahkan menggunakan *Regular Expressions* (Regex) yang diimplementasikan dengan modul re di Python. Pola-pola Regex didesain untuk mencari kata kunci (misalnya, "NAMA :", "NIM :") dan menangkap teks yang mengikutinya.

E. EVALUASI KINERJA

Akurasi sistem diukur dengan membandingkan hasil ekstraksi OCR dengan data asli (*ground truth*) pada kartu. Metrik yang digunakan adalah akurasi per *field* yang dihitung berdasarkan jumlah karakter yang dikenali dengan benar dibagi dengan total karakter asli

IV. HASIL DAN PEMBAHASAN

A. REKOMENDASI LAINNYA

Pengujian dilakukan pada sebuah gambar Kartu Tanda Mahasiswa (KTM) UMKLA seperti yang terlihat pada Gambar 1, dengan kondisi pencahayaan dan resolusi yang baik. Sistem berhasil mengekstrak informasi dari gambar tersebut seperti pada Gambar 2, dan hasilnya dibandingkan dengan data asli (*ground truth*) untuk mengukur akurasi. Gambar 2 menampilkan hasil ekstrasi yang dilakukan dalam penelitian yang menggunakan Tesseract untuk ektraksi teks yang terdapat pada kartu mahasiswa. Berikut hasil ektrasi teks pada kartu mahasiswa.

Gambar 1. Gambar Kartu Mahasiswa

Gambar 2. Gambar Hasil Deteksi Teks pada Kartu Mahasiswa

B. PEMBAHASAN

Berdasarkan hasil pengujian pada Tabel I, sistem menunjukkan performa yang sangat tinggi dengan **rata-rata akurasi di atas 95%**. Sebagian besar *field* seperti **Nama**, **NIM**, **Program Studi**, dan **Fakultas** berhasil diekstraksi dengan akurasi sempurna (100%). Hal ini menunjukkan bahwa kombinasi pra-pemrosesan citra (grayscaling dan binarisasi) dan Tesseract OCR sangat efektif untuk mengenali teks dengan format cetak yang jelas dan standar.

TABEL I
EVALUASI HASIL DETEKSI TEKS PADA KARTU MAHASISWA

Field	Teks Asli	Hasil OCR	Akurasi
Nama	AGIEL FAIZ MUFAZZAL	AGIEL FAIZ MUFAZZAL	100%
Nim	202307001	202307001	100%
TTL	*******	*********	96%
Program Studi	S1 Teknologi Informasi	S1 Teknologi Informasi	100%
Fakultas	FAKULTAS KESEHATAN DAN TEKNOLOGI	FAKULTAS KESEHATAN DAN TEKNOLOGI	100%

Satu-satunya kesalahan terdeteksi pada *field* **TTL (Tempat, Tanggal Lahir)**, di mana angka tahun "**2004**" salah dikenali sebagai "**3004**". Ini adalah contoh klasik dari **kesalahan substitusi karakter** dalam OCR, di mana bentuk angka '2' yang sedikit tidak sempurna pada gambar mungkin dikenali sebagai angka '3' oleh Tesseract. Meskipun kesalahan ini kecil, hal ini menyoroti kelemahan OCR yang sensitif terhadap variasi kecil pada bentuk karakter, bahkan dalam gambar yang berkualitas baik.

Temuan ini mengkonfirmasi bahwa meskipun Tesseract sangat andal, ia tidak sepenuhnya sempurna dan masih memerlukan mekanisme validasi atau koreksi lebih lanjut, terutama untuk data numerik yang krusial. Penggunaan algoritma pasca-pemrosesan yang lebih cerdas, seperti yang disarankan oleh Brillian & Agustin [14], dapat menjadi solusi untuk memverifikasi dan memperbaiki kesalahan semacam ini di masa depan.

V. KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan sistem untuk ekstraksi data otomatis dari Kartu Tanda Mahasiswa (KTM) Universitas Muhammadiyah Klaten menggunakan Tesseract OCR Engine dengan dukungan pustaka OpenCV dan Python. Sistem yang dikembangkan mampu melakukan pra-pemrosesan citra dan mengekstrak informasi tekstual dengan tingkat akurasi yang sangat tinggi, di mana sebagian besar *field* data berhasil dikenali dengan sempurna.

Hasil pengujian menunjukkan bahwa metode yang diusulkan sangat efektif untuk data dengan format teks yang jelas dan standar. Namun, ditemukan adanya kesalahan substitusi karakter pada data numerik, yang menggarisbawahi bahwa performa Tesseract masih sensitif terhadap variasi kecil dalam bentuk karakter, bahkan pada citra berkualitas baik.

Dapat disimpulkan bahwa implementasi OCR berbasis Tesseract merupakan solusi yang **layak dan efisien** untuk otomatisasi entri data KTM, yang berpotensi mengurangi waktu kerja manual dan human error. Untuk pengembangan di masa depan, disarankan untuk mengimplementasikan **metode pasca-pemrosesan yang lebih canggih**, seperti algoritma koreksi teks atau validasi berbasis aturan, untuk menangani dan memperbaiki kesalahan pengenalan karakter secara otomatis, sehingga dapat meningkatkan keandalan sistem secara keseluruhan

REFERENSI

[1] C. Padole, U. S. Verma, P. Gujral, M. Kumar, I. Bajpai, and D. Mitra, "Information Extraction from Visiting Cards Using OCR and Post-Processing in Python," International Journal of Scientific and Technical Research in Engineering (IJSTRE), vol. 7, no. 5, hlm. 1-7, Sep-Okt 2022.

[2] G. Sugiarta, D. P. Andini, and S. Hidayatullah, "Ekstraksi Informasi/Data e-KTP Menggunakan Optical Character Recognition Convolutional Neural Network," JTERA (Jurnal Teknologi Rekayasa), vol. 6, no. 1, hlm. 1-6, Jun. 2021.

[3] Wahyuddin and A. Hasim, "APLIKASI EKSTRAKSI DATA KARTU VAKSIN BERBASIS WEB MENGGUNAKAN METODE OCR," JURNAL SINTAKS LOGIKA, vol. 3, no. 2, hlm. 52-57, Mei 2023.

[4] O. O. Patience, E. M. Amaechi, O. George, and O. N. Isaac, "Enhanced Text Recognition in Images Using Tesseract OCR within the Laravel Framework," Asian Journal of Research in Computer Science, vol. 17, no. 9, hlm. 58-69, 2024.

[5] M. Haris, M. G. Suryanata, and M. Yetri, "Implementasi OCR Menggunakan Algoritma Template Matching Correlation Pada Pengarsipan e-KTP," Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD, vol. 6, no. 2, hlm. 281-289, Jul. 2023.

[6] S. Fritz, V. Srikanthan, R. Arbai, C. Sun, J. Ovtcharova, and H. Wicaksono, "Automatic Information Extraction from Text-Based Requirements," International Journal of Knowledge Engineering, vol. 7, no. 1, hlm. 8-13, Jun. 2021.

[7] Y. Li, "Synergizing Optical Character Recognition: A Comparative Analysis and Integration of Tesseract, Keras, Paddle, and Azure OCR," M.S. thesis, School of Computer Science, Univ. of Sydney, Sydney, NSW, 2024.

[8] M. N. Darpito, K. Firdausy, and A. Fadlil, "Perbandingan Unjuk Kerja Library Optical Character Recognition (OCR) dalam Pengenalan Teks pada Dokumen Digital," JIP (Jurnal Informatika Polinema), vol. 11, no. 3, hlm. 273-281, Mei 2025. [9] T. Octaviani, H. Setiawan, and O. H. Kelana, "PERBANDINGAN PYTESSERACT DAN TEMPLATE MATCHING UNTUK OTOMATISASI INPUT DATA KTP," Jurnal Buana Informatika, vol. 14, no. 2, hlm. 147-156, Nov. 2023.

[9] R. S. Bahri and I. Maliki, "PERBANDINGAN ALGORITMA TEMPLATE MATCHING DAN FEATURE EXTRACTION PADA OPTICAL CHARACTER RECOGNITION," Jurnal Komputer dan Informatika (KOMPUTA), vol. 1, no. 1, hlm. 29-35, Mar. 2012.

[10] Y. Reswan, R. Raffles, A. Wijaya, and Y. Apridiansyah, "PENERAPAN ALGORITMA OCR UNTUK EKSTRAKSI INFORMASI DARI CITRA KARTU TANDA MAHASISWA (KTM)," JATI (Jurnal Mahasiswa Teknik Informatika), vol. 8, no. 5, Okt. 2024.

[11] R. Surya, "Peningkatan Akurasi OCR dalam Pemrosesan Formulir Keuangan melalui Fine-Tuning Transformer dan Strategi Pra-pemrosesan Data," Jurnal Inovasi Informatika (JII), vol. 7, no. 2, hlm. 1-12, Apr. 2025.

[12] Y. Sun, X. Mao, S. Hong, W. Xu, and G. Gui, "Template matching-based method for intelligent invoice information identification," IEEE Access, vol. 7, hlm. 28392-28401, 2019. [14] R. R. Brillian and S. Agustin,

"Pencocokan Kata dalam Optical Character Recognition Menggunakan Metode Hamming Distance," Universitas Muhammadiyah Gresik, 2023.

[13] A. M. Syahputri, B. Harijanto, and C. Rahmad, "IMPLEMENTASI OPTICAL CHARACTER RECOGNITION (OCR) UNTUK MENINGKATKAN AKURASI DAN KECEPATAN INPUT DATA DI POSYANDU," JIP (Jurnal Informatika Polinema), vol. 11, no. 1, hlm. 45-50, Nov. 2024.

[14] A. Meirza and N. R. Puteri, "Implementasi Metode YOLOV5 dan Tesseract OCR untuk Deteksi Plat Nomor Kendaraan," Jurnal Ilmu Komputer dan Desain Komunikasi Visual, vol. 9, no. 1, hlm. 424-435, Jul. 2024.

[15] A. K. Siliwangi and Y. D. Prabowo, "Pencarian Informasi Berbasis Teks dalam Komik Digital Menggunakan OCR," KALBISIANA: Jurnal Mahasiswa Institut Teknologi dan Bisnis Kalbis, vol. 8, no. 2, hlm. 1886-1894, Mei 2022.

[16] S. M. Angela and A. Eviyanti, "Development Of Optical Character Recogition Technology In Flutter For Text Detection In Images," Universitas Muhammadiyah Sidoarjo, 2024.

[17] A. Rozi, et al., "Improving OCR Performance on Low-Quality Image Using Pre-processing and Post-processing Methods," International Journal of Engineering and Technology, vol. 71, no. 6, 2023.

[18] S. Kundu, et al., "A Novel Pipeline for Improving Optical Character Recognition through Post-processing Using Natural Language Processing," arXiv preprint, 2023.