Handwriting OCR
세부 개발 요청사항 정리

테라젠이텍스
임상빅데이터팀
김현민

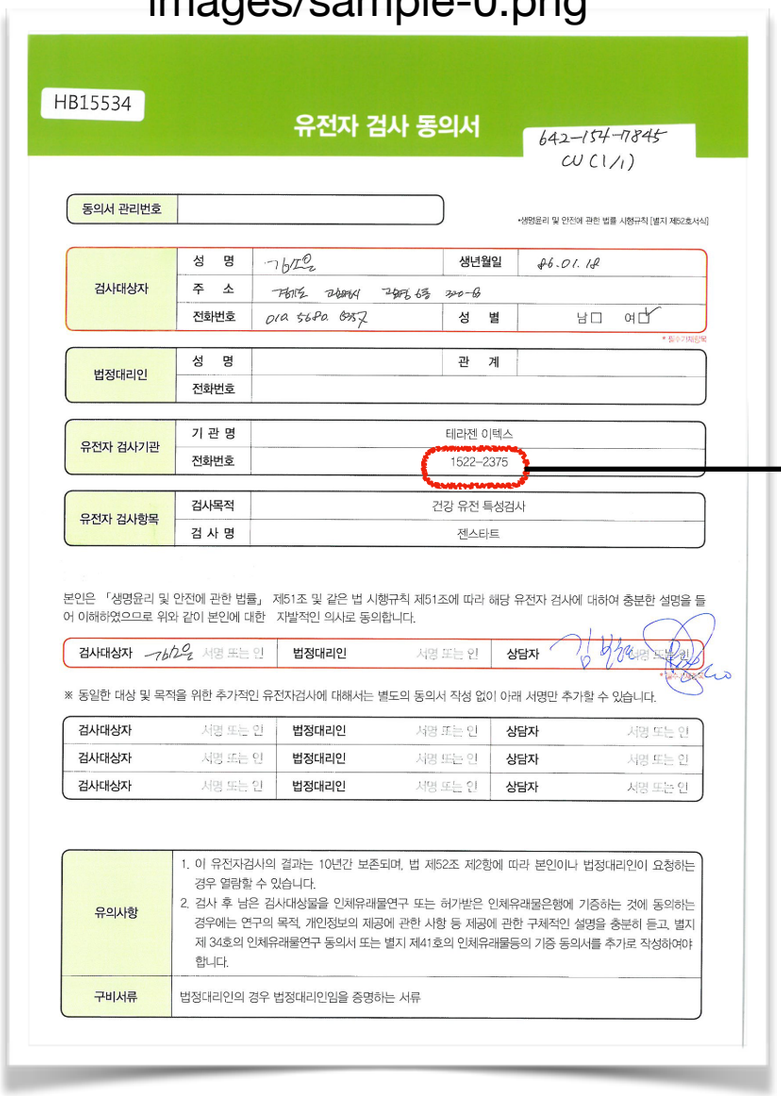
OpenCV

- Along with well-established companies like **Google**, **Yahoo, Microsoft, Intel, IBM, Sony, Honda, Toyota** that employ the library.

- It has C++, Python, Java and MATLAB interfaces and supports Windows, Linux, Android and Mac OS. OpenCV leans mostly towards **real-time vision applications** and takes advantage of MMX and **SSE instructions** when available. A full-featured **CUDA** and OpenCL interfaces are being actively developed right now.

Tesseract OCR (Google)

images/sample-0.png


HB15534

유전자 검사 동의서

642-154-71845
CU (1/1)

동의서 관리번호				
검사대상자	성 명	강O은	생년월일	96.01.18
	주 소	경기도 과천시 [illegible] 320-8		
	전화번호	010 5680 8972	성 별	남□ 여☑
법정대리인	성 명		관 계	
	전화번호			
유전자 검사기관	기 관 명	테라젠 이텍스		
	전화번호	1522-2375		
유전자 검사항목	검사목적	건강 유전 특성검사		
	검 사 명	젠스타트		

본인은 「생명윤리 및 안전에 관한 법률」 제51조 및 같은 법 시행규칙 제51조에 따라 해당 유전자 검사에 대하여 충분한 설명을 들어 이해하였으므로 위와 같이 본인에 대한 자발적인 의사로 동의합니다.

검사대상자	강O은 서명 또는 인	법정대리인	서명 또는 인	상담자	서명 또는 인

※ 동일한 대상 및 목적을 위한 추가적인 유전자검사에 대해서는 별도의 동의서 작성 없이 아래 서명만 추가할 수 있습니다.

검사대상자	서명 또는 인	법정대리인	서명 또는 인	상담자	서명 또는 인
검사대상자	서명 또는 인	법정대리인	서명 또는 인	상담자	서명 또는 인
검사대상자	서명 또는 인	법정대리인	서명 또는 인	상담자	서명 또는 인

유의사항	1. 이 유전자검사의 결과는 10년간 보존되며, 법 제52조 제2항에 따라 본인이나 법정대리인이 요청하는 경우 열람할 수 있습니다. 2. 검사 후 남은 검사대상물을 인체유래물연구 또는 허가받은 인체유래물은행에 기증하는 것에 동의하는 경우에는 연구의 목적, 개인정보의 제공에 관한 사항 등 제공에 관한 구체적인 설명을 충분히 듣고, 별지 제 34호의 인체유래물연구 동의서 또는 별지 제41호의 인체유래물등의 기증 동의서를 추가로 작성하여야 합니다.
구비서류	법정대리인의 경우 법정대리인임을 증명하는 서류

cropped.jpg

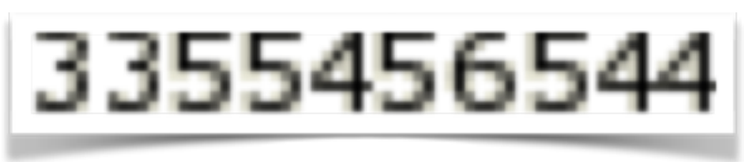

1522-2375

Google Open Source

Tesseract OCR

github.com/tesseract-ocr/tesseract

```
$ tesseract -l eng cropped.jpg output
$ cat output.txt
1522?2375
```


3355456544

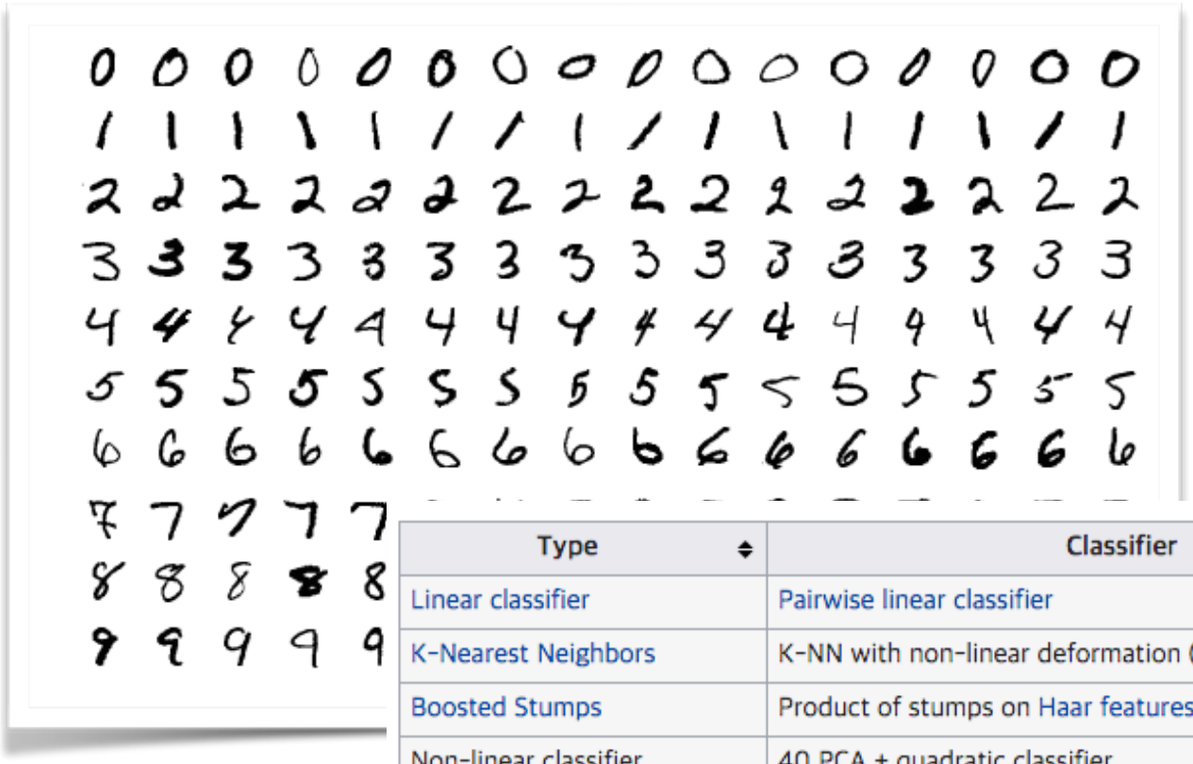
Tesseract OCR

python™

3355456544

손글씨 숫자 알고리즘

Sample images from MNIST test dataset.



- 다양한 숫자 인식 모델이 나와있으며 각 모델별로 에러율이 다양함.
- 즉, 최선의 답은 없음.

Type	Classifier	Distortion	Preprocessing	Error rate (%)
Linear classifier	Pairwise linear classifier	None	Deskewing	7.6[9]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	None	Shiftable edges	0.52[18]
Boosted Stumps	Product of stumps on Haar features	None	Haar features	0.87[19]
Non-linear classifier	40 PCA + quadratic classifier	None	None	3.3[9]
Support vector machine	Virtual SVM, deg-9 poly, 2-pixel jittered	None	Deskewing	0.56[20]
Neural network	2-layer 784-800-10	None	None	1.6[21]
Neural network	2-layer 784-800-10	elastic distortions	None	0.7[21]
Deep neural network	6-layer 784-2500-2000-1500-1000-500-10	elastic distortions	None	0.35[22]
Convolutional neural network	6-layer 784-40-80-500-1000-2000-10	None	Expansion of the training data	0.31[15]
Convolutional neural network	6-layer 784-50-100-500-1000-10-10	None	Expansion of the training data	0.27[16]
Convolutional neural network	Committee of 35 CNNs, 1-20-P-40-P-150-10	elastic distortions	Width normalizations	0.23[8]
Convolutional neural network	Committee of 5 CNNs, 6-layer 784-50-100-500-1000-10-10	None	Expansion of the training data	0.21[17]

This is a table of some of the machine learning methods used on the database and their error rates, by type of classifier:

https://en.wikipedia.org/wiki/MNIST_database

Specification from PGS 운영팀

OCR 프로그램 요구사항 정의서

1. 일반용지 내의 바코드 또는 QR코드를 인식하여 자동으로 데이터화 되어야함.
   - 관리를 위해 자체적으로 설문지에 바코드를 삽입함 => 프로그램 인식시 해당 바코드 데이터화 필요

2. 여러개의 설문지 양식을 셋팅 후 원하는 설문지 양식을 인식 할 수 있어야 함.
   - A사, B사의 설문지 양식이 틀리기 때문에 각각의 설문지 양식을 자동으로 인식하여 데이터화 필요.
   - 단면, 양면에 따라 자동으로 인식 필요

3. 프로그램에서 CSV, 엑셀 포멧의 파일로 저장 가능 해야 함.

4. 스켄된 이미지의 파일명 규칙 설정/변경 가능해야함
   ex) “바코드번호.pdf”

5. 손글씨(필기체) 중 숫자 인식 정합율이 90% 이상이어야 함.

6. 판독불가 또는 판독 정확성이 떨어지는 경우에는 해당 부분이 화면에 표시가 되고 직접 교정할 수 있어야 함

7. 각종 동의서 및 설문지 서식을 직접 등록 가능 해야함.

8. 각종 동의서 및 설문지의 서명 여부를 판독 가능해야함.

9. 스켄된 이미지의 저장경로를 설정 가능해야함.

10. 스켄된 이미지의 암호화 처리 설정가능해야함.

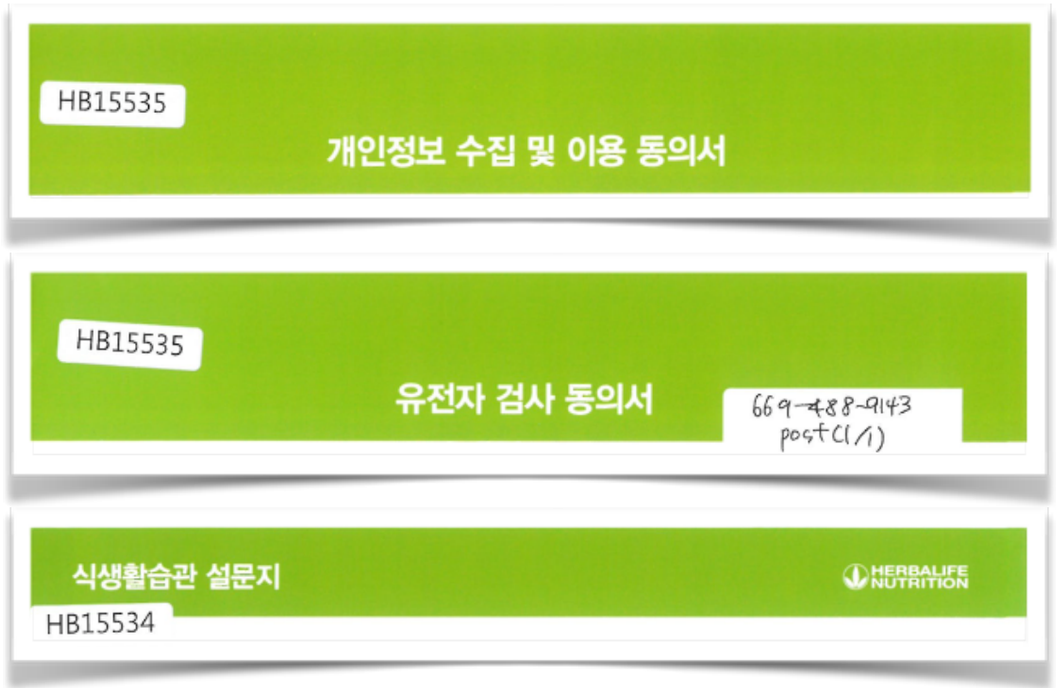
개발 소요 탐색

No.	내용	개발 아이디어 / Comment	개발 우선 순위 (난이도 고려)
1	일반용지 내의 바코드 또는 QR코드를 인식하여 자동으로 데이터화 되어야함. - 관리를 위해 자체적으로 설문지에 바코드를 삽입함 => 프로그램 인식시 해당 바코드 데이터화 필요	QR 코드 관련 개발 별도 조사 단계 필요	후 순위
2	여러개의 설문지 양식을 셋팅 후 원하는 설문지 양식을 인식 할 수 있어야 함. - A사, B사의 설문지 양식이 틀리기 때문에 각각의 설문지 양식을 자동으로 인식하여 데이터화 필요. - 단면, 양면에 따라 자동으로 인식 필요	• 문서 분류를 위해 한글 인식 기술이 필요 혹은 • 설문지 양식 변경을통해 숫자 Labeling 요구됨. • 1번 내용과 통합하여 OCR로 해결 가능	선 순위
3	프로그램에서 CSV, 엑셀 포멧의 파일로 저장 가능 해야 함.	No problem.	선 순위
4	스켄된 이미지의 파일명 규칙 설정/변경 가능해야함 ex) “바코드번호.pdf”	No problem.	선 순위
5	손글씨(필기체) 중 숫자 인식 정합율이 90% 이상이어야 함.	• 손글씨 숫자는 이미 알고리즘 존재 • 이미지 프로세싱 단계가 소요많이 될 것으로 예상 • 에러율은 연구된 알고리즘 기준 이미	선 순위
6	판독불가 또는 판독 정확성이 떨어지는 경우에는 해당 부분이 화면에 표시가 되고 직접 교정할 수 있어야 함	소프트웨어 개발 측면임	후 순위
7	각종 동의서 및 설문지 서식을 직접 등록 가능 해야함.	소프트웨어 개발 측면임	후 순위
8	각종 동의서 및 설문지의 서명 여부를 판독 가능해야함.	인식 기술의 문제로 판별 가능할 것으로 판단됨. (머신러닝알고리즘)	중 순위
9	스켄된 이미지의 저장경로를 설정 가능해야함.	소프트웨어 개발 측면임	후 순위
10	스켄된 이미지의 암호화 처리 설정가능해야함.	소프트웨어 개발 측면임. 별도 조사가 더 필요할 수 있음.	후 순위

*DB화는 각 항목 구현 가능시 추후 가능한 부분이라 판단되며 위 Comment 및 개발 순위에서는 고려하지 않음.

요구사항 #2

- 문서의 카테고리화



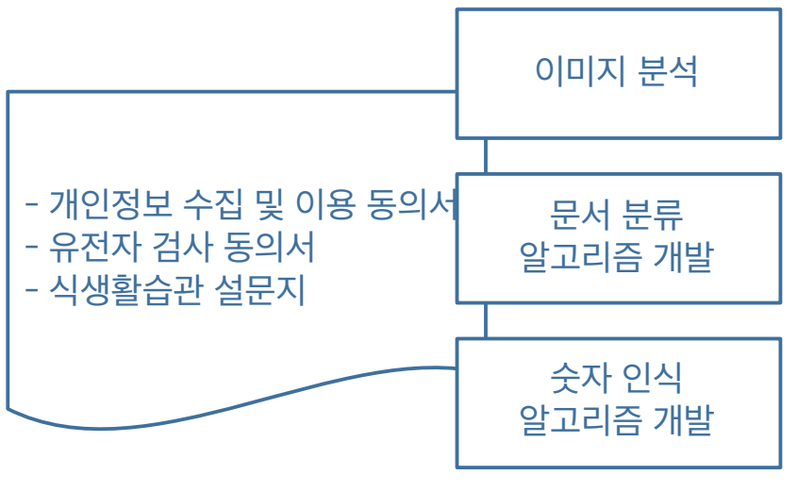
No.	방식	Solution 예제
1	설문지 변경	• 각 문서에 Page 번호 Labeling하여 구분 • Barcode 사용하여 구분 • 그외 다양한 아이디어 존재
2	개발 소요	• 이미지 분류방식을 통해 구분 (머신러닝) • 한글 인식을 통해 구분 • 그외 다양한 아이디어 존재
3		any idea..?

Next action

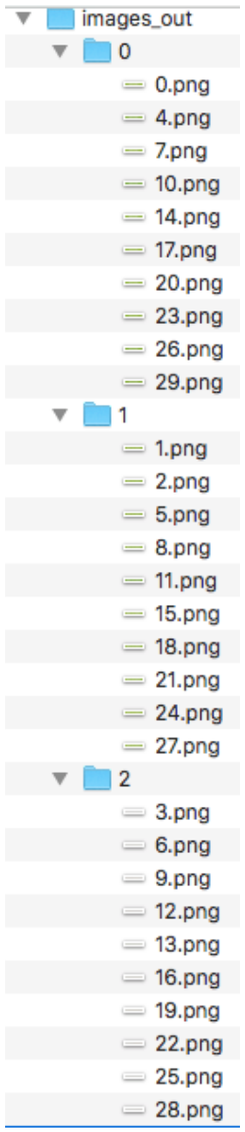
1. 문서 분류 알고리즘 개발

2. 숫자로 작성된 손글씨 인식 알고리즘 개발

3. 개발 소요 탐색에 대한 comment 확정 후 개발 기간 산정 및 검토 (with PGS팀)



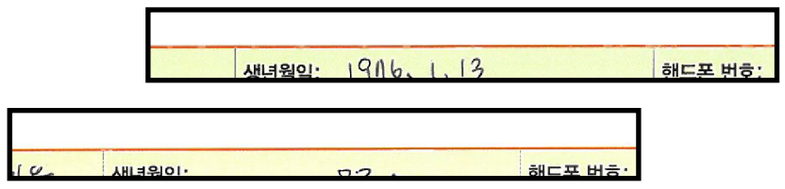
알고리즘 개발 사항



Label: 0

Label: 1

Label: 2



* 각 라벨별로 10개씩의 이미지 데이터가 존재함.

* 이 데이터로 이미지 인식 머신러닝 알고리즘 구현 (코드 개발) 요청

References

- https://opencv.org/about.html
- https://github.com/openpaperwork/pyocr
- https://ko.wikipedia.org/wiki/OpenCV
- http://opencv-python.readthedocs.io/en/latest/