키오스크 환경에서 얼굴 인식 기반 자동 언어 선택을 위한 인종 분류 시스템에 관한 연구

장수현, 최윤영, 강지헌
덕성여자대학교
tngus7238@duksung.ac.kr, cyy0509@duksung.ac.kr, jhkang@duksung.ac.kr

A Study on Race Classification System for Automatic Language Selection Based on Face Recognition in Kiosk Environments

SuHyeon Jang, YunYeong Choi, Jiheon Kang
School of Software, Duksung Women's University

요 약

본 논문은 사용자 맞춤형 키오스크 UX(User Experience) 개선에 기여할 수 있는 얼굴 인식 기술의 적용 가능성을 제시한다. 특히 얼굴 인식 기반 인종 분류를 통해 키오스크 사용자의 언어를 자동으로 설정함으로써 다문화, 다인종 사용자 적합한 비대면 인터페이스를 구현하고자 한다. 이를 위해 공공 데이터세트와 자체 수집한 데이터세트를 활용하여 4가지 인종 분류 모델을 비교하고 Fine tuning을 통해 실제 키오스크 환경에서의 적용 가능성을 실험하고 평가하였다.

Ⅰ. 서 론

비대면 서비스가 일상화되면서 키오스크(kiosk)는 다양한 장소에서 필수적인 사용자 인터페이스 도구로 자리잡고 있다. 특히 다문화 환경과 다국적 여행자들이 확산됨에 따라 키오스크 사용자들이 사용하는 언어는 점차 다양해지고 있으며, 이에 따라 언어 자동 전환 기능의 필요성이 대두되고 있다. 기존 언어 선택 시스템은 사용자의 수동 입력에 의존하거나 단순한 위치 기반 설정에 한정되어 있어 개인화된 인터페이스 제공에는 한계가 있다.

본 연구는 키오스크에서 사용자의 얼굴 이미지를 기반으로 인종을 자동 분류하고 해당 인종에 맞는 언어로 인터페이스를 전환하는 인종 기반 언어 자동 변환 시스템을 구현하는 것을 목표로 한다. 핵심 기술은 정확한 인종 분류 모델 구축에 있으며 이를 위해 총 4개의 딥러닝 기반 모델 아키텍처를 구현하여 비교 실험을 수행하였다. 사전학습에는 공공 데이터셋 FairFace[1] 내 4개국 인종 약 5만 장 이미지를 표 1과 같이 추출하여 활용하고, 파인튜닝에는 실제 키오스크 사용 환경을 고려해 수집한 4개 국가의 약 3,700장 이미지로 구성된 자체 데이터셋을 사용하였다. 본 연구는 모델 비교를 통해 실제 환경에서의 적용 가능성과 인종 인식 정확도 향상을 위한 최적 구조를 제안하고자 한다.

Classification	# of dataset		
	Total	Train	Test
East_Asian	13,813	12,265	1,548
Indian	13,824	12,308	1,516
Middle_Eastern	10,408	9,200	1,208
White	18,578	16,498	2,080
Total	56,623		

[표 1] FairFace 기반 4개국 인종 이미지 데이터셋 구성

Ⅱ. 본론

적용 가능성과 인종 인식 정확도 향상을 위한 최적 구조를 제안하고자 한다. 본 연구에서는 한국(Korean), 중국(China), 일본(Japan), 미국(US) 국적의 얼굴 이미지를 웹 크롤링을 통해 직접 그림 1과 같이 수집하여 약 3,700장의 데이터셋을 구축하였으며, 전체 데이터셋의 분포는 표 2와 같다. 수집된 이미지 중 저해상도이거나 얼굴이 제대로 인식되지 않는 샘플은 전처리 과정에서 제거하여 데이터의 품질을 확보하였다.

[그림 1] 각 클래스별 대표 얼굴 이미지 시각화
(Korean, Chinese, Japanese, US)

Classification	# of dataset		
	Total	Train	Test
Korean	969	775	194
Chinese	905	724	181
Japanese	799	639	160
US	1,037	829	208
Total	3,710		

[표 2] 직접 수집한 4개국 인종 이미지 데이터셋 구성

본 연구의 목적은 얼굴 이미지 기반 인종 분류 모델의 정확도를 향상시키는 것이며, 이를 위해 다양한 구조의 딥러닝 기반 모델을 설계하고 그 성능을 비교하였다. 먼저, ArcFace 모델로부터 추출한 512차원의 임베딩 벡터를 이용해 SVM 분류기를 학습시켰다. ArcFace는 주로 개인 식별 목적에 최적화되어 있어 동일인의 다양한 얼굴에서도 일관된 특징을 유지하

는 데 중점을 둔다[2]. 그러나 개인 식별 방식과는 다르게 인종 분류를 위해서는 동일 인종 내 다양한 얼굴 간의 공통점을 강조하고 서로 다른 인종 간의 특징의 구분을 극대화하는 방향으로 학습되어야 하므로 ArcFace 임베딩 단독 사용에는 구조적 한계가 있다. 실제로 korean, china, japan, us 4개 국가의 얼굴 이미지 약 3,700장을 학습시킨 결과 SVM 분류기의 정확도는 74% 수준에 그쳤다. 다음으로 동일한 데이터셋에 대해 EfficientNet-B4 기반의 모델을 학습하였다. 이 CNN 모델은 얼굴 이미지를 입력으로 받아 깊은 계층에서 다양한 해상도와 비율의 시각적 특징을 추출할 수 있는 구조로 인종 분류에서 유의미한 정보를 더 잘 포착하였다. 이를 통해 약 76%의 정확도를 달성하였으며 ArcFace보다 개선된 결과를 보였다. 하지만 여전히 키오스크에 적용하기 위한 일반화 성능에는 한계가 있었다.

개별적 두 모델을 보완하기 위해 ArcFace 임베딩과 EfficientNet 특징 벡터를 융합한 멀티모달 모델을 제안하였다. ArcFace의 512차원 벡터와 EfficientNet의 1792차원 벡터를 병합(concatenate)하여 총 2304차원의 통합 특징 벡터를 구성하고 이를 fully connected layer에 입력하여 인종 분류를 수행하였다. 이 모델은 두 입력의 상호보완적 특성을 활용하여 정확도 86%을 기록하였다. ArcFace가 사람 간의 미세한 차이를 반영하는 데 효과적이라면 EfficientNet은 이미지 자체에서 추출한 시각적 인종 특징을 보완적으로 제공함으로써 더 넓은 범주의 분류 기준을 학습할 수 있었다.

또한, 본 연구에서는 모델의 일반화 능력을 강화하기 위해 학습 시 데이터 증강(data augmentation) 기법을 도입하였으며, 수평 뒤집기(RandomHorizontalFlip), 회전(RandomRotation), 랜덤 크롭(RandomResizedCrop), 밝기 및 채도 변화(ColorJitter)와 같은 시각적 변형을 포함시켜 동일 얼굴 이미지 내 다양한 변화가 포함할 수 있도록 하였다. 이를 통해 실제 상황에서 다양한 조명, 각도, 표정 변화에 대응할 수 있는 견고한 모델을 구축할 수 있었다.

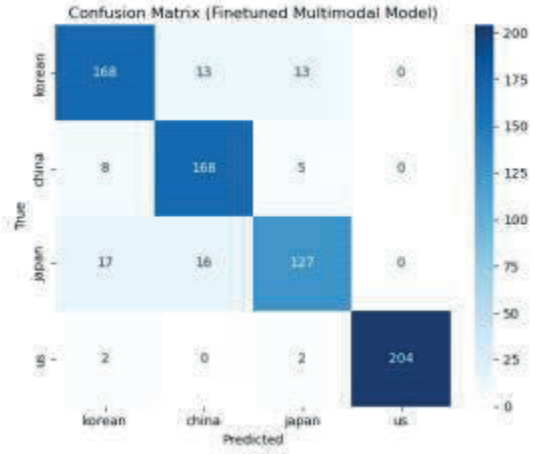
이후 정확도 향상을 위해 사전학습(pretraining) 및 미세조정(fine-tuning) 전략을 적용하였다. FairFace 대규모 데이터셋에서 East_Asian, Indian, Middle_Eastern, White 인종에 해당하는 약 5만장의 이미지를 활용하여 멀티모달 모델을 사전학습하였고 이를 통해 일반적인 인종 특성을 학습한 후 웹 크롤링한 Korean, China, Japan, US 4개 국가의 얼굴 이미지 데이터셋으로 파인튜닝을 수행하였다. ArcFace 임베딩은 그대로 사용하지 않고 별도의 ArcFace 임베딩을 변환하는 projection layer를 통해 인종 분류에 유용한 정보만 가공하도록 설계하였다. 해당 레이어는 512차원의 ArcFace 벡터를 256차원으로 차원 축소 인코딩 후 Batch Normalization과 ReLU를 통해 인종 간 구분을 위한 새로운 표현 공간으로 투영하는 역할을 수행하였다.

모델(Model)	정확도 (Accuracy)
ArcFace SVM	74.0%
EfficientNet-B4	76.0%
EfficientNet+ArcFace(Pretrain only)	86.0%
EfficientNet+ArcFace(Fine-tuning 후)	90.0%

[표 3] 4가지 인종 분류 모델 간 정확도 비교

표 3은 본 연구에서 실험한 네 가지 모델의 성능을 비교한 결과이다. 단일 ArcFace 또는 EfficientNet 기반 모델보다 멀티모달 결합 모델이 높은 성능을 보였으며 특히 사전학습 된 멀티모달 모델은 4개국 데이터에 대해 가장 높은 정확도인 90%를 기록하였다. 이는 이미지와 얼굴 임베딩을 결합한 표현이 인종 분류에 효과적임을 보여준다.

최종적으로, 사전학습 기반의 멀티모달 파인튜닝 모델은 90%의 정확도를 달성하였으며 그림 2의 Confusion Matrix를 통해 각 클래스 간 분류 정확도와 주요 오분류 경향을 직관적으로 확인할 수 있다. 이는 기존 단일 모델 대비 약 15%p 이상의 성능 향상을 의미한다. 특히 다양한 인종 간 차이를 일반화된 표현으로 학습한 사전학습 단계가 파인튜닝 시 적은 수의 샘플에서도 높은 일반화 성능을 확보하는 데 기여했음을 보여준다.



[그림 2] 파인튜닝된 멀티모달 모델의 Confusion Matrix (Korean, China, Japan, US 대상)

Ⅲ. 결론

본 연구에서는 얼굴 이미지를 기반으로 인종을 분류하기 위한 다양한 모델 구조를 실험적으로 비교하였다. 특히 ArcFace의 특징이 인종 분류에는 직접적으로 적합하지 않다는 점을 분석하였고 EfficientNet 기반 시각 정보와의 융합을 통해 보완하는 구조를 제안하였다.

단일 임베딩(SVM 기반 ArcFace)보다 단일 이미지 분류 모델(EfficientNet)의 성능이 더 높았고 두 소스를 결합한 멀티모달 모델은 그보다 높은 성능을 기록하였다. 나아가 FairFace 기반 사전학습과 인종 분류 목적의 파인튜닝을 통해 분류 정확도는 90%로 향상되었으며 이는 멀티모달 사전학습이 인종 간 표현 학습에 효과적임을 보여준다.

제안한 멀티모달 인종 분류 모델은 비대면 서비스 환경에서 사용자의 인종을 정확하게 예측함으로써 키오스크 시스템의 언어 자동 전환 기능에 직접적으로 적용 가능할 것을 기대한다. 더불어, 입력 없이도 개인화된 언어 설정을 제공할 수 있어 다문화 사회에서의 접근성 향상과 서비스 이용 편의성을 크게 높일 수 있다.

참 고 문 헌

[1]Mikaela Karkkainen, Jungseock Joo, “FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation”, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[2]Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, Stefanos Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019