OCR 기반 화장품 성분표 분석 시스템

*강범진 *육찬기 *이진영 *오혜빈 *이의진

서울과학기술대학교

*qjawls1399@seoultech.ac.kr *yookcg2222@seoultech.ac.kr *jin9497young@seoultech.ac.kr
*sew538@seoultech.ac.kr[1] *yeejinlee@seoultech.ac.kr

OCR-based Cosmetics Ingredients Labeling Analysis System

*Beom-jin Kang *Chan-gi Yook *Jin-yeong Lee *Hye-been Oh *Yeejin Lee

Seoul National University of Science and Technology

요약

본 논문에서는 화장품의 효율적 구매를 위한 화장품 성분표를 분석하고 정보를 전달하는 기능의 시스템을 제안한다. 이 시스템에서는 화장품 성분표에 최적화시킨 OCR (Optical Character Recognition) 모델을 사용해 화장품 성분표를 촬영한 영상에서 인식한 문자 데이터를 추출한다. 이 문자 데이터를 통해 얻은 화장품 성분이 사용자 피부 유형에 적합한지 구축된 데이터베이스와의 비교를 통해 소비자에게 최종 전달된다. 200개의 화장품 성분표 영상을 사용해 제안하는 화장품 성분표 분석 모델의 성능을 평가한 결과 80.348%의 정확도를 보였다.

1. 서론

화장품 구매에 대한 소비자의 수요가 점점 늘어나고 있으며, 상품의 종류도 갈수록 다양해지고 있다. 따라서 개인의 피부에 적합한 화장품을 선택하는 것이 더 중요해졌고, 이에 따라 화장품 성분표를 분석하고 소비자의 피부 유형에 따라 적합한지 판단해주는 서비스의 필요성이 대두되었다.

화장품 성분표 분석 서비스를 구현하기 위해서는 먼저 성분표의 글자 인식을 위해 정확도가 높은 OCR (Optical Character Recognition) 모델, 즉 광학 문자 인식 모델이 필요하다. 이는 성분표의 글자 인식이 제대로 이루어지지 않는다면 구축된 성분 데이터베이스와의 비교가 정확하게 이루어지지 않아 사용자에게 정확한 정보가 전달되지 않기 때문이다.

기존 OCR 모델의 예로는 다양한 운영 체제에서 사용 가능한 광학 문자 인식 엔진 Tesseract[1]가 존재한다. 하지만 화장품 성분표 분석 서비스를 목적으로 하지 않아 화장품 성분표에서는 성능이 다소 떨어지는데, 그 이유는 성분표의 글자가 매우 작고 맞춤형 데이터셋이 존재하지 않기 때문이다. 따라서 본 논문에서는 화장품 성분표 분석만을 위한 맞춤 데이터셋을 제작하고 YOLOv5[2]와 SRN (Sequence Recurrent Network)[3]을 결합한 새로운 모델의 구현을 통해 화장품 성분표 분석에 특화된 OCR 모델을 제안한다.

2. 시스템 구조

본 논문에서 제안하는 화장품 성분 분석 서비스는 그림 1과 같이 글자 검출(text detection), 글자 인식(text recognition)의 2가지 모듈로 구성된다. 입력으로 성분표 영상을 넣으면 글자 검출 모델에서 단어 이미지를 검출해 바운딩 박스(bounding box)로 표시한다. 검출해낸 단어 단위의 바운딩 박스를 입력 영상에서 추출하여 글자 인식 모델의 입력으로 하면 글자를 인식해 출력한다.

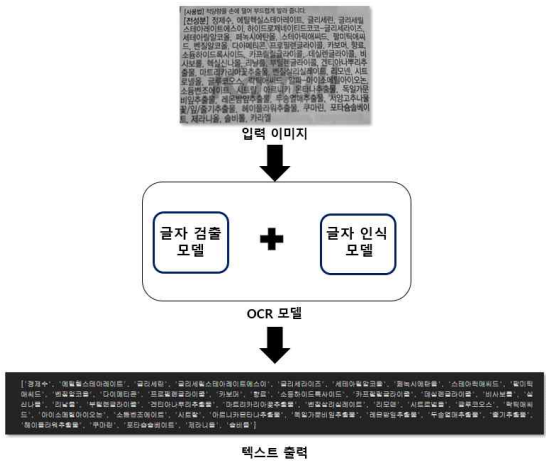

그림 1. 전체 시스템 블록도

[1] These authors contributed equally to this work (corresponding author: Yeejin Lee).

2.1 글자 검출 모델

객체 탐지 모델에는 YOLO (You Only Look Once)[2], EAST (Efficient and Accurate Scene Text detector)[4], EfficientDet[5] 등이 있다. 그 중 YOLO는 CNN (Convolution Neural Network) 기반 오픈 소스 객체 탐지 모델로, 빠른 속도와 높은 정확도로 주목받고 있다. 또한 원하는 데이터를 사용해 학습 및 평가가 가능해 글자 검출에 쉽게 적용 가능하다. 그림 2는 YOLOv5 모델들과 EfficientDet의 성능을 분석한 그래프이다. 분석 결과 YOLOv5s의 수행 속도가 EfficientDet보다 빠르지만 성능 차이는 크지 않다. 또한, 단순 일반적인 문자열로 구성된 문서 형태에 대해서는 YOLO와 EAST가 동일한 성능을 보이지만 생활 속 이미지 문자열에서는 YOLO가 더 나은 성능을 보여준다[6]. 따라서 본 실험에서는 모델 성능과 복잡도를 고려해 YOLO 모델을 객체 탐지 모듈에 사용했으며, 데이터셋의 크기와 학습 환경 등을 고려해 YOLOv5 중 성능이 좋으면서도 경량화된 YOLOv5s 모델을 사용했다. 이와 더불어 한글 단어 검출 성능을 향상시키기 위하여 그림 3과 같이 맞춤 데이터셋을 획득하고 라벨링 작업을 통해 구축하여 학습과 평가에 사용했다.

성분표에는 검출해야 할 객체가 여러 개이고 인접해 있기 때문에 검출 결과 여러 개의 바운딩 박스가 검출된다. 이 과정에서 바운딩 박스가 겹쳐서 출력되는 문제가 생길 수 있다. 이는 같은 단어를 여러 번 출력하는 결과로 이어지므로, 단어 한 개에 바운딩 박스가 한 개만 검출되도록 할 필요가 있다. 본 실험에서는 이 문제를 해결하기 위해 기준 IoU-Threshold 파라미터 이상의 바운딩 박스를 제거하는 NMS (Non-Maximum Suppression) 기법[7]을 사용했다. IoU 값은 검출한 바운딩 박스가 실제 정답 바운딩 박스와 겹치는 비율을 의미한다. 이 기법에서는 IoU-Threshold를 낮출수록 겹치는 영역에서의 바운딩 박스를 제거하는 임계값(threshold)이 낮아지기 때문에 결과적으로 가장 높은 IoU(Intersection over Union)값을 갖는 바운딩 박스를 제외하고 겹치는 바운딩 박스를 제거할 수 있다. 그림 4.(b)와 그림 4.(c)는 각각 IoU-Threshold가 0.9, 0.2일 때의 검출 결과를 보여준다.

2.2 글자 인식 모델

본 연구에서는 SRN (Sequence Recognition Network)과 JAMOEmbedding 방식을 결합해 글자 인식 모듈을 구성했다. 본 논문에서 사용한 글자 인식 모델은 CRNN (Convolutional Recurrent Neural Network), Seq2Seq (Sequence to Sequence), Attention Mechanism을 결합하여 구성하였다[12]. 그림 5와 같이, 입력 영상의 특징은 컨볼루션 층을 사용하여 추출하며, 추출된 특징 맵을 인코더 (Encoder)의 입력 형태로 변환하는 시퀀스 매핑 과정이 수행한다. 위와 같은 과정을 통해 추출된 시퀀스는 인코더로 입력되어 BLSTM (Bidirectional Long Short Term Memory) 유닛간 상호작용에 의해 히든 스테이트(hidden state)로 변환된다[3,9,11].

그림 6에서 보는 바와 같이, 인코더를 통해 얻은 히든 스테이트와 디코더(decoder)의 GRU (Gate Recurrent Unit)을 통해 추출된 스테이트를 기반으로 어텐션 층에서 히든 스테이트와 디코더 스테이트간 유사도 점수를 계산한다. 이 과정을 통해 높은 점수를 가지는 스테이트는 레이블과 유사한 정보를 가진 스테이트이며, 가장 유사한 스테이트를 문맥 벡터(context vector)로 추출한다. 유사도는 디코더 스테이트 벡터 $(s_1, s_2, \dots s_k)$와 히든 스테이트 벡터간 내적으로 정의하며, 추출된 문맥 벡터 c_k는 아래 식(1)과 같이 표현된다.

$$c_k = \max(s_k \cdot h_1, s_k \cdot h_2 \dots, s_k \cdot h_m) \quad (1)$$

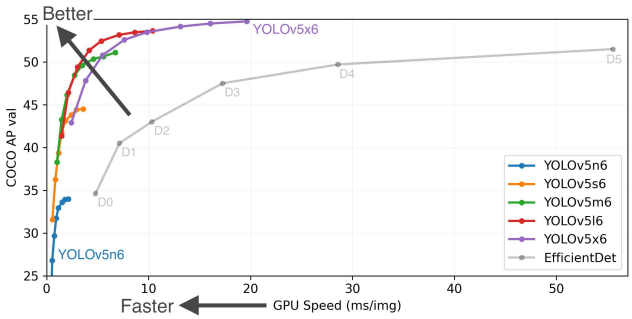

그림 2. YOLOv5 모델들과 EfficientDet의 성능 분석[2]

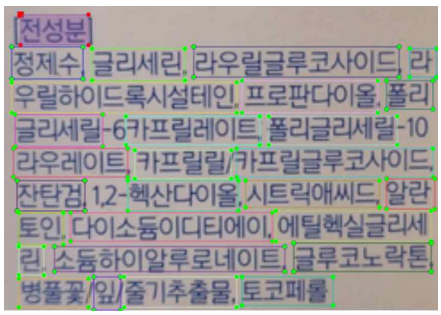

그림 3. 맞춤 데이터셋

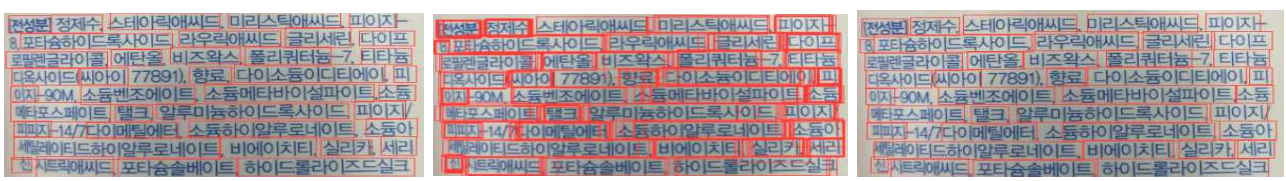

그림 4. 평가 데이터셋 (a) 평가 영상 (b) 임계값(threshold) = 0.5 (c) 임계값(threshold) = 0.2

식 (1)에서 k는 디코더의 GRU의 개수로 디코더 스테이트의 크기이며, m은 인코더의 BLSTM 유닛의 개수로 히든 스테이트의 크기이다. 유사도 계산과정을 통해 획득한 문맥 벡터는 분류기를 통해 유니코드 숫자로 변환되며, 변환된 숫자는 다시 한글로 변환되어 추출된다. 출력된 문맥 벡터는 초성/중성/종성 샘플이 오직 하나의 클래스에 속하며, 정수형 분류과정을 거치기 때문에 희소 교차 엔트로피 손실함수(Sparse_Categorical_Cross_entropy)를 사용하여 모델을 학습하였다. 자모(초성/중성/종성) 스테이트가 분리되는 자모임베딩 기법 특성으로 인해, 자모마다 손실값 계산과정이 필요하며, 손실값은 식 (2)와 같이 정의한다[13].

$$J = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})] \qquad (2)$$

식 (2)의 N은 문맥 벡터의 크기, y_i는 정답 레이블, $\hat{y_i}$는 모델의 예측 레이블이며 자모마다 손실함수 값의 합이 모델의 손실값이 된다.

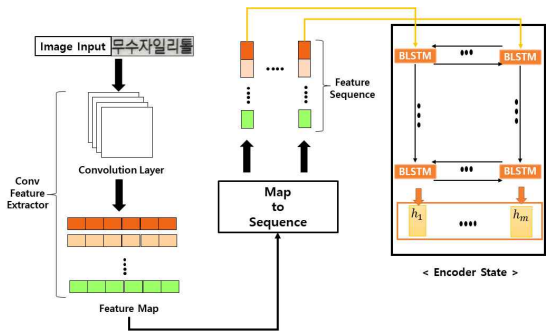

그림 5. 글자 특징 추출 및 벡터화 [14]

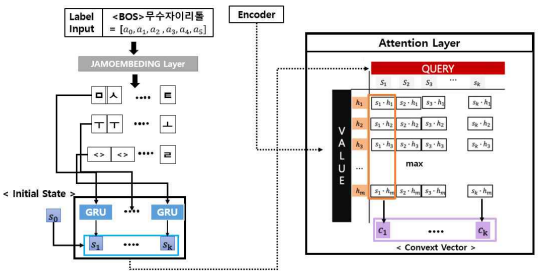

그림 6. 어텐션 메커니즘을 이용한 벡터추출 [15]

3. 실험 및 결과

제안하는 글자 검출 모델 성능 평가에는 mAP (mean Average Performance)를 사용하여 수행하였다[16]. 맞춤 데이터셋의 모든 바운딩 박스에 대한 AP (Average Precision)의 평균을 식 (3)과 같이 하며, AP는 정밀도(Precision)와 재현율(Recall)에 의해서 결정된다. 정밀도란 모델이 예측한 바운딩 박스 중 정확하게 예측한 바운딩 박스의 비율을 의미한다. 재현율은 정확한 바운딩 박스(ground-truth bounding box) 중 정확하게 예측된 바운딩 박스의 비율을 의미한다. 재현율과 정밀도를 이용한 정밀도-재현율 그래프에서의 선 아래의 면적으로 AP를 계산한다. 모든 재현율 레벨에서의 정밀도의 평균을 계산하여 P-R 곡선을 요약하는 방식이다. mAP는 하나의 클래스 마다 계산된 AP의 값을 전체 클래스 개수에 대해 AP를 계산해 평균을 낸 값으로, 본 논문에서는 단일 클래스 검출 과정이므로 AP의 평균이 mAP가 된다.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \qquad (3)$$

글자 검출 모델 성능은 학습 데이터셋에 의해서 영향을 크게 받으므로, 화장품 성분표와 유사한 공공행정문서 데이터셋(AI Hub)과 맞춤 식료품 데이터셋(Grocery)을 사용해 학습한 모델과 맞춤 데이터셋(Custom)을 사용하여 학습 모델의 성능을 비교하였다. 표 1은 학습 데이터셋별 성능을 비교한 결과다. 본 연구의 목적에 부합한 맞춤 데이터셋을 이용하였을 때 mAP = 0.957로 가장 좋은 성능을 보인다.

또한, 검출 모델의 성능은 IoU-Threshold에도 영향을 받으므로 서로 다른 임계값을 사용하여 바운딩 박스를 검출한 성능을 비교하여 표 2에 정리하였으며, 임계값이 0.2 일 때 가장 좋은 성능을 보인다. 표 3은 2.1장에서 추출한 단어 단위의 입력을 사용하여 여러 글자 인식 모델을 사용하여 글자를 인식한 결과이다.

데이터셋	정밀도 (%)	재현율 (%)	mAP
Custom	96.7	91.2	0.957
AI Hub	63.2	53.2	0.482
Grocery	66.6	83.2	0.753

표 1. 서로 다른 데이터셋을 사용하여 학습한 모델의 성능 비교

IoU-Threshold	정밀도 (%)	재현율 (%)	mAP
0.2	96.7	91.2	0.957
0.4	97.2	90.2	0.949
0.6	97.5	88.6	0.938
0.8	91.7	86.2	0.913
0.9	60.2	76.7	0.679

표 2. 임계값에 따른 모델의 성능 비교(맞춤 데이터셋 기준)

모델	정밀도 (%)	처리 속도 (ms/image)
Tessearact	75.367	18.9
CRNN	78.027	4.8
SRN	80.348	21.6

표 3. 글자 검출 결과에 대한 글자 인식 모델 성능 비교

4. 결론

본 논문에서는 글자 검출 모델과 글자 인식 모델을 결합한 OCR 모델을 이용하여 화장품 성분표를 분석하고 소비자의 피부 유형에 따라 적합한지 판단해주는 서비스를 제안한다. 본 시스템은 소비자에게 있어서

화장품 구매 시 성분에 대한 정보를 제공함으로써 소비자별 상이한 피부 유형에 맞추어 구매를 유도해 소비자의 효율적인 화장품 구매와 안전한 사용을 가능하게 한다. 서비스에 사용한 OCR 모델 실험 결과, 검출된 글자 영역에 대한 글자 인식의 모델의 정확도는 약 80.348% 이다. 본 논문에서 제안하는 모델은 한글 성분에 초점을 맞추어 개발을 진행하였으나, 추후 특수문자, 영문, 숫자까지 포함한 데이터를 구축하여 학습시킨다면 더 발전된 성능을 가진 시스템을 구현할 수 있을 것으로 예상한다. 또한 화장품 성분표가 아닌 식료품 성분표 혹은 다른 데이터셋에 적합한 데이터를 이용하여 학습을 진행하고 개발한다면 화장품 성분표 뿐만 아니라 식료품 및 다른 성분표 분석에도 모델을 쉽게 적용시킬 수 있을 것으로 기대된다.

참 고 문 헌

[1] R. Smith, "An Overview of the Tesseract OCR Engine," Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 629-633, doi: 10.1109/ICDAR.2007.4376991.

[2] G. Jocher, K. Nishimura, T. Mineeva, R. Vilariño, GitHub repository [Internet], https://github.com/ultralytics/yolov5

[3] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, 1 Nov. 2017, doi: 10.1109/TPAMI.2016.2646371.

[4] Zhou, Xinyu & Yao, Cong & Wen, He & Wang, Yuzhi & Zhou, Shuchang & He, Weiran & Liang, Jiajun. (2017). EAST: An Efficient and Accurate Scene Text Detector.

[5] Mingxing Tan, Ruoming Pang, Quoc V. Le; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10781-10790

[6] C. Y. Park et al., "YOLO, EAST : Comparison of Scene Text Detection Performance, Using a Neural Network Model," KIPS Transactions on Software and Data Engineering, vol. 11, no. 3, pp. 115-124, Mar. 2022.

[7] P. Henderson, V. Ferrari "End-to-end training of object class detectors for mean average precision (ACCV 2016) pp.198-213, Mar.2017.

[8] N. H. Aung, H. Htun. Y. K. Thu, S. S, Maung "CRNN Based OCR for American and British Sign Language Fingerspelling", 2021 13th International Conference on Knowledge and System Engineering(KSE2021), 2021

[9] I. Sutskever, O. Vinyals, Q. V. Le " Sequence to Sequence Learning with Neural Networks", Advances in Neural information Processing System 27(NIPS 2014)

[10] X. He, C. Yao "Scene Text Detection and Recognition", International Journal of Computer Vision, pp.161-182, Aug.2021

[11] Minh-Thang Luong, Hieu Pham, Christopher D. Manning "Effective Approaches to Attention-based Neural Machine Translation", EMNLP 2015, pp.11

[12] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, Xiang Bai, "Robust Scene Text Recognition With Automatic Rectification", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4168-4176

[13] GOWDRA, Nidhi, et al. Maximum Categorical Cross Entropy (MCCE): A noise-robust alternative loss function to mitigate racial bias in Convolutional Neural Networks (CNNs) by reducing overfitting. 2020.

[14] Huang Z, Lin J, Yang H, Wang H, Bai T, Liu Q, Pang Y. "An Algorithm Based on Text Position Correction and Encoder-Decoder Network for Text Recognition in the Scene Image of Visual Sensors." Sensors. 2020; 20(10):2942

[15] Zhaoyang Niu, Guoqiang Zhong, Hui Yu, "A review on the attention mechanism of deep learning", Neurocomputing, Volume 452, 2021, pp. 48-62, ISSN 0925-2312.

[16] Everingham, M., Van Gool, L., Williams, C.K.I. et al, "The PASCAL Visual Object Classes (VOC) Challenge." Int J Comput Vis(IJCV) 88, 303-338, 2010.