

TABLE RECOGNITION UNTUK OPTICAL CHARACTER RECOGNITION (OCR)

Lola

Program Studi Teknik Informatika, FTI, Institut Teknologi Budi Utomo Jakarta,
lola.rezak@gmail.com

Abstrak

Table Recognition merupakan salah satu aspek penting dalam *Optical Character Recognition* (OCR) yang memungkinkan untuk pengenalan karakter dengan tetap mempertahankan struktur tabel dari gambar. Seringkali, mempertahankan struktur teks sangatlah penting untuk mendapatkan informasi yang lengkap. Penelitian ini menyoroti tantangan-tantangan utama yang terkait dengan variasi format tabel dan struktur dokumen yang beragam. Berbagai pendekatan seperti *deep learning* dan pengolahan citra (*image processing*) sudah banyak digunakan pada penelitian ini. Pada makalah ini akan digunakan pendekatan *image processing* untuk melakukan *table recognition*.

Kata kunci : *component; table recognition, table extraction, optical character recognition, OCR.*

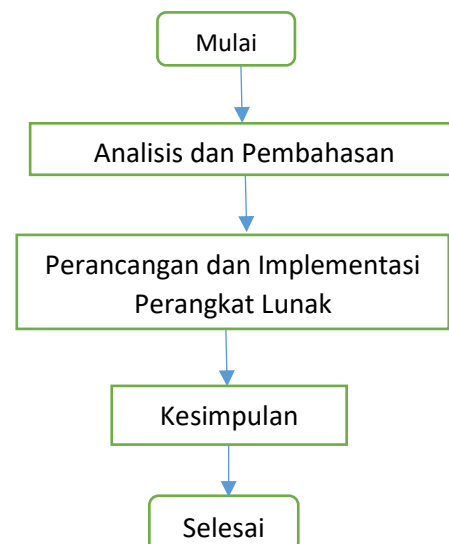
1. PENDAHULUAN

Table recognition adalah proses mengekstraksi struktur dan konten tabel dari gambar atau dokumen. Proses ini melibatkan tugas-tugas seperti *table detection*, *table structure recognition*, dan *table content recognition*. *Table detection* akan menemukan tabel pada gambar. *Table structure recognition* akan mengidentifikasi baris, kolom, dan posisi dari sel. *Table content recognition* mengidentifikasi dan mengekstrak konten teks dari tabel (Rashid, 2022). Pengenalan tabel biasanya menggunakan model *deep learning* atau pengolahan citra sederhana untuk mendeteksi tabel (*table detection*). *Table recognition* telah digunakan di beberapa bidang seperti *document analysis*, *data extraction*, dan *information organization*. *Table detection* dapat dilakukan dengan beberapa pendekatan, seperti metode *deep learning*, maupun metode *image processing*. Meskipun pendekatan *deep learning* telah mendapatkan perhatian yang signifikan dalam beberapa tahun terakhir, metode *image processing* masih relevan dan memiliki keunggulan tersendiri. Teknik pengolahan citra seperti dilasi, erosi, dan deteksi kontur dapat digunakan untuk mengidentifikasi tabel dan mengekstrak informasi tabel. Data tabel seringkali tersebar dalam berbagai gambar dan dokumen yang dapat ditemukan secara luas disektor bisnis maupun di internet. Namun, pada saat

melakukan proses pengenalan atau pengambilan informasi dari tabel tersebut, penting untuk memiliki pemahaman yang kuat terhadap struktur tabel itu sendiri. Hal ini memungkinkan kita untuk memperoleh informasi secara komprehensif dan akurat. Hal tersebut dapat sangat berguna dalam berbagai aplikasi seperti analisis data hingga pengolahan informasi lebih lanjut. Oleh karena itu, *table recognition* diperlukan untuk mengekstrak informasi dari tabel pada gambar atau dokumen.

2. METODOLOGI

Metodologi penelitian digambarkan dalam bentuk diagram alir sebagai berikut :



Gambar 1. Diagram Alir Metodologi Penelitian

Sumber :

https://www.researchgate.net/publication/338235695_Metode-Metode_Penelitian_Dalam_Penulisan_Jurnal_Ilmuiah_Elektronik

3. HASIL DAN PEMBAHASAN

3.1 HASIL

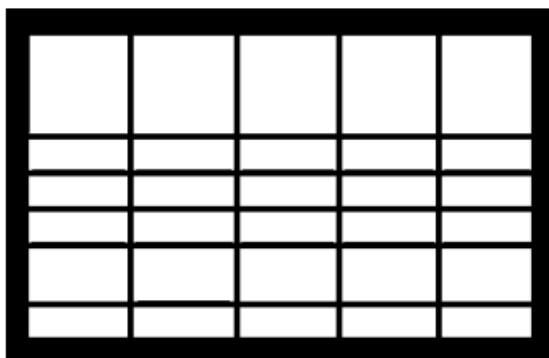
Berikut akan digunakan beberapa gambar untuk menguji hasil *table recognition* yang sudah dikembangkan. Citra yang akan diuji adalah sebagai berikut.

Language	No. of chars (millions)	No. of words (millions)	Char error rate (%)	Word error rate (%)
English	39	4	0.5	3.72
EFIGSD	213	26	0.75	5.78
Russian	38	5	1.35	5.48
Simplified Chinese	0.25	NA	3.77	NA
Hindi	1.4	0.33	15.41	69.44

Gambar 2. Citra Masukan Pertama

Sumber :

<https://d3i71xaburhd42.cloudfront.net/0a482d33dc0c7ca08cc333158c5c0e82d4c04560/8-Table1-1.png>



Gambar 3. Citra Output Pipeline *Region Extraction* Citra Pertama

Sumber : Munir, 2023

```
[
  [
    "Language",
    "No. of chars (millions)",
    "No. of words (millions)",
    "Char error rate (*)",
    "Word error rate (%)"
  ],
  ["English", "39", "4", "0.5", "3.72"],
  ["EFIGSD", "213", "26", "0.75", "5.78"],
  ["Russian.", "38", "5", "1.35", "5.48"],
  ["Simplified Chinese", "0.25", "NA", "3.77", "NA"],
  ["Hindi", "1.4", "0.33", "15.41", "69.44"]
]
```

Gambar 4. Citra Output Pipeline OCR Citra Pertama

Sumber : Munir, 2023

Study Time vs. Grades

Student	Study Time (hours)	Grade
Bob	2	84
Carlos	4	91
Cindy	5	92
Florence	3	89
Kim	4	88
Lori	4	93
Marisa	1	78
Pat	2	89
Thomas	5	94
Wendy	2.5	87

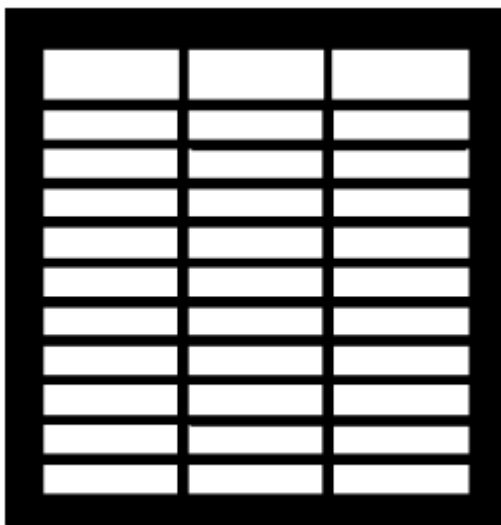
Gambar 5. Citra Masukan Kedua

Sumber : https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSrrpHWRlQdbO-SfMg6puUFew44paOOX78C_D-yeDiM502vryJ5

https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSrrpHWRlQdbO-SfMg6puUFew44paOOX78C_D-yeDiM502vryJ5

Student	Study Time (hours)	Grade
Bob	2	84
Carlos	4	91
Cindy	5	92
Florence	3	89
Kim	4	88
Lori	4	93
Marisa	1	78
Pat	2	89
Thomas	5	94
Wendy	2.5	87

Gambar 6. Citra Output Pipeline *Table Extraction*
Citra Kedua
Sumber : Munir, 2023



Gambar 7. Citra Output Pipeline *Region Extraction*
Citra Kedua
Sumber : Munir, 2023

```
[
  ["Student", "Study Time (hours)", "Grade"],
  ["Bob", "2", "84"],
  ["Carlos", "4", "91"],
  ["Cindy", "5", "92"],
  ["Florence", "3", "89"],
  ["Kim", "4", "88"],
  ["Lori", "4", "93"],
  ["Marisa", "1", "78"],
  ["Pat", "2", "89"],
  ["Thomas", "5", "94"],
  ["Wendy", "2.5", "87"]
]
```

Gambar 8. Citra Output Pipeline OCR Citra Kedua
Sumber : Munir, 2023

3.2 PEMBAHASAN

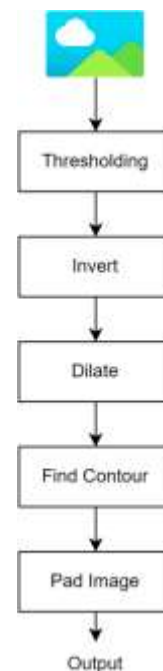
Untuk melakukan *Table Recognition* pada gambar dan dokumen, maka gambar atau dokumen akan melalui tiga pipeline yaitu *Table Extraction*, *Region Extraction*, dan OCR. *Pipeline* tersebut dapat diilustrasikan pada gambar berikut:



Gambar 9. *Pipeline Table Recognition*.
Sumber : Rashid, 2022

A. *Table Extraction*

Pipeline ini bertujuan untuk mengekstraksi tabel dari gambar / dokumen. Pada *pipeline table extraction* akan dilakukan beberapa proses seperti pengambangan (*thresholding*), *invert*, *dilate* (*dilatasi*), *find contour*, dan *pad image* yang dapat dilihat pada gambar berikut.



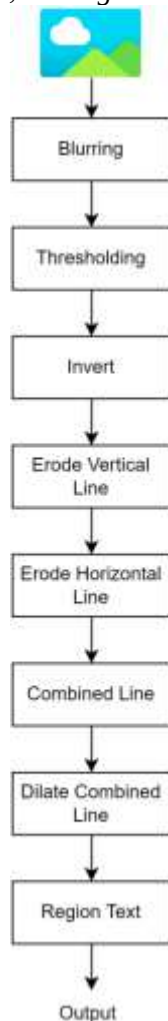
Gambar 10. Alur Proses Pada *Pipeline Table Extraction*
Sumber : Rashid, 2022

Thresholding akan menghasilkan citra biner dari masukan gambar / dokumen. Hasil *thresholding* tersebut akan dilakukan *Invert* untuk mengubah nilai pixel pada gambar menjadi pixel yang berlawanan. *Dilatasi* akan menebalkan atau melebarkan objek pada citra

yang nantinya akan dicari kontur berbentuk persegi yang merupakan representasi dari tabel. Setelah tabel ditemukan, maka gambar akan di-crop pada bagian tabel tersebut dan akan dilakukan padding sebesar 5% dari tinggi gambar masukan.

B. Region Extraction

Pipeline kedua yaitu *region extraction*. *Pipeline* ini bertujuan untuk mengekstraksi region-region dari teks. Pada *pipeline* tersebut akan dilakukan beberapa proses seperti *blurring*, *thresholding*, *invert*, *erode vertical line*, *erode horizontal line*, *combined line*, *dilate combined line*, dan *region text*



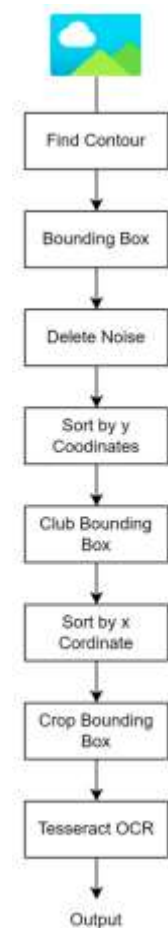
Gambar 11. Alur Proses Pada *Pipeline Region Extraction*

Sumber : Rashid, 2022

Thresholding akan menghasilkan citra biner dari masukan gambar / dokumen hasil dari *pipeline* sebelumnya. Hasil *thresholding* tersebut akan dilakukan *Invert* untuk mengubah nilai pixel pada gambar menjadi pixel yang berlawanan. Berdasarkan hasil *invert* tersebut akan dilakukan erosi untuk mendapatkan garis horizontal dan garis vertical. Garis horizontal dan garis vertical akan digabungkan dan akan dilakukan dilatasi untuk menebalkan garis-garis tersebut. Berdasarkan hasil tersebut akan di-invert kembali dan menghilangkan semua elemen diluar tabel untuk mendapatkan *region text*.

C. OCR

Pipeline terakhir yaitu OCR. *Pipeline* ini bertujuan untuk melakukan pengenalan karakter pada tabel setiap selnya dan menyusun hasil OCR menjadi informasi terstruktur. Pada *pipeline* ini akan dilakukan beberapa proses seperti *find-contour*, *bounding box*, *delete noise*, *sort by y coordinate*, *club bounding box*, *sort by x coordinate*, *crop bounding box*, dan *tesseract OCR*.



Gambar 12. Alur Proses Pada *Pipeline OCR*
Sumber : Smith, 2023

Find Contour bertujuan mencari kontur berdasarkan hasil dari *pipeline Region Extraction* yang sudah dilakukan sebelumnya. Berdasarkan hasil kontur-kontur yang ditemukan akan dibentuk *bounding box* disetiap konturnya. Terkadang terdapat beberapa *noise* dari pembentukan *Bounding Box* sehingga perlu dihilangkan dari daftar *bounding box*. Daftar *bounding box* akan diurutkan berdasarkan nilai koordinat y dan dikelompokkan. Setelah dikelompokkan akan diurutkan berdasarkan koordinat x. Setiap *bounding box* pada citra akan di-crop berdasarkan koordinat dari *bounding box* tersebut. Setelah di-crop akan dilakukan pengenalan menggunakan *tesseractOCR* secara satu-persatu.

KESIMPULAN

Metode pengolahan citra dapat digunakan untuk melakukan *table recognition*. Dengan

menggunakan metode tersebut didapatkan hasil yang cukup baik. Terdapat beberapa hal yang bisa ditingkatkan seperti mengubah parameter yang terdapat pada kode seperti konfigurasi OCR yang digunakan, serta parameter-parameter pada pengolahan citra.

DAFTAR PUSTAKA

- Munir, Rinaldi., (2023). Bahan Kuliah IF4073 Interpretasi dan Pengolahan Citra. Program Studi Informatika ITB.
- Rashid, S. F., Akmal, A., Adnan, M., Aslam, A. A., & Dengel, A. (2022). Table Recognition in Heterogeneous Documents Using Machine Learning. In 2022 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2022 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE.
<https://doi.org/10.1109/icdar.2017.132>
- Smith, R. (2023). An Overview of the Tesseract OCR Engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2023) Vol 2. Ninth International Conference on Document Analysis and Recognition (ICDAR 2023) Vol 2. IEEE.
<https://doi.org/10.1109/icdar.2007.4376991>.