우리 FISA 2025.09.26
AI 엔지니어링 – 2차 기술세미나

컴퓨터는 어떻게 텍스트 이미지를 읽을까?

OCR 알아보기

발표자 : 김성현

목차

01
주제 선택 배경

왜 OCR을??

Black Carbon (BC)

블랙카본

BC(Black Carbon), 미세먼지 중 가장 강력하게 빛을 흡수하는 성분으로, 석유, 석탄 등 화석연료 등의 불완전연소 그 때문에 생기는

MHB

산적 상태에서만 위험한 물질

MHB (Material Hazardous in Bulk), 국제해상위험물규칙의 포장 위험물 분류 기준을 충족하지는 않으나, 산적 형태로 운송되는 경우에 화학적 위험성을 가지는 물질. 만약 어떤 물질이 아래와 같이 한 가지 이상의 위험성을 나타내는 것으로 확인되는 경우에는 산적 상태에서만 위험한 물질(MHB)로 분류하여 운송해야 함

화학적 위험성	위험성 표기방법
가연성 고체	CB
	H
	/F
	/T
	X
	R
	/H

MP

해양오염물질

MP (Marine Pollutant), 해양 유출 시 또는 수생 환경에 위해를 줄 수 있는 물질, 재료 또는 제품. 수생 생물에게 급성(acute) 또는 만성(chronic) 독성이 있는 경우 해양오염 물질로 분류됨.

PDF에서 텍스트를 추출해야 되는 상황

배운 CNN을 활용해서 충분히 가능!

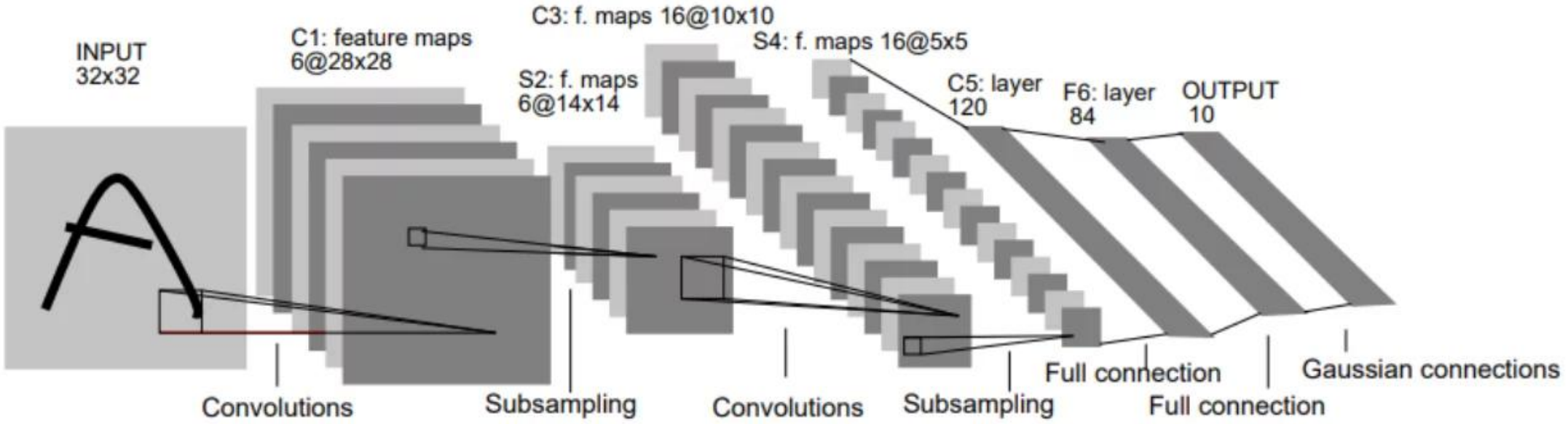
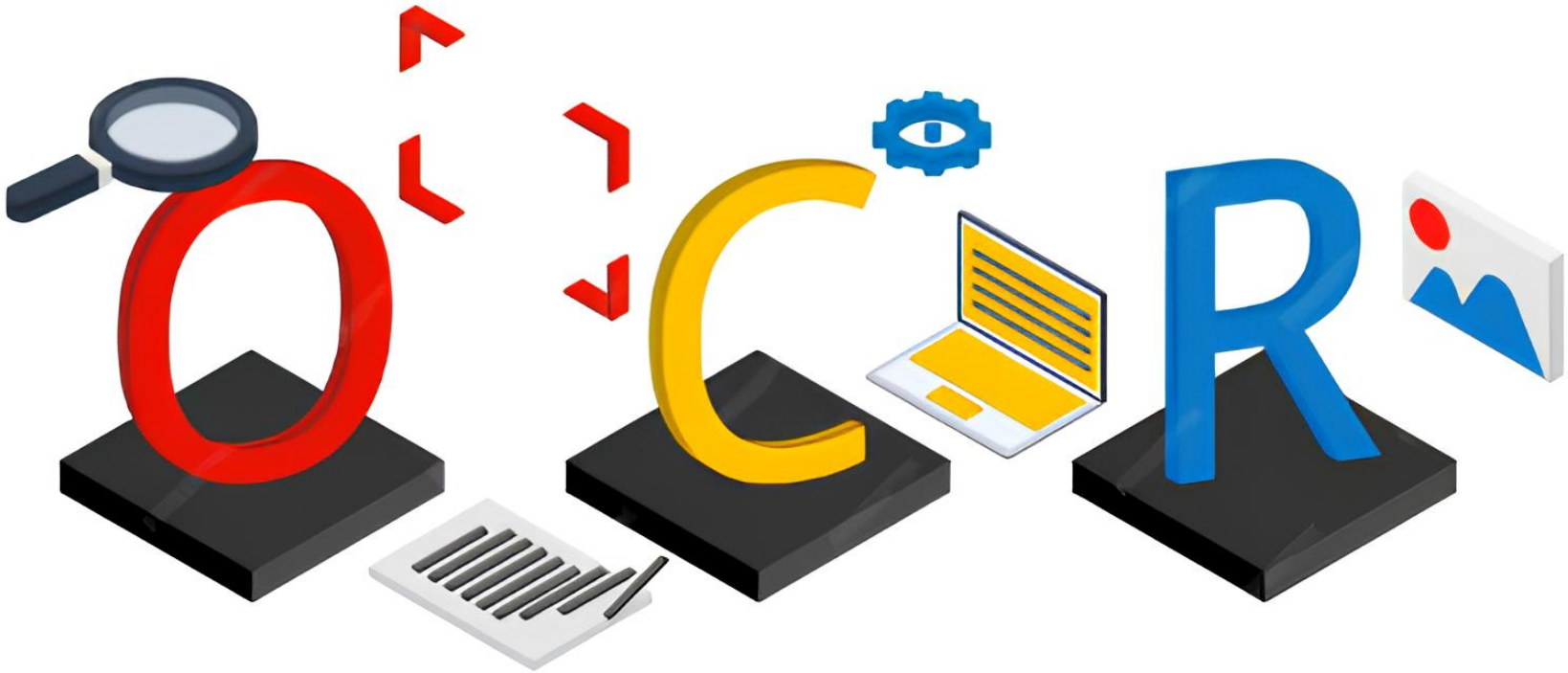


Figure 1: LeNet-5

https://healthyinsight.co.kr/%EC%9D%B4%EB%AF%B8%EC%A7%80-%EB%B6%84%EB%A5%98-cnn/

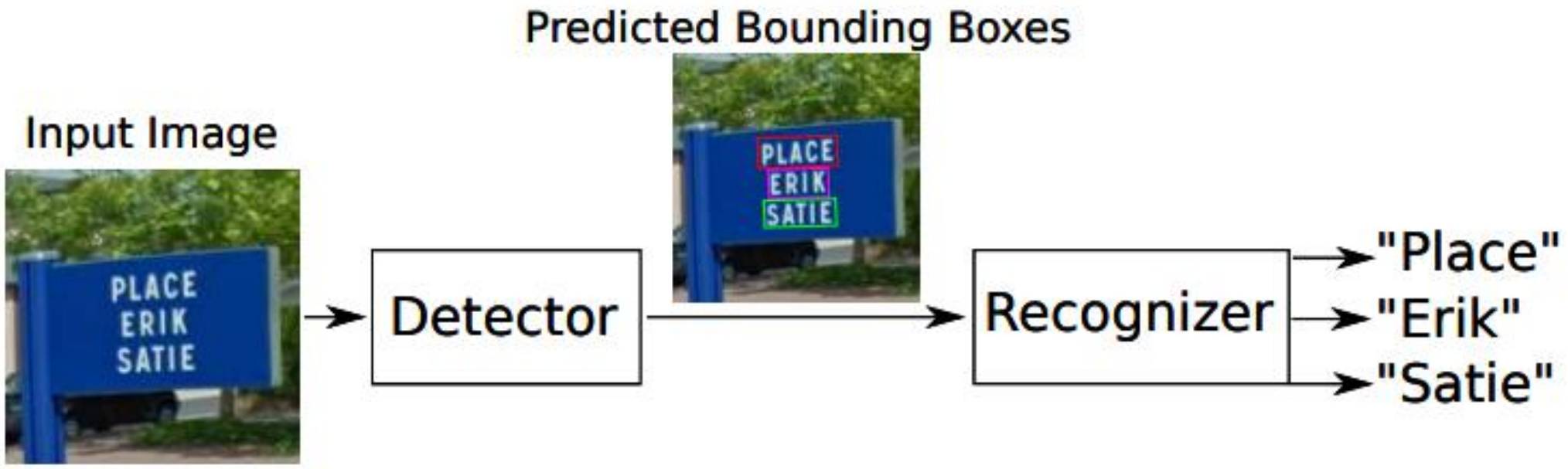
02
OCR이란 어떤 기술인가?

광학 문자 인식



OCR의 동작 방식

1단계. Text Detection ⟶ 2단계. Text Recognition



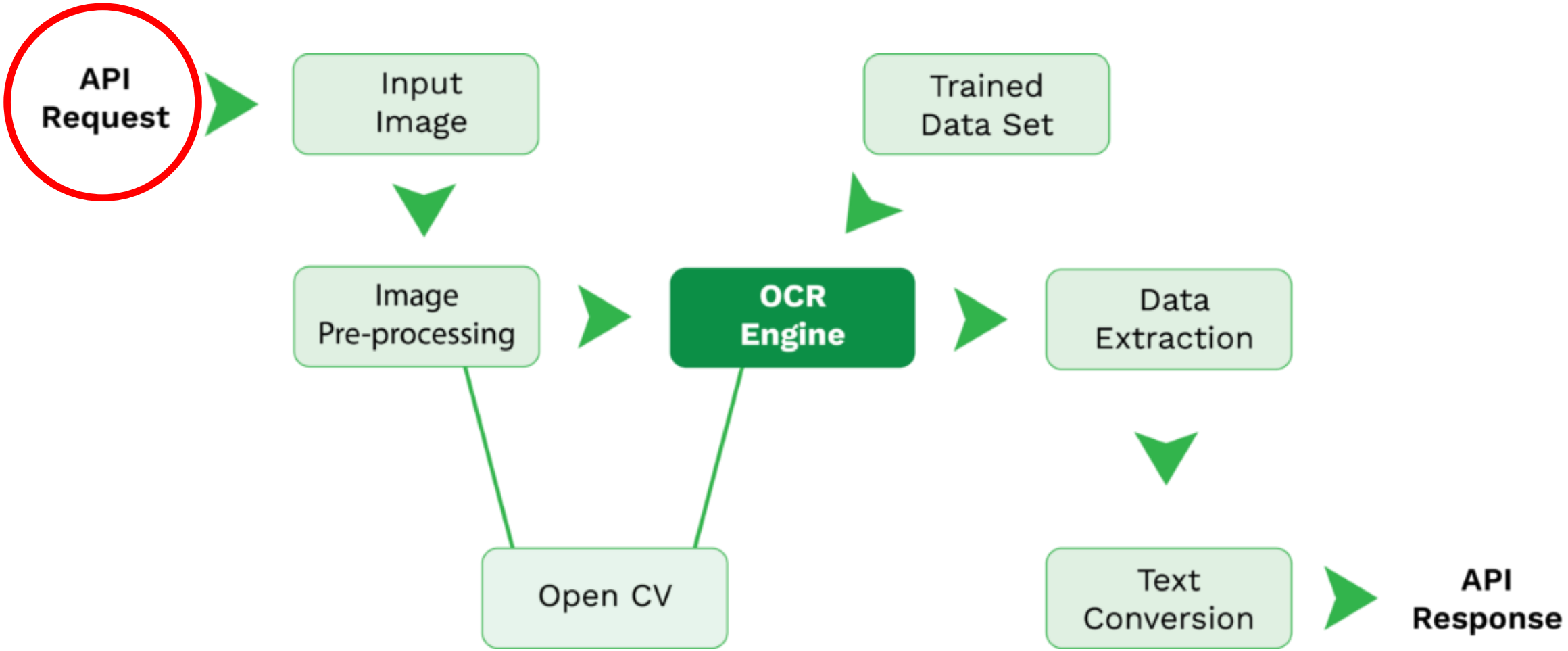
https://velog.io/@xpelqpdj0422/11.-OCR-%EA%B8%B0%EC%88%A0%EC%9D%98-%EA%B0%9C%EC%9A%94

핵심 엔진 Tesseract

Tesseract OCR

https://github.com/tesseract-ocr/tesseract

동작 방식



https://www.klippa.com/en/blog/information/tesseract-ocr/

COMPANY

Company B.V.
Straatweg 4
2244AZ Utrecht
Netherlands

info@bedrijfsnaam.com

KvK: 33344555
BTW: NL000099998857
Bank: NL 01 BANK 0043 5678 22

Klippa App B.V.
t.a.v. J. Doe
Lübeckweg 2 9723 HE Groningen
Netherlands

Invoice 2021_0567

Description	Quantity	Price	Total	VAT
Object 097	2	€50,00	€100,00	21%
Object 524	3	€10,00	€30,00	21%
Object 8032	6	€4,50	€27,00	21%

Subtotal	€157,00
21% BTW	€32,97
Total	€189,97

CRNN

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition

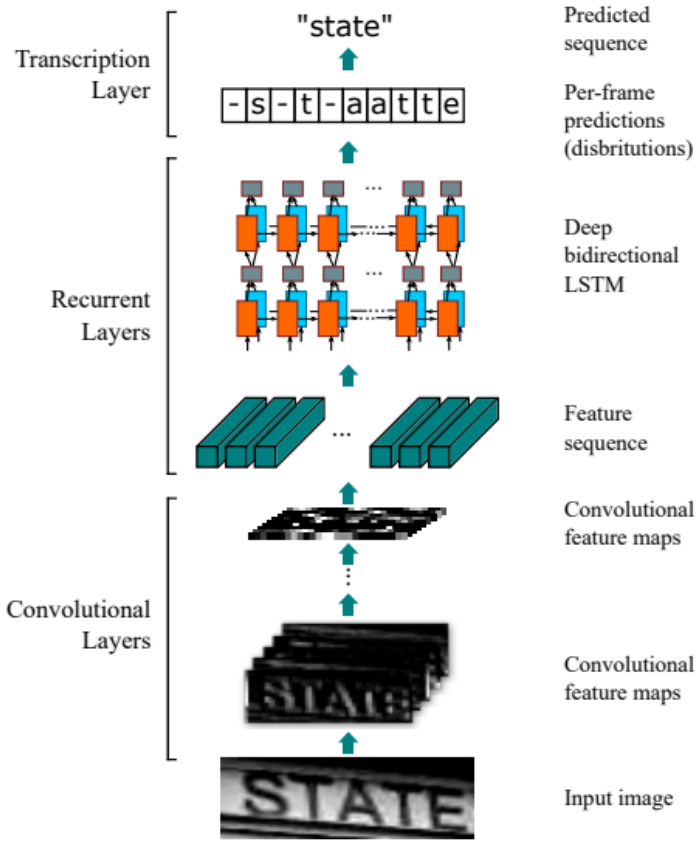


Figure 1. The network architecture. The architecture consists of three parts: 1) convolutional layers, which extract a feature sequence from the input image; 2) recurrent layers, which predict a label distribution for each frame; 3) transcription layer, which translates the per-frame predictions into the final label sequence.

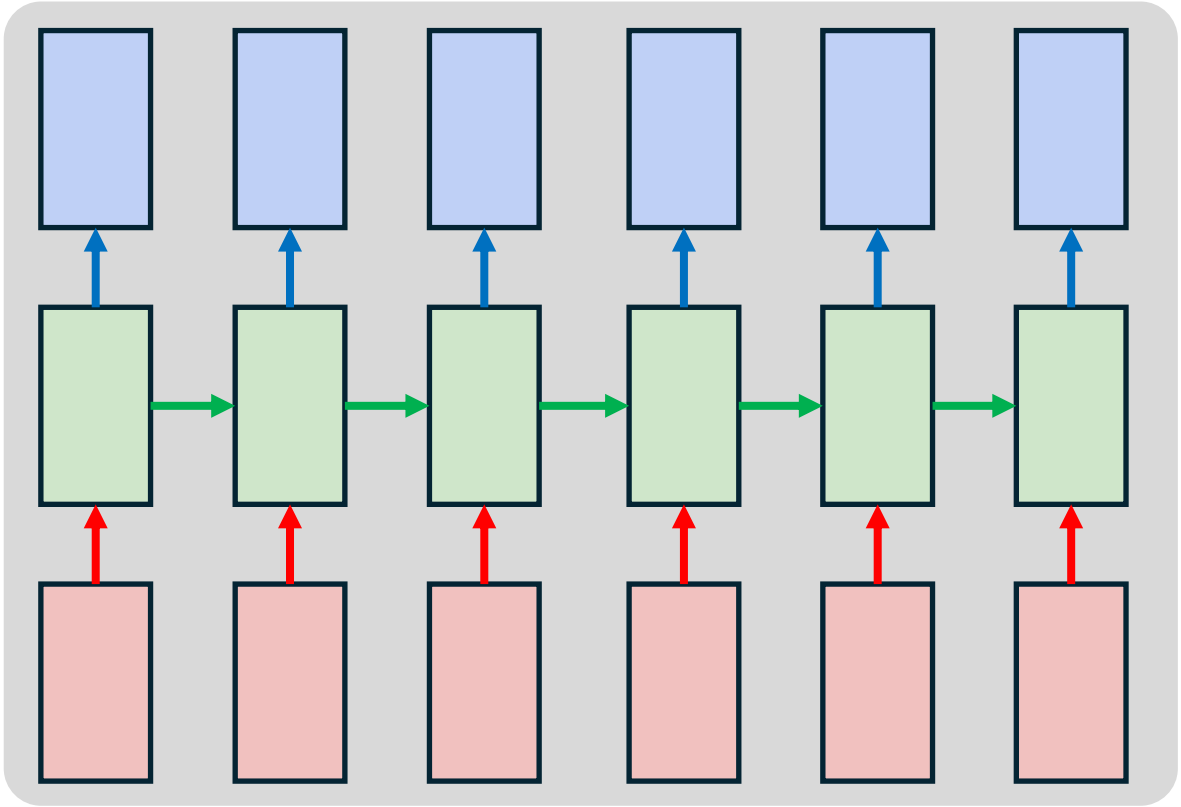
1. Convolutional Layer (합성곱 계층 - 특징 추출)

2. Recurrent Layers (순환 계층 - 순서 모델링)

3. Transcription Layer (전사 계층 - 텍스트 변환)

CTC (Connectionist Temporal Classification)

Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks



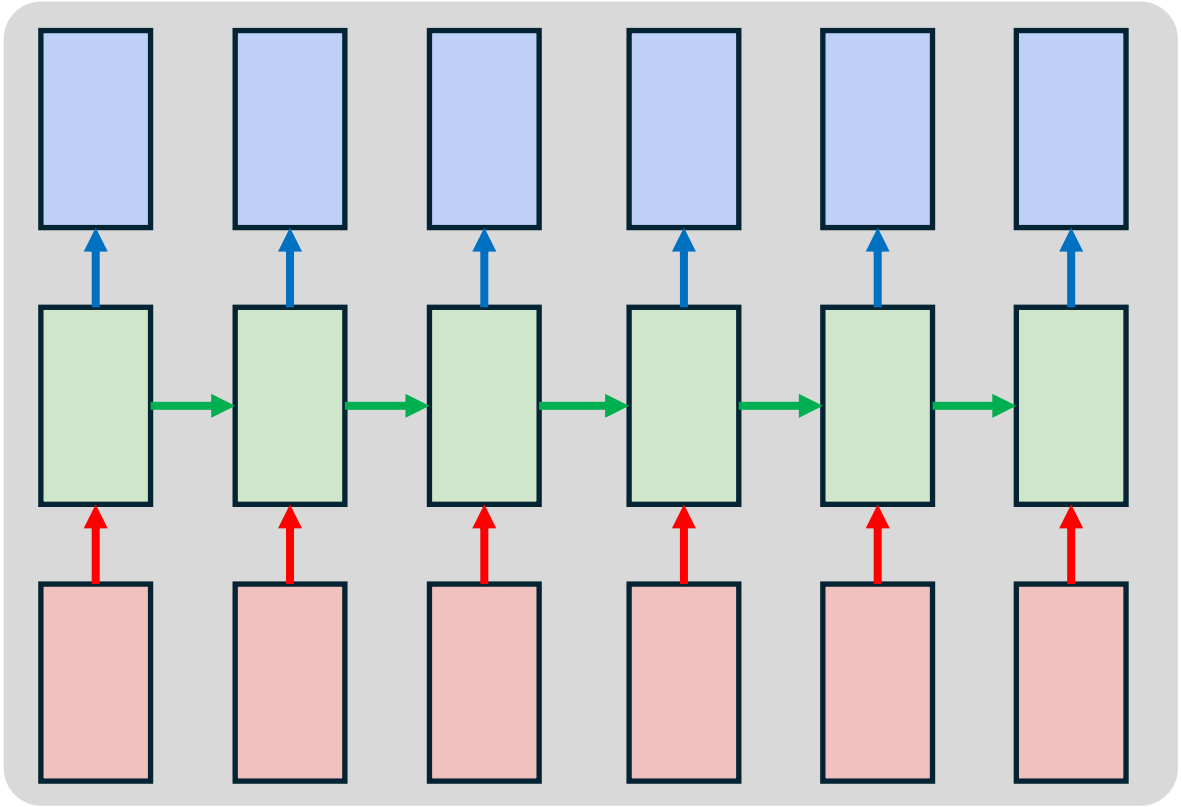
HHEEELLLLLLOOO

=> HELLO ??

Blank '–' 추가!

CTC (Connectionist Temporal Classification)

Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks



-HEEE-LL-LLOOOO--

-HE-L-LOO--------

-HEEE-LLLL-LO----

→ -H-E-L-LO-

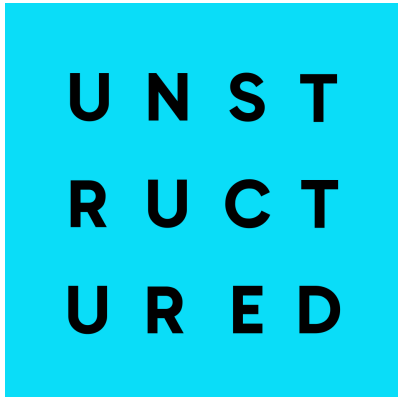
→ HELLO

03
여러가지 데이터로 OCR 해보기

OCR 라이브러리

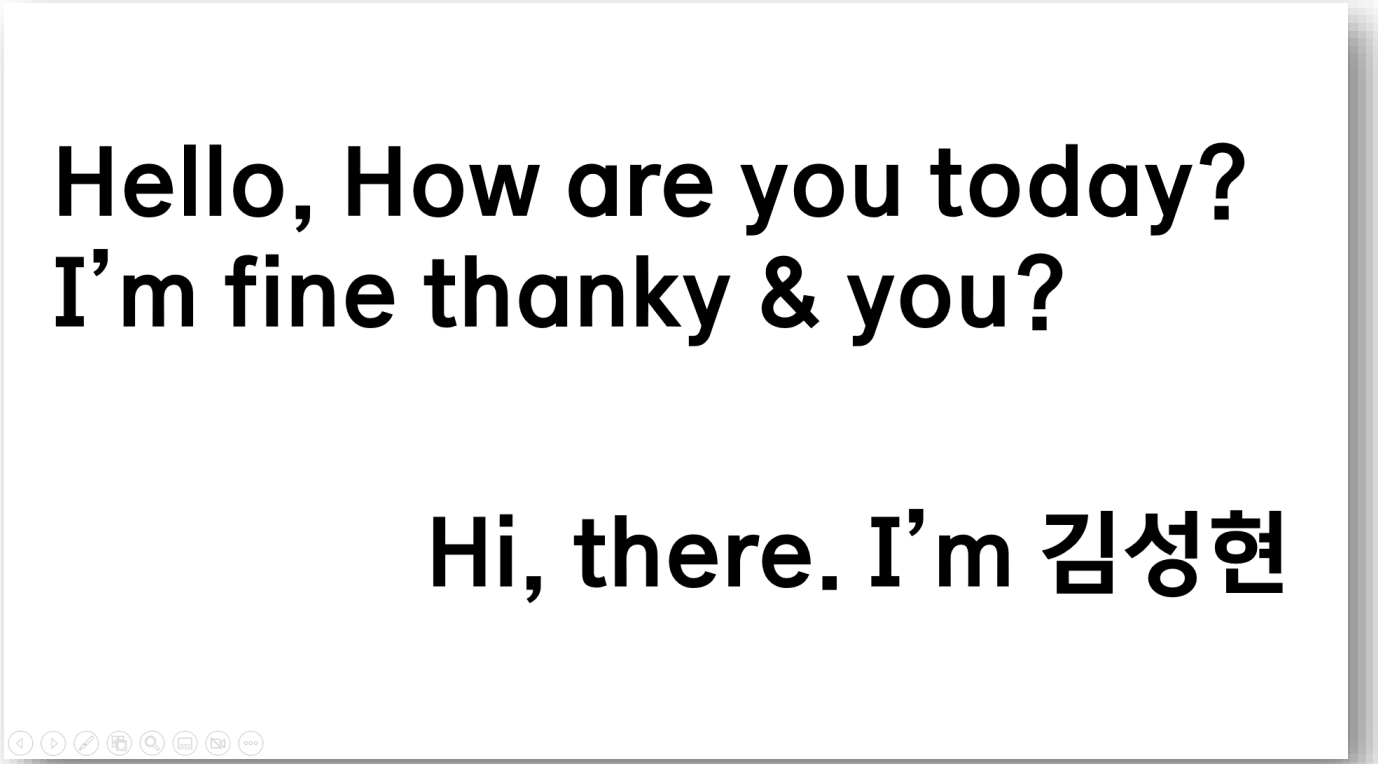
PyTesseract

EasyOCR

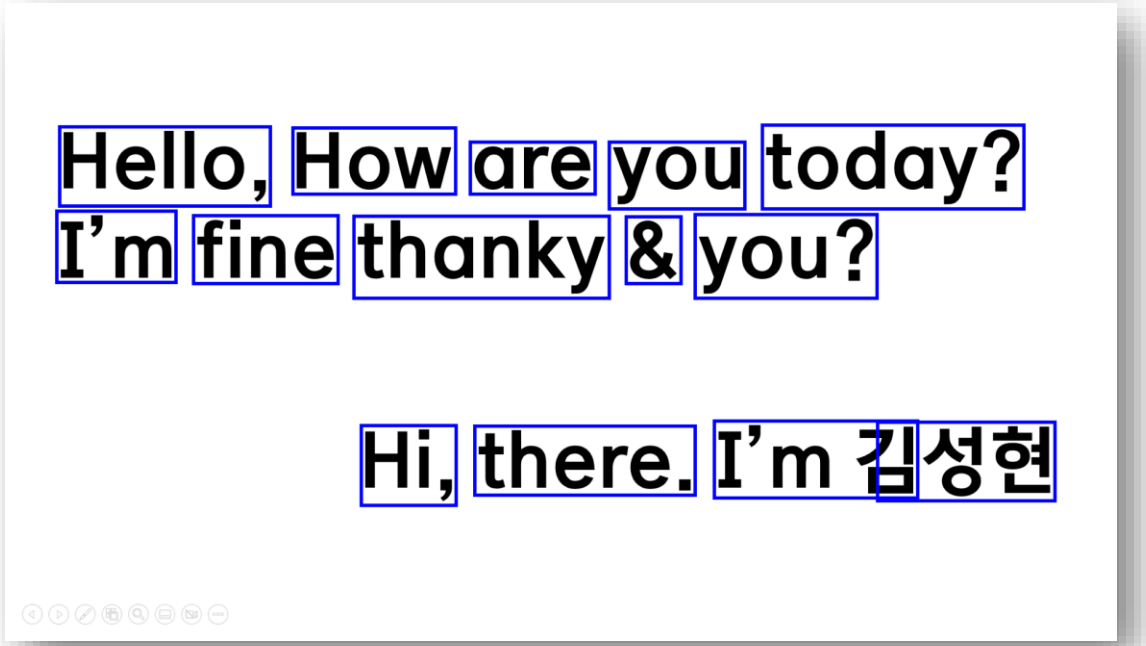


Unstructured

Test Image



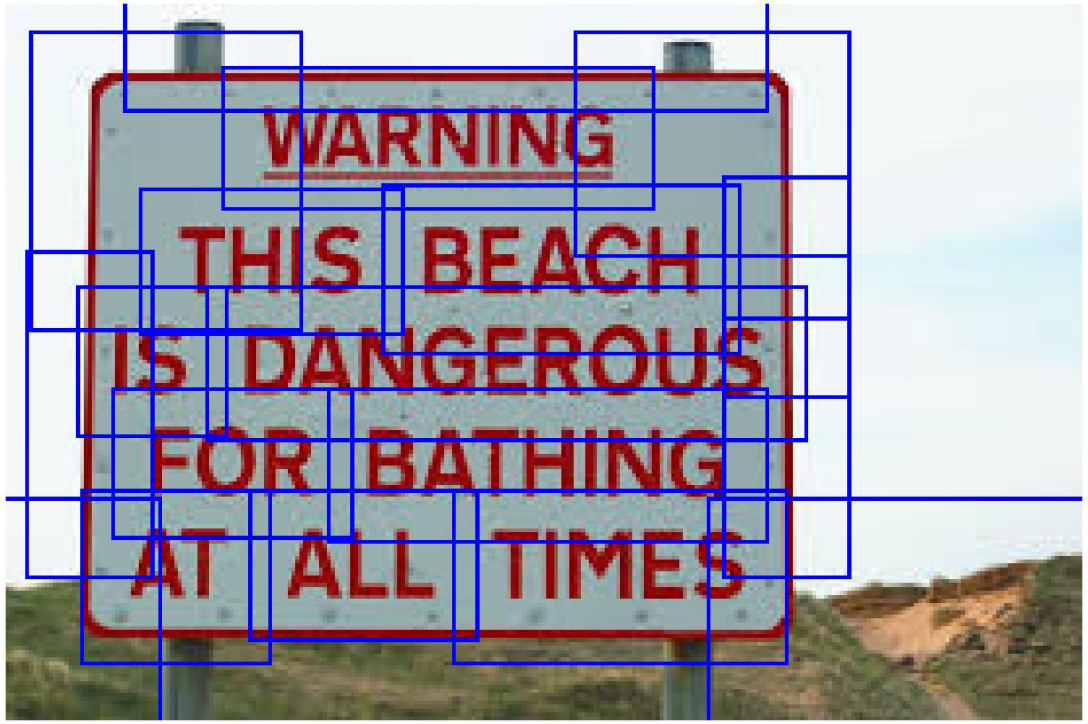
Text Extraction



Hello, How are you today?
I'm fine thanky & you?

Hi, there. 'm 김성현

PyTesseract의 한계점



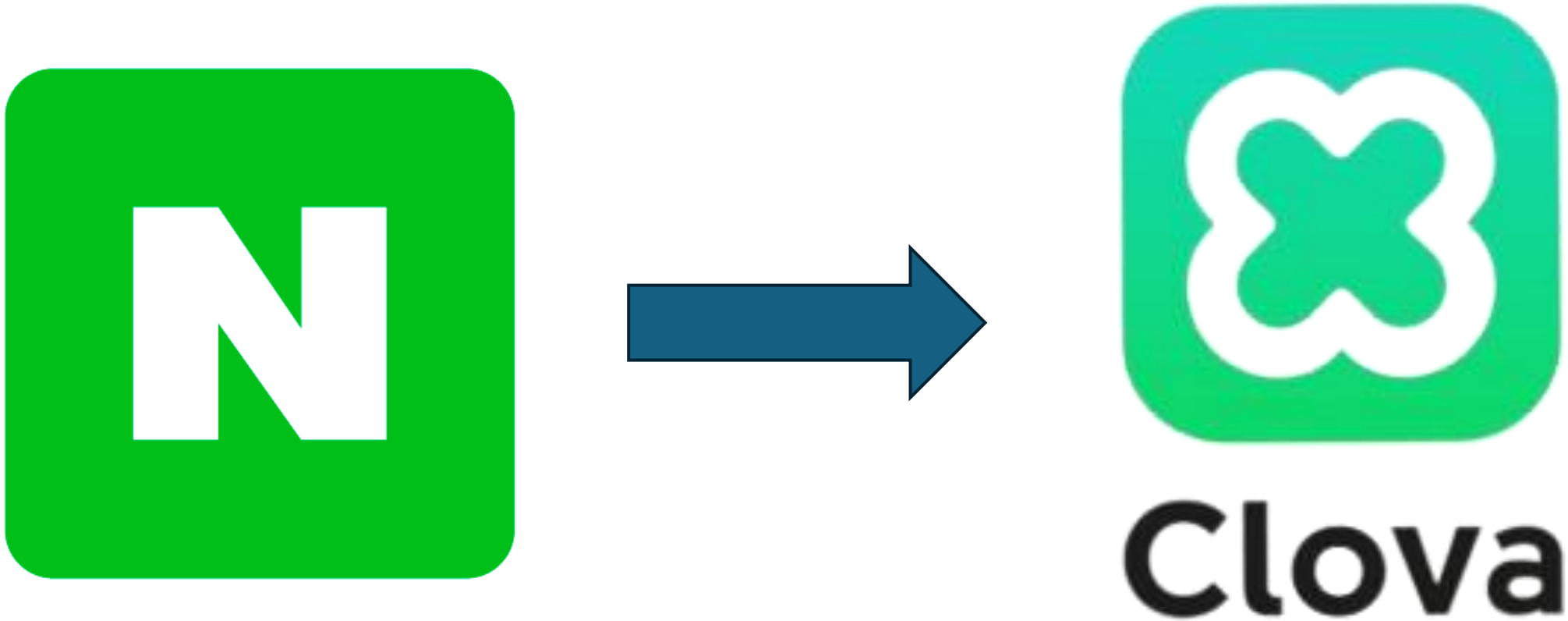
PyTesseract

PyTesseract

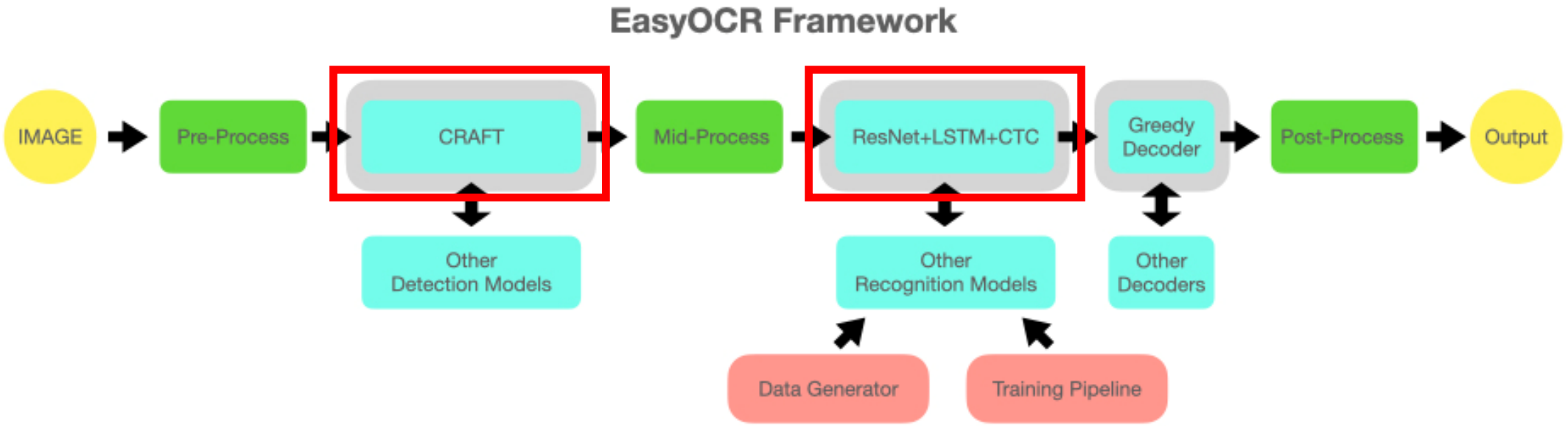
EasyOCR

Unstructured

Easy OCR



Easy OCR



https://github.com/JaidedAI/EasyOCR

CRAFT

Character Region Awareness for Text Detection

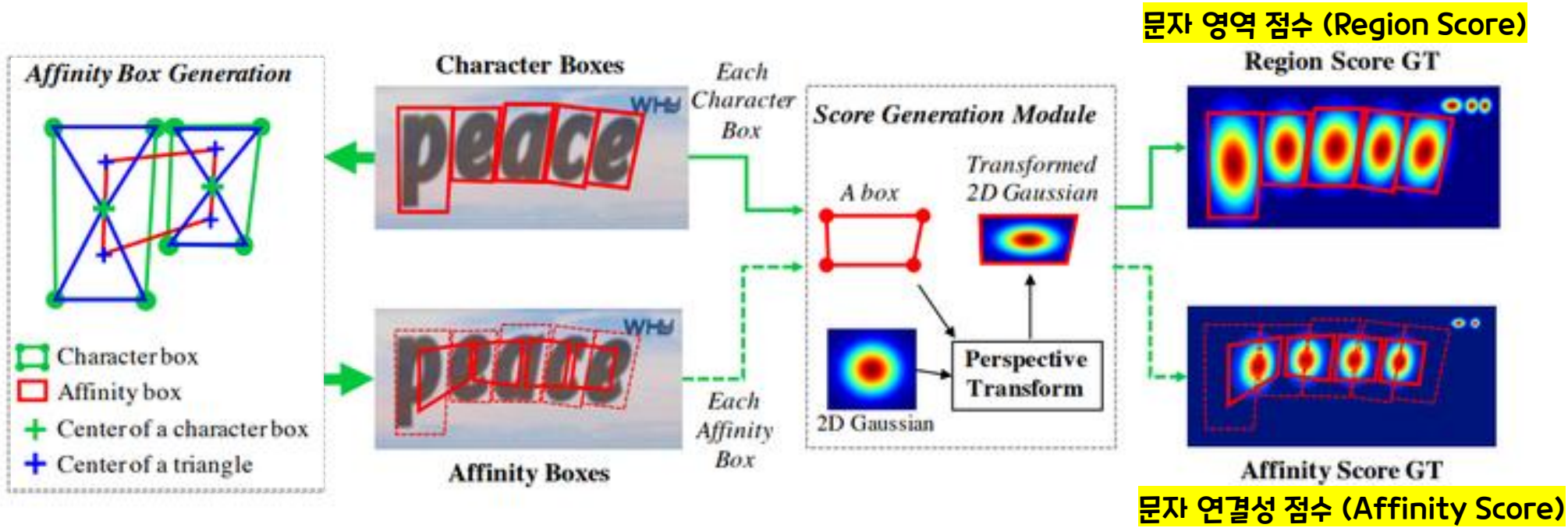


Figure 3. Illustration of ground truth generation procedure in our framework. We generate ground truth labels from a synthetic image that has character level annotations.

CRAFT - Weakly Supervised Learning

Character Region Awareness for Text Detection

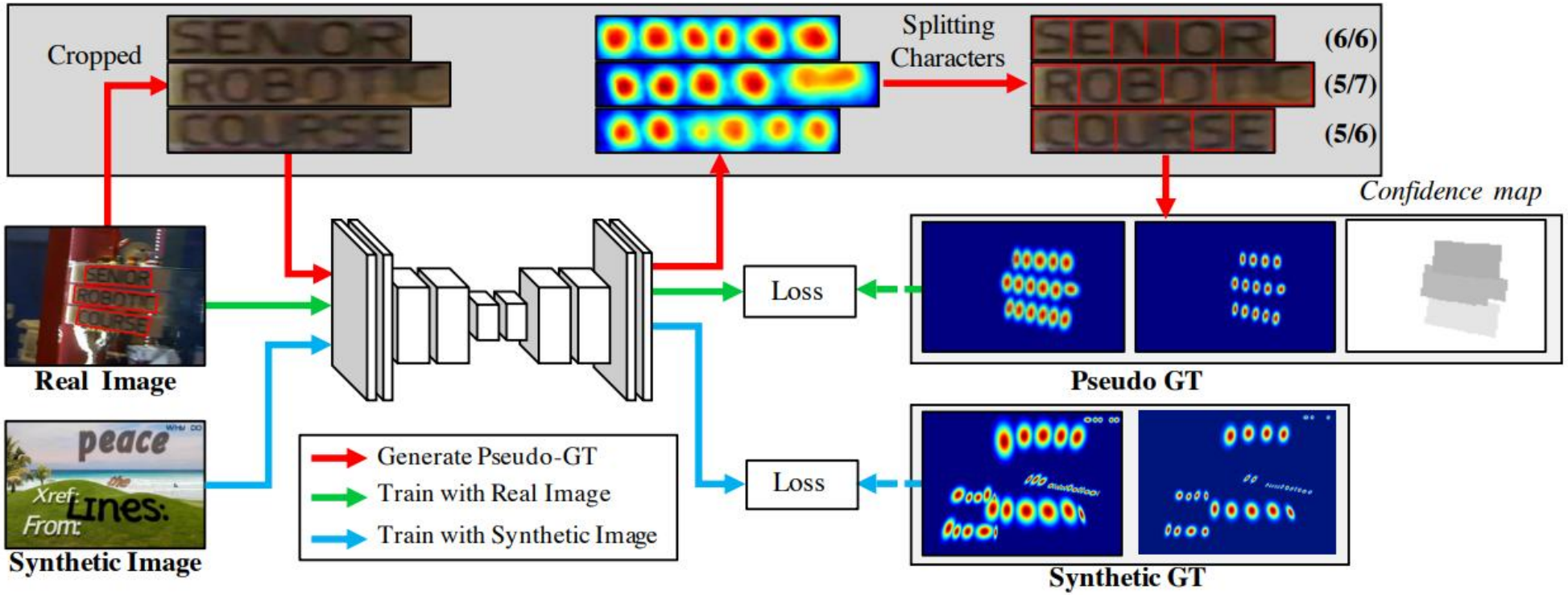
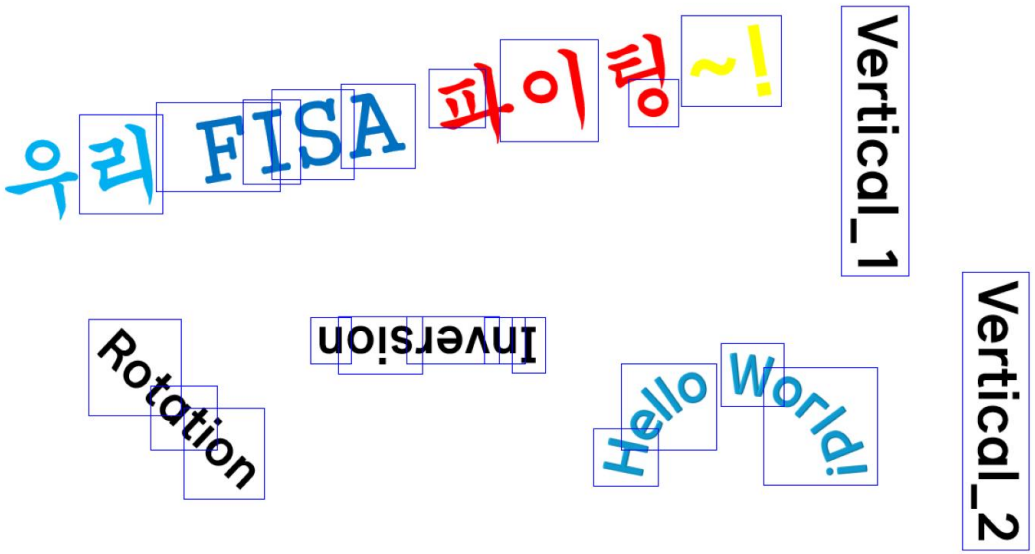
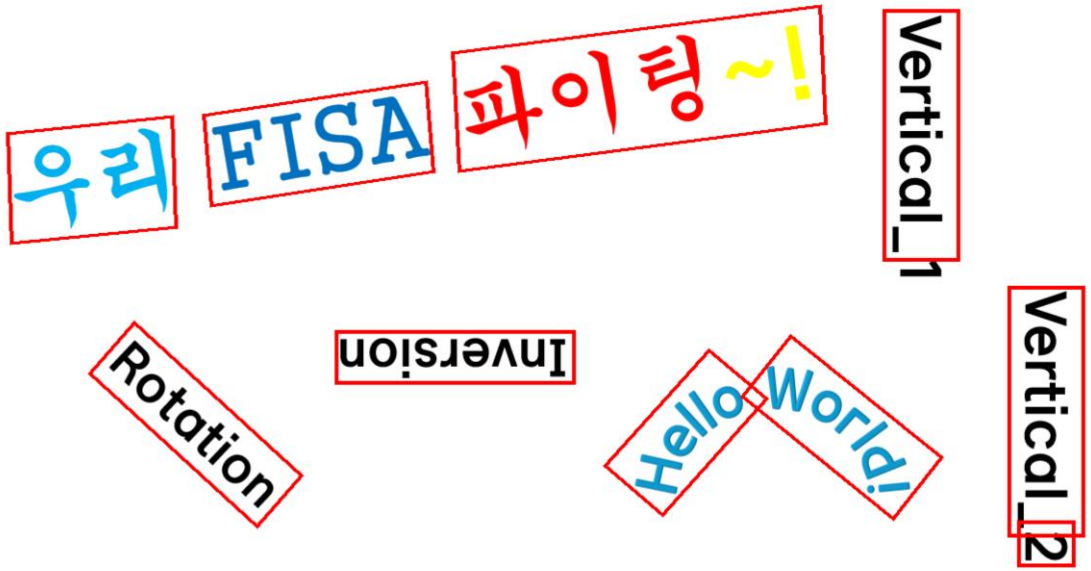


Figure 4. Illustration of the overall training stream for the proposed method. Training is carried out using both real and synthetic images in a weakly-supervised fashion.

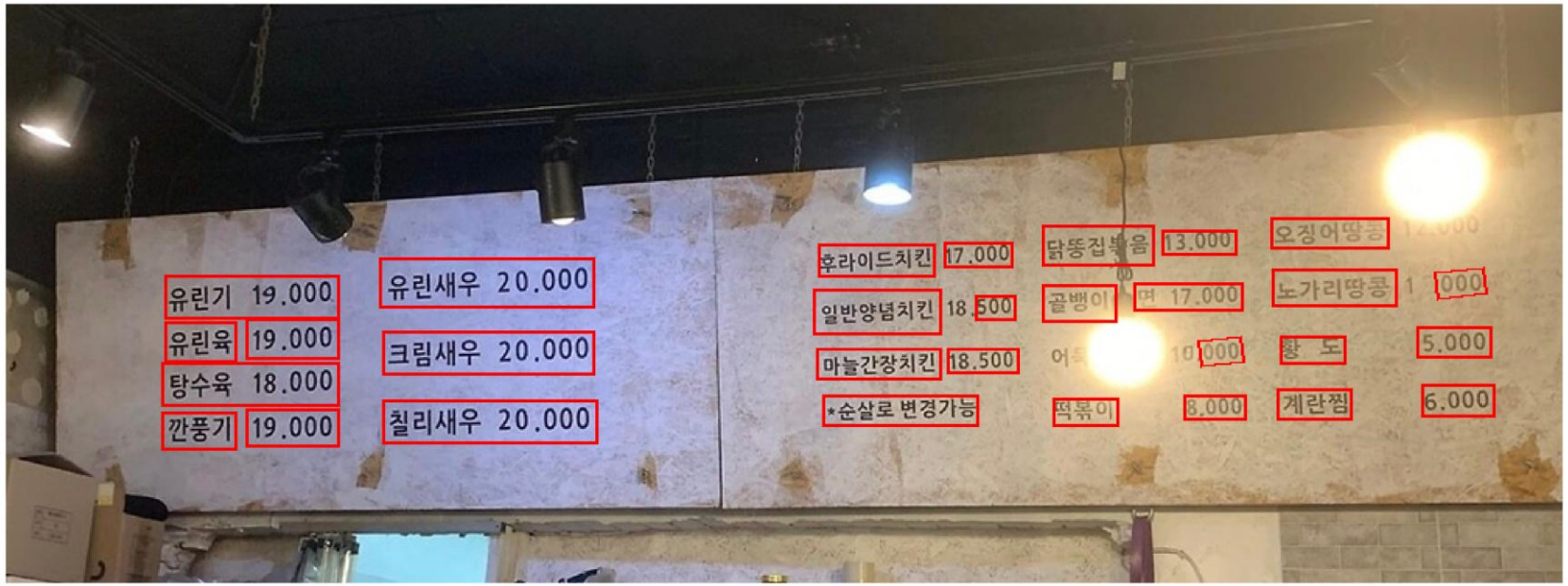
PyTesseract



Easy OCR




광한루원
춘향테마파크
Chunhyang Thema Park
법원.검찰청
고속도로
Express way
남원역
Namwon Stn
시청.시의회
City Hall
200m


유린기 19.000
유린육 19.000
탕수육 18.000
깐풍기 19.000
유린새우 20.000
크림새우 20.000
칠리새우 20.000
후라이드치킨 17.000
일반양념치킨 18,500
마늘간장치킨 18,500
*순살로 변경가능
닭똥집볶음 13.000
골뱅이 17.000
떡볶이 8,000
오징어땅콩
노가리땅콩
계란찜 6.000

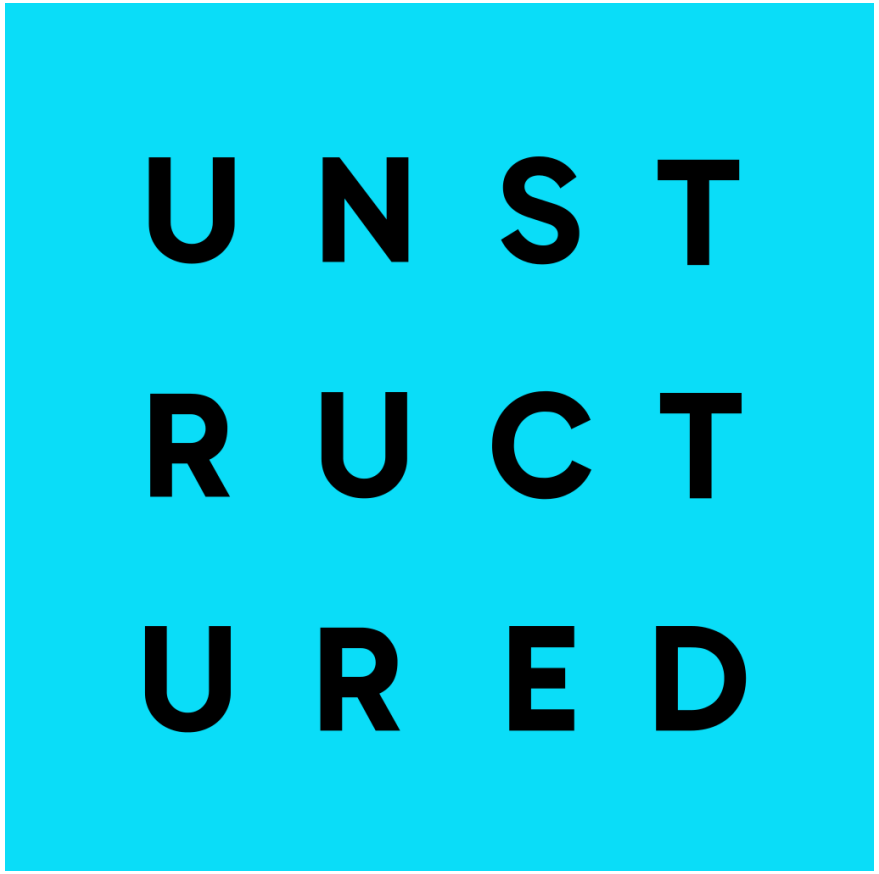
PyTesseract

PyTesseract

EasyOCR

Unstructured

Unstructured



문서 이해 및 구조화

Partitioning

Cleaning

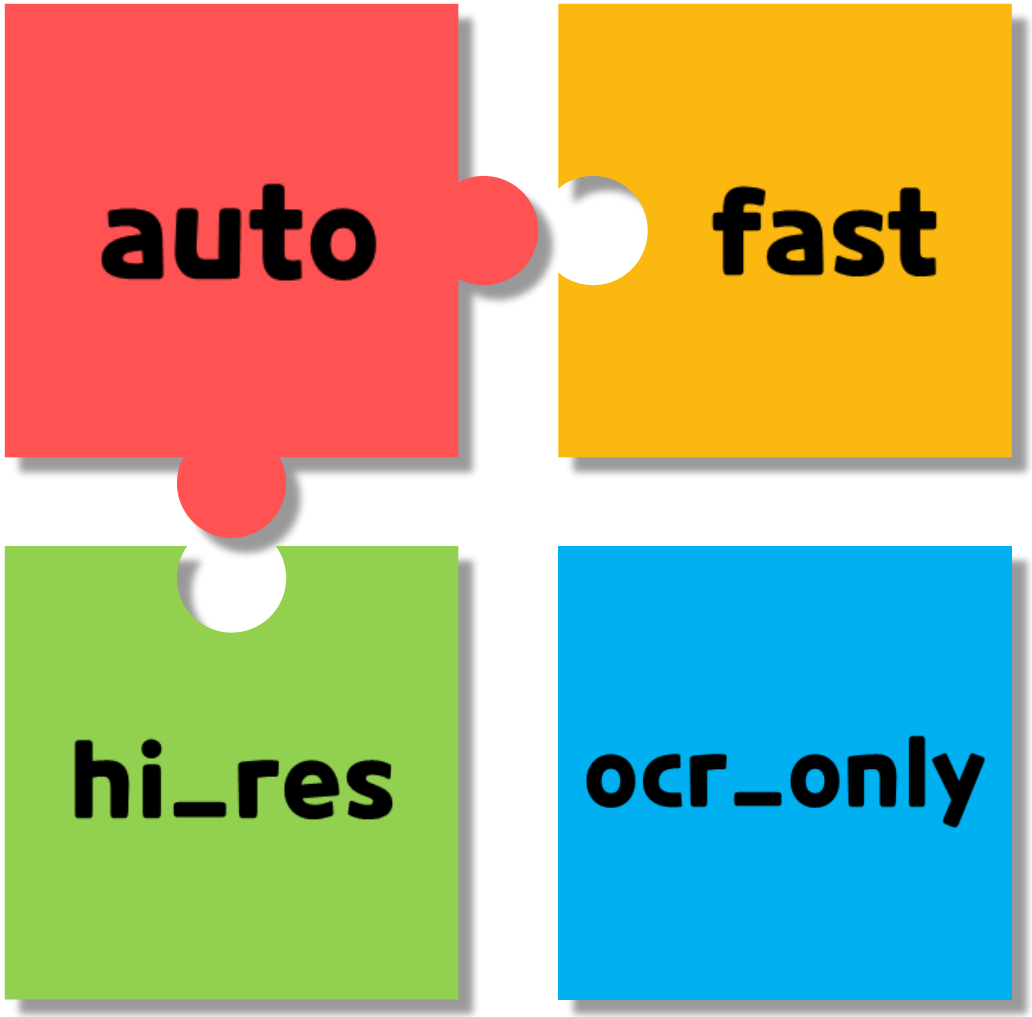
Extracting

Staging

Chunking

Embedding

Partitioning



Partitioning

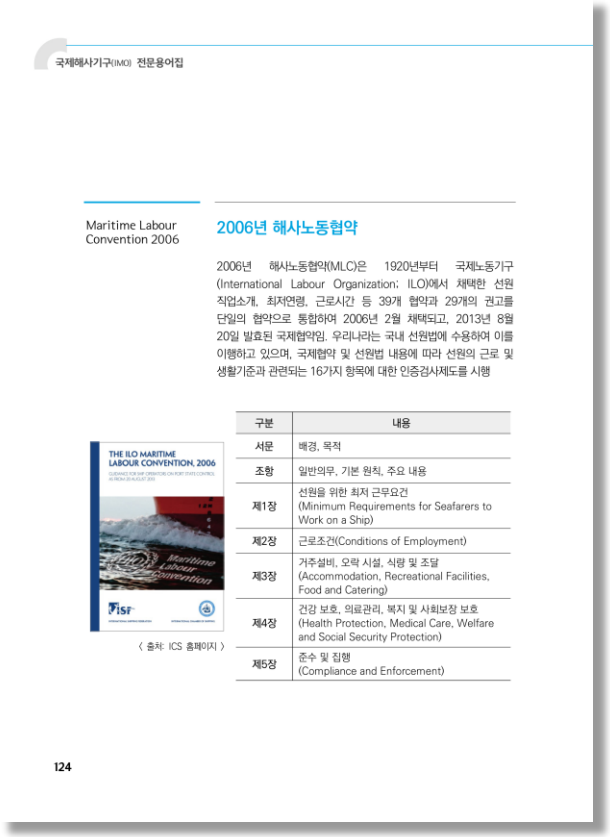
auto

fast

hi_res

ocr_only

지능적인 최적화 문서 처리



hi_res

Partitioning

신속한 텍스트 중심 문서 처리

Maritime Labour Convention 2006

2006 해사노동협약

2006년 해사노동협약(MLC)은 1920년부터 국제노동기구(International Labour Organization; ILO)에서 채택한 선원 직업소개, 최저연령, 근로시간 등 39개 협약과 29개의 권고를 단일의 협약으로 통합하여 2006년 2월 채택되고, …

fast

Maritime Labour Convention 2006

2006 해사노동협약

2006년 해사노동협약(MLC)은 1920년부터 국제노동기구(International Labour Organization; ILO)에서 채택한 선원 직업소개, 최저연령, 근로시간 등 39개 협약과 29개의 권고를 단일의 협약으로 통합하여 2006년 2월 채택되고, …

Partitioning

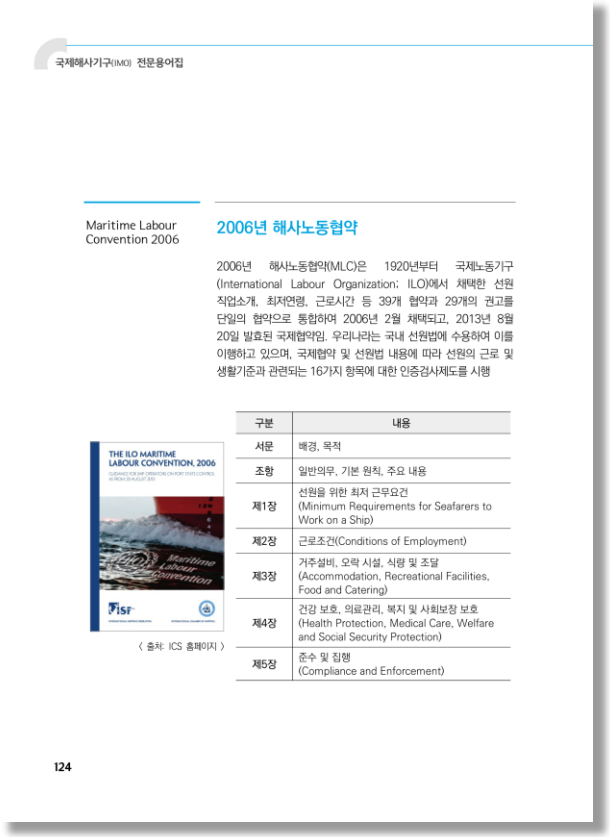
auto

fast

hi_res

ocr_only

정밀한 레이아웃 분석 기반 문서 처리



hi_res

Partitioning

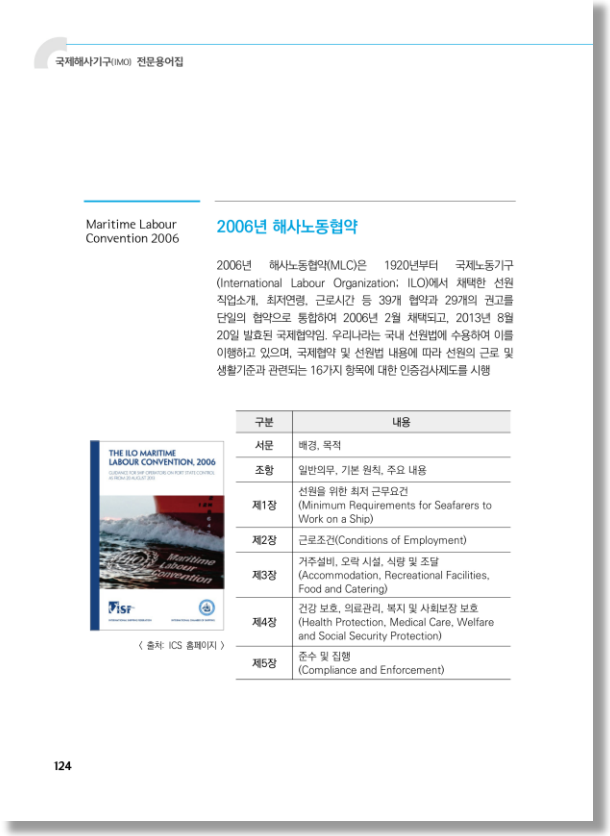
auto

fast

hi_res

ocr_only

이미지 텍스트 추출 전문 전략



ocr_only

Partitioning

	장점	단점
auto	기본 설정으로 편리하고, 균형 잡힌 성능을 보임	세밀한 제어가 불가하고, 내부 로직에 의존함
fast	빠른 분석 속도	text 문서만 가능 (OCR 기능이 없음)
hi_res	매우 높은 정확도, 표, 이미지 추출에 효과적	느린 분석 속도, 고사양 자원 필요(GPU)
ocr_only	이미지/스캔 문서에 필수적	단순 텍스트 나열

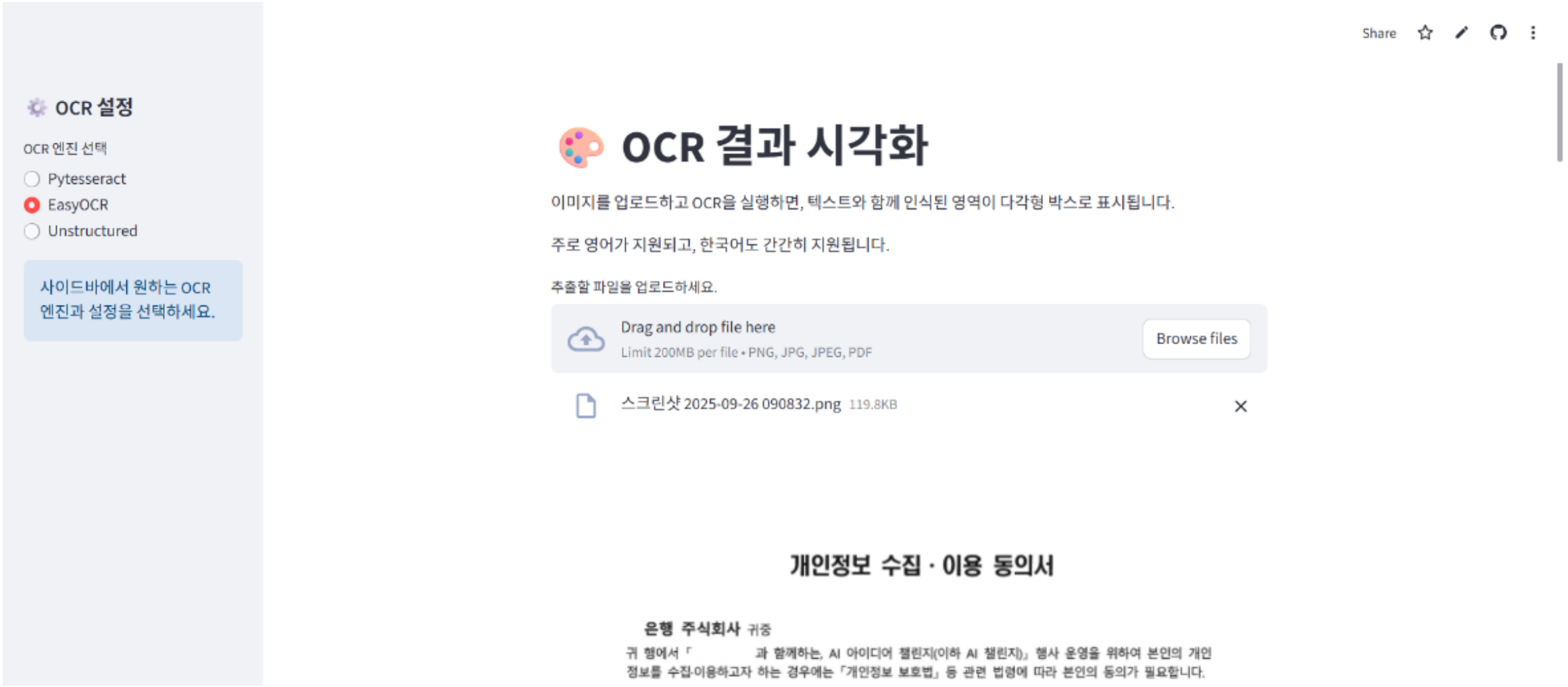
Streamlit page

https://ocrprojectforsecondtechseminar.streamlit.app/

Streamlit page

Woori FISA Fighting ~!

Streamlit page



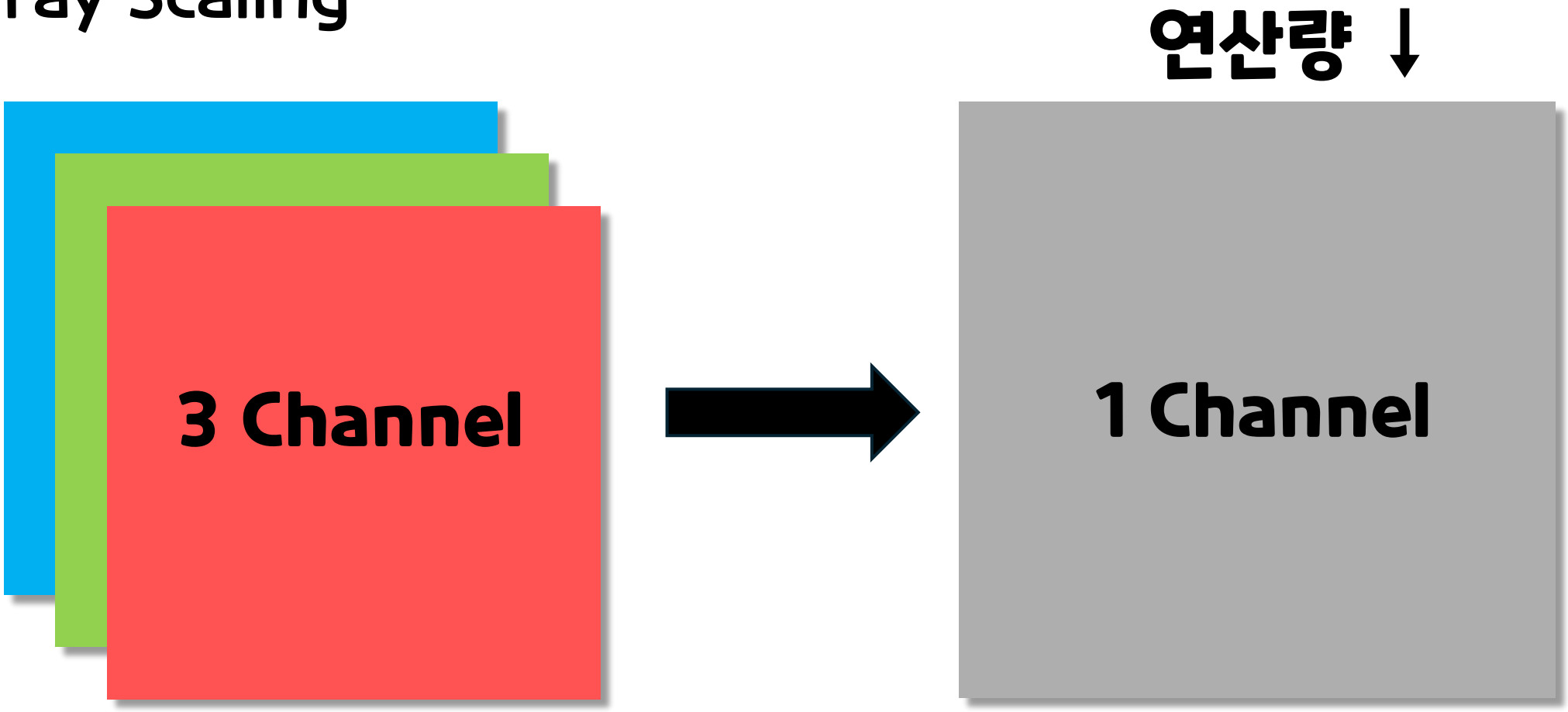
04
결론 및 향후 방향성

이미지 전처리

새로운 알고리즘

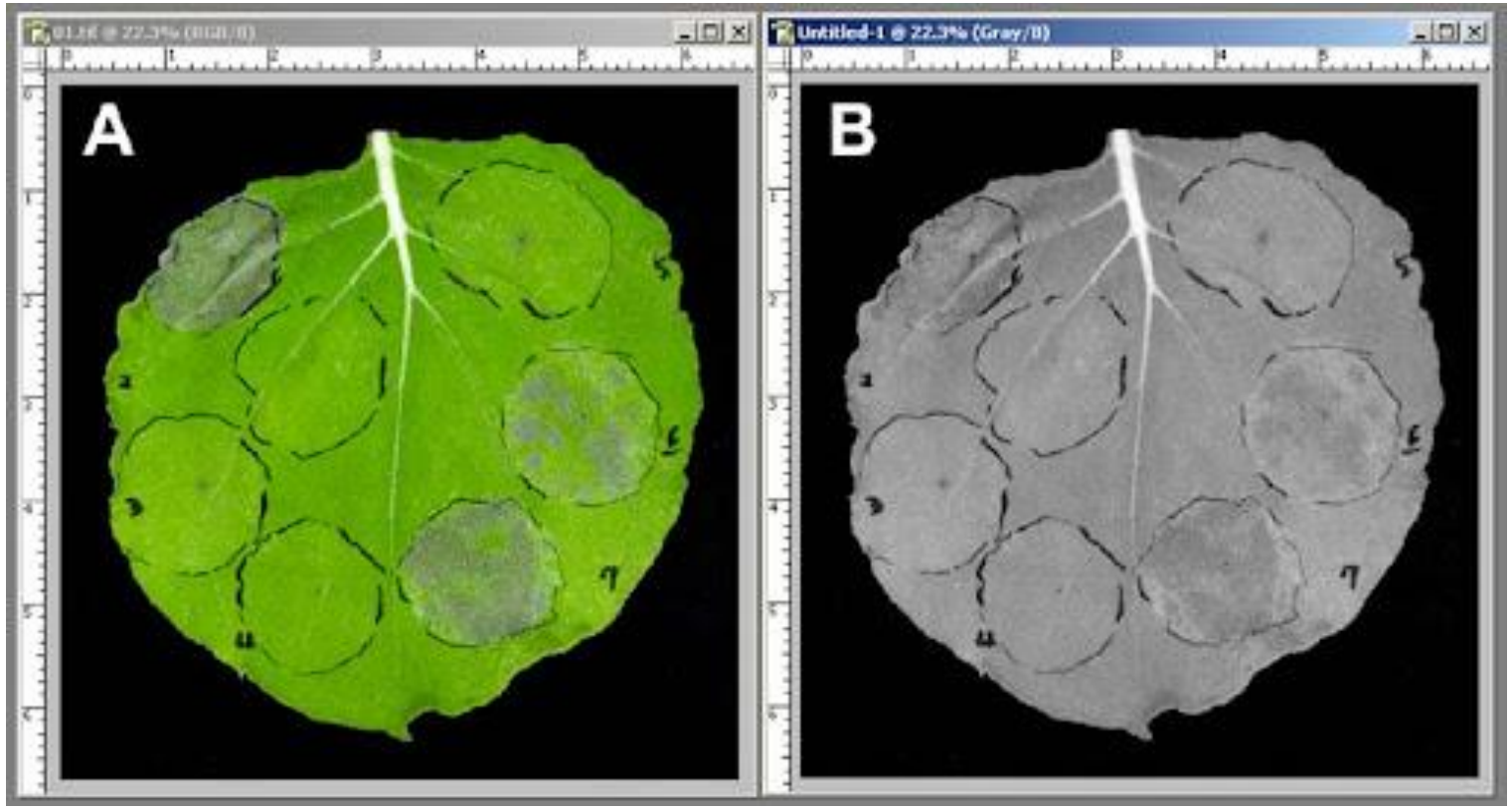
성능 향상을 위한 이미지 전처리

1. Gray Scaling



성능 향상을 위한 이미지 전처리

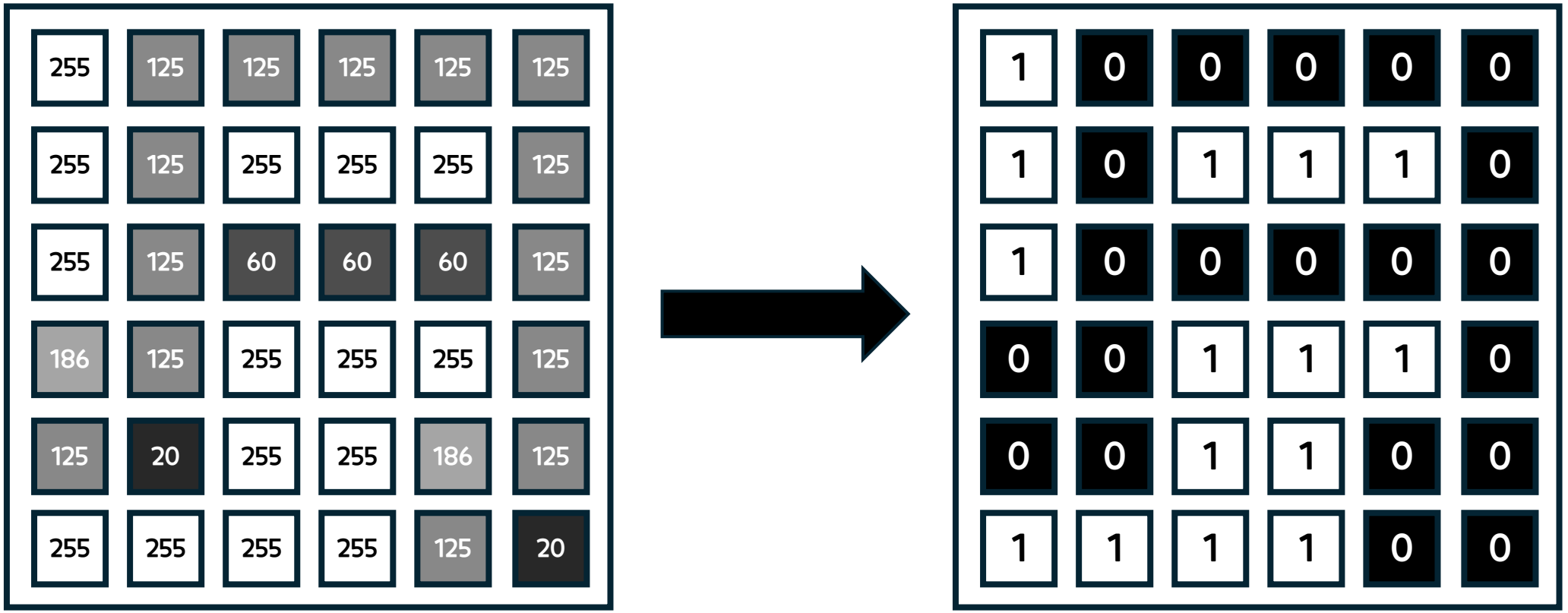
1. Gray Scaling



성능 향상을 위한 이미지 전처리

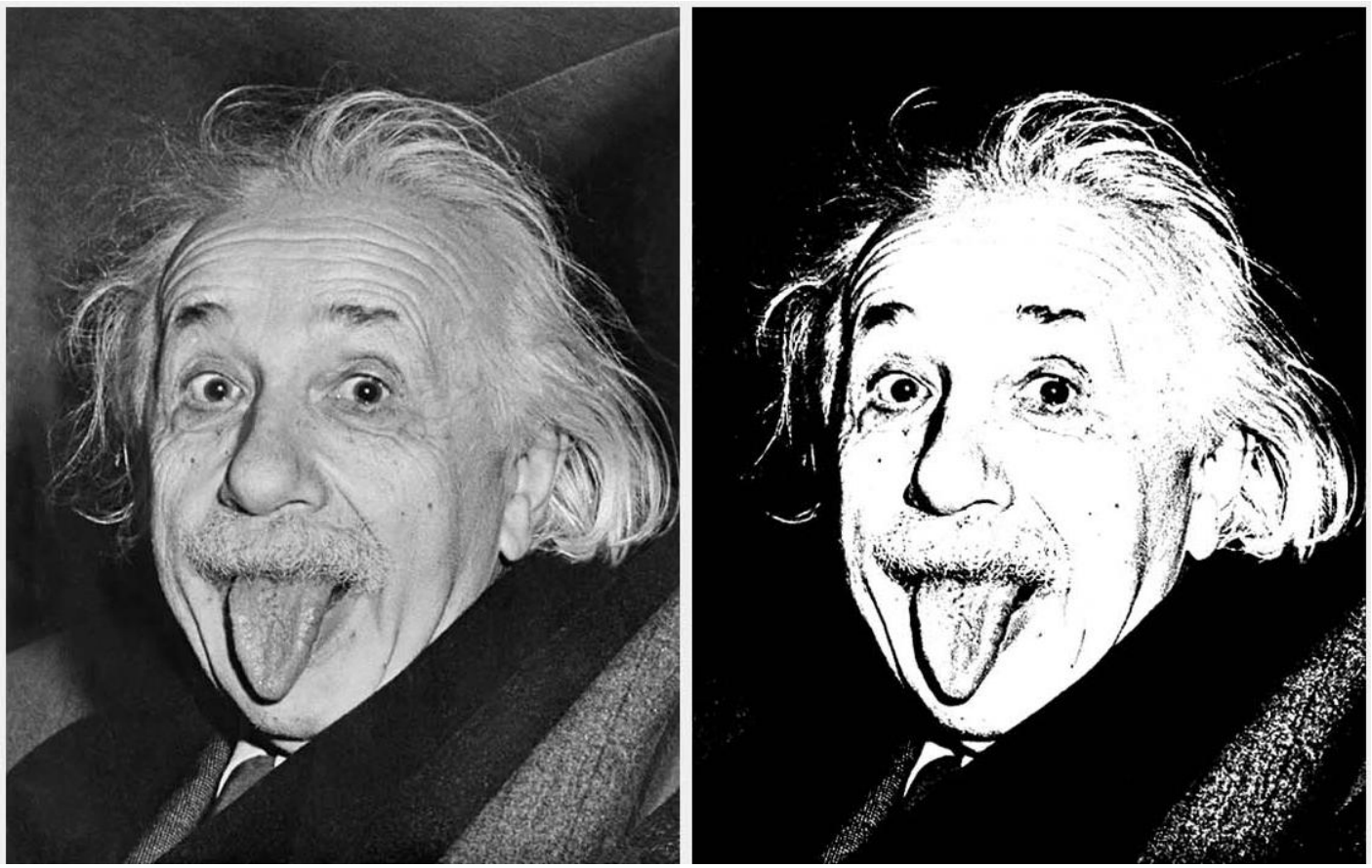
2. Binarization

연산량 ↓ ↓



성능 향상을 위한 이미지 전처리

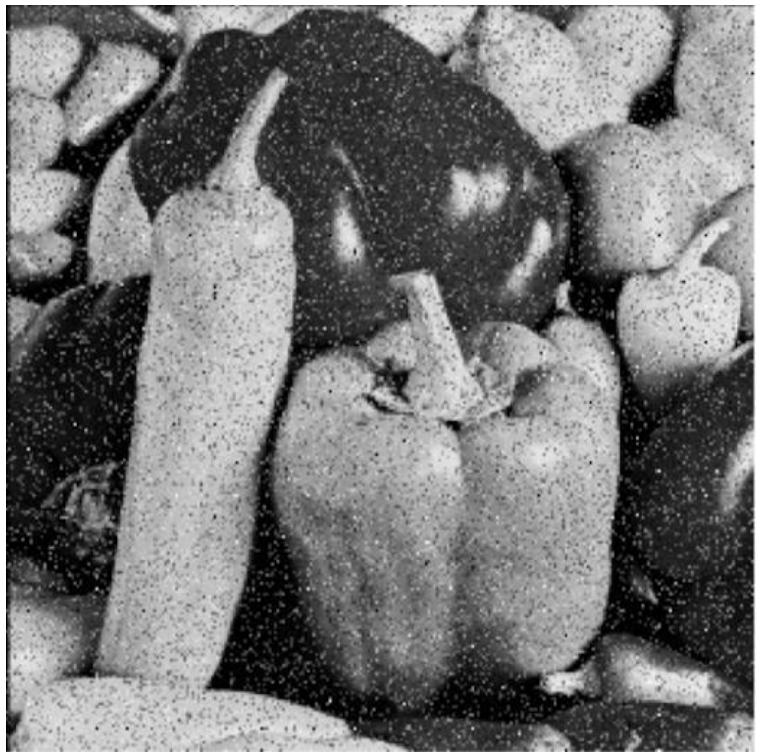
2. Binarization



https://analyticsindiamag.com/ai-trends/how-differential-binarization-can-improve-real-time-scene-text-detection/

성능 향상을 위한 이미지 전처리

3. DeNoising



Noise

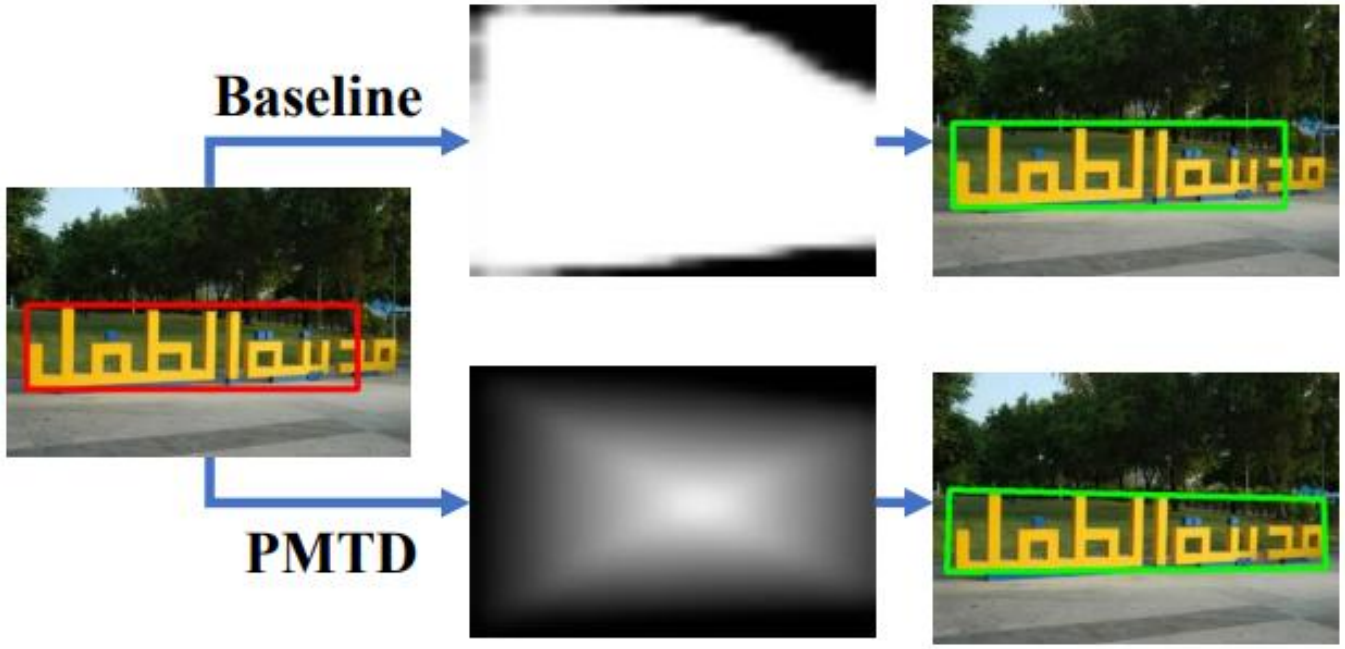
Denoise

이미지
전처리

새로운
알고리즘

Pyramid Mask Text Detector

Pyramid Mask Text Detector



(b) The red box is the predicted bounding box and the green box refers to the predicted text box. The existing Mask R-CNN based methods suffer from the errors of bounding box detection while PMTD can regress more accurate text box with the help of the informative soft text mask.

Pyramid Mask Text Detector

Pyramid Mask Text Detector

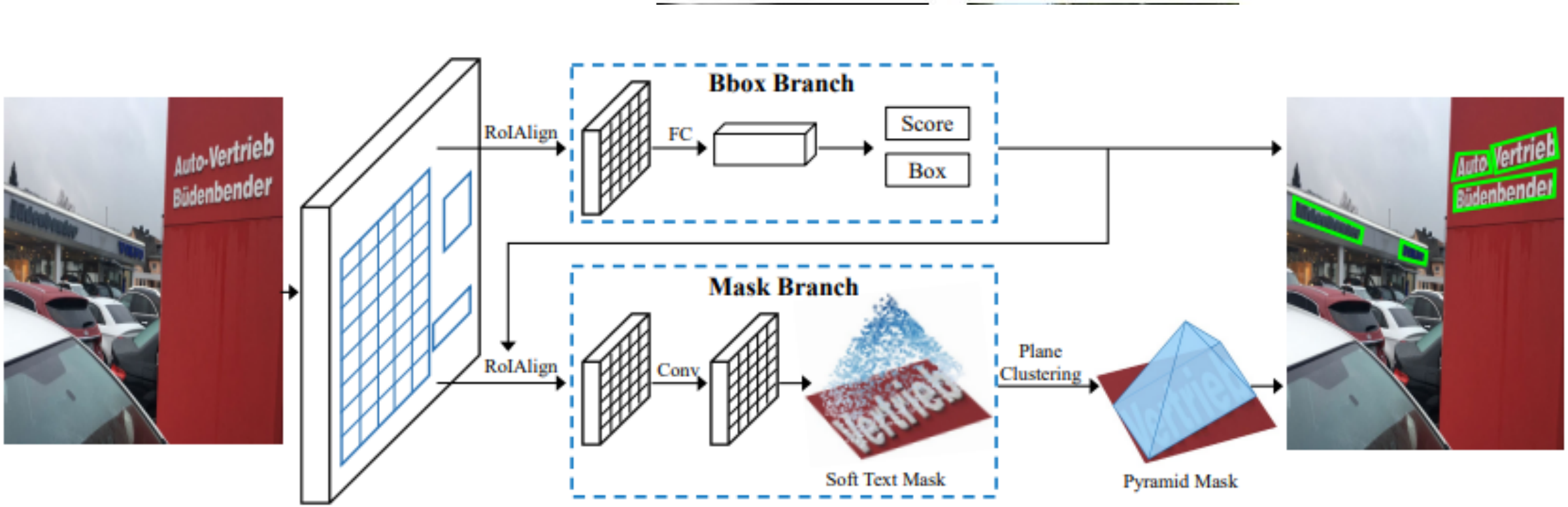


Figure 3: Overall architecture of PMTD.
accurate text box with the help of the informative soft text mask.

Pyramid

Pyramid Mask Text Detector

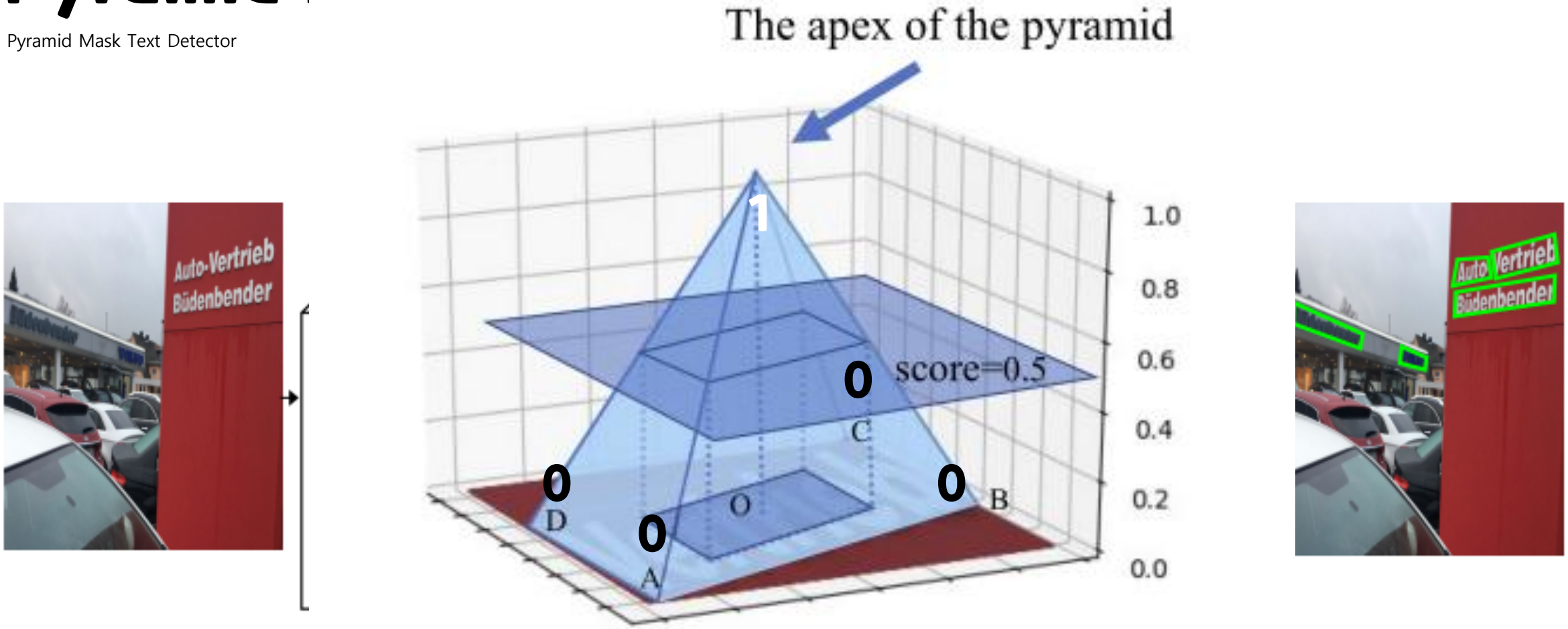
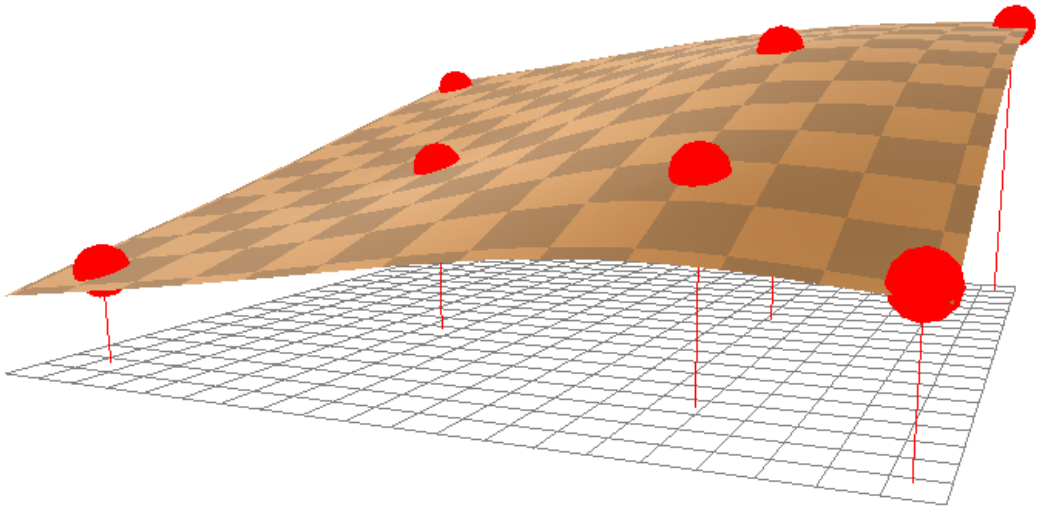
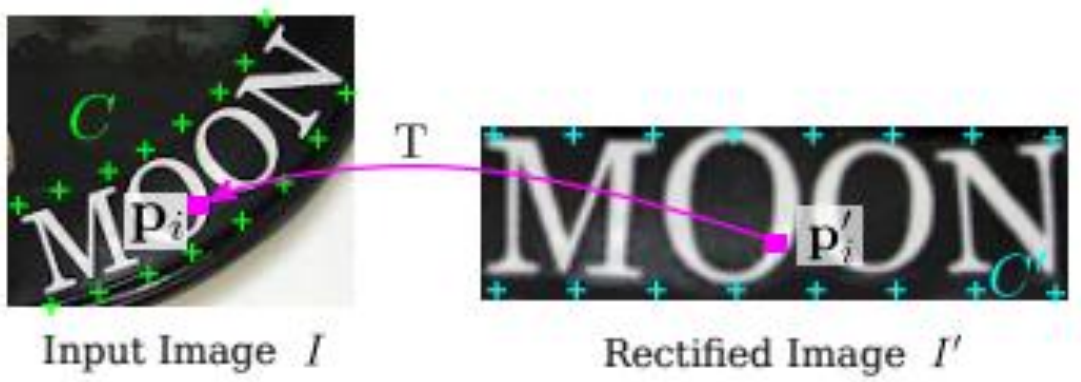


Figure 4: Generation of soft pyramid label. For a pixel in the text area, its label is the height of the pyramid.

TPS(Thin Plate Spline) transformation

Robust Scene Text Recognition with Automatic Rectification





휘어진 이미지를 평면 투영

TPS(Thin Plate Spline) transformation

Robust Scene Text Recognition with Automatic Rectification

휘어진 이미지를 평면 투영

느낀점

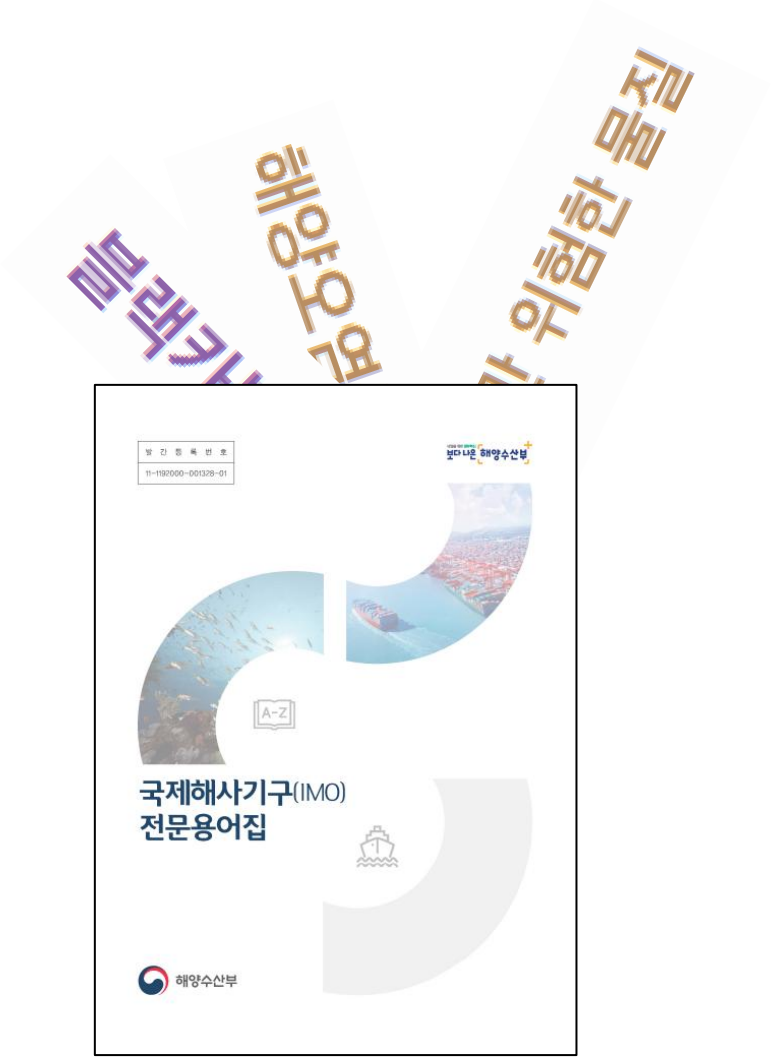
이점

- 이미지에서 텍스트를 추출함으로 번거롭게 옮기는 작업을 줄일 수 있다
- 다양한 알고리즘들을 배울 수 있어서 좋았다

단점

- 원하는 대로 분리가 잘 안되는 경우가 빈번하다
- 시간이 오래 걸린다
- 이미지 자료형에만 한정된다

이후 수집한 데이터는...



SLM

참고자료

참고 사이트

https://healthyinsight.co.kr/%EC%9D%B4%EB%AF%B8%EC%A7%80-%EB%B6%84%EB%A5%98-cnn/
https://velog.io/@xpelqpdj0422/11.-OCR-%EA%B8%B0%EC%88%A0%EC%9D%98-%EA%B0%9C%EC%9A%94
https://github.com/tesseract-ocr/tesseract
https://www.klippa.com/en/blog/information/tesseract-ocr/
https://github.com/JaidedAI/EasyOCR
https://docs.unstructured.io/examplecode/codesamples/apioss/table-extraction-from-pdf
https://velog.io/@smile_b/Grayscaling-Images
https://analyticsindiamag.com/ai-trends/how-differential-binarization-can-improve-real-time-scene-text-detection/
https://canvas4sh.tistory.com/334

논문

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition
Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks
Character Region Awareness for Text Detection
Pyramid Mask Text Detector
Robust Scene Text Recognition with Automatic Rectification

감사합니다 :>