Vision Transformer 기반 얼굴 연령 분류 기법의 성능 분석

박준휘[O], 김남중[*], 박창준[**], 이재현[***], 곽정환[*]
[O]국립한국교통대학교 AI로봇공학과,
[*]국립한국교통대학교 소프트웨어학과,
[**]국립한국교통대학교 교통에너지융합학과,
[***]국립한국교통대학교 컴퓨터공학과
e-mail: objectdetection@kakao.com[O], jgwak@ut.ac.kr[*]

Performance Analysis of Human Facial Age Classification Method Based on Vision Transformer

Junhwi Park[O], Namjung Kim[*], Changjoon Park[**], Jaehyun Lee[***], Jeonghwan Gwak(Corresponding Author)[*]
[O]Dept. of AI·Robotics Engineering, Korea National University of Transportation,
[*]Dept. of Software, Korea National University of Transportation,
[**]Dept. of IT·Energy Convergence, Korea National University of Transportation,
[***]Dept. of Computer Engineering, Korea National University of Transportation

● 요 약 ●

얼굴 연령 분류 기법은 신원 확인 시스템 고도화, 유동 인구 통계 자동화 시스템 구축, 연령 제한 콘텐츠 관리 시스템 고도화 등 다양한 분야에 적용할 수 있는 확장 가능성을 가진다. 넓은 확장 가능성을 가지는 만큼 적용된 시스템의 안정성을 위해서는 얼굴 연령 분류 기법의 높은 정확도는 필수적이다. 따라서, 본 논문에서는 Vision Transformer(ViT) 기반 분류 알고리즘의 얼굴 연령 분류 성능을 비교 분석한다. ViT 기반 분류 알고리즘으로는 최근 널리 사용되고 있는 ViT, Swin Transformer(ST), Neighborhood Attention Transformer(NAT) 세 가지로 선정하였으며, ViT의 얼굴 연령 분류 정확도 65.19%의 성능을 확인하였다.

키워드: 얼굴 연령 분류(Facial Age Classification), Vision Transformer(ViT), 이미지 분류(Image Classification)

I. Introduction

얼굴 연령 분류 기법이란, 사람의 얼굴 이미지를 통해 연령을 예측하는 방법을 의미하며 다음과 같은 넓은 확장성을 가진다. 기존 신분증을 통한 신원 확인 시스템의 경우, 사용자가 신분증의 실제 소유자인지 진위 여부를 고려하지 않고, 신분증의 소지 여부만을 확인하는 맹점이 존재한다. 연령 제한 콘텐츠 관리 시스템 또한, 콘텐츠 열람을 시도하는 수요자에 대한 확인 없이 연령을 충족하는 타인의 개인정보 입력을 통해 간단하게 열람할 수 있다는 문제점을 가진다. 추가적으로, 현재 유동 인구 통계 시스템의 경우 통화 기지국 신호, 인터넷 접속 정보와 같은 외부 데이터로부터 유동 인구 정보 집계를 진행하기에 오차 발생 가능성이 있다는 한계를 가진다. 앞서 언급한 다양한 시스템의 한계는 얼굴 연령 분류 시스템 적용을 통해 각 시스템에 대해 신뢰도 제고 및 정확도 개선을 통해 극복할 수 있다. 얼굴 연령 분류 기법 적용을 통한 시스템의 개선을 위해서는 높은 정확도가 필수적으로 요구된다. 얼굴 이미지의 경우 눈, 코, 주름 등과 같은 얼굴 내 전역적인 모든 구성요소의 연관성을 고려해야 한다. Vision Transformer(ViT)[1]기반 알고리즘의 경우 패치(Patch) 단위의 상관관계를 Self-Attention을 통해 학습하기에 합성곱 신경망 기반 알고리즘보다 이미지의 전역적인 특징을 통한 분류에 용이하다. 따라서, 본 논문에서는 ViT 기반 얼굴 연령 분류 기법의 성능을 비교 분석한다. ViT, Swin Transformer(ST)[2], Neighborhood Attention Transformer(NAT)[3] 세 가지 알고리즘의 얼굴 연령 분류 기법 성능을 비교 분석을 통해 ViT가 가장 적합한 알고리즘임을 입증한다.

II. Proposed Method

본 논문에서는 얼굴 연령 분류에 가장 적합한 ViT 기반 알고리즘을 확인하기 위해 분류 성능 비교 분석 실험을 제안한다. Fig. 1과 같이

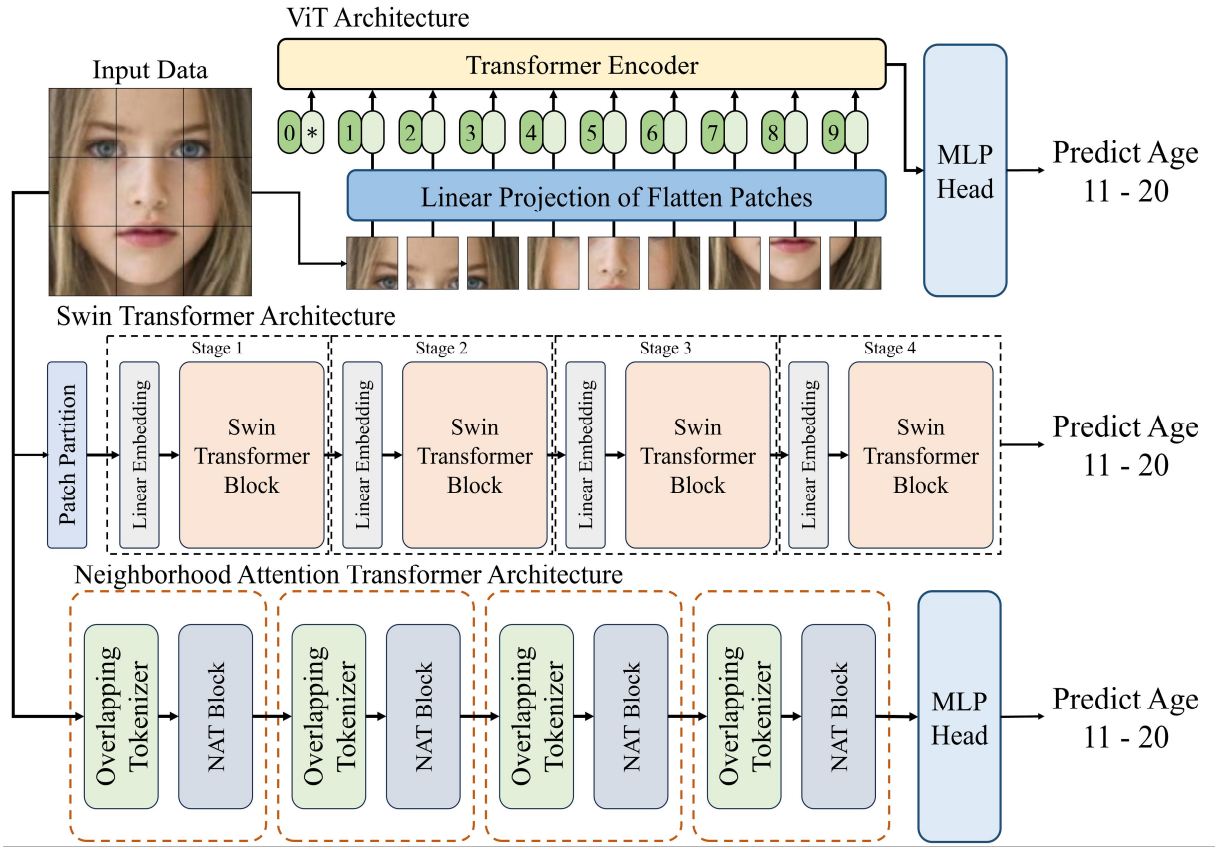

Fig. 1. Proposed Method Architecture

비교 분석 실험을 위한 알고리즘으로는 최근 다양한 Task에 널리 사용되는 ViT, ST, NAT 세가지로 선정하였다. ViT의 경우 입력 이미지를 Patch 단위로 나눈 후에, Self-Attention 알고리즘을 통해 Patch 내 모든 픽셀 간 상관관계를 통해 얼굴 연령 분류를 진행한다. ST의 경우 Patch 간 영역 중첩을 허용하는 Shifted Window based Self-Attention 방식을 통해 인접하지 않은 patch 간의 상관관계 분석에 용이하다. NAT의 경우 Neighborhood Attention을 통해 동적 이동이 가능한 Patch를 통해 넓은 수용영역을 기반으로 각 Patch 관 상관관계를 이용한다. 따라서, 본 논문에서는 얼굴 연령 분류에 가장 적합한 ViT 기반 알고리즘을 확인하기 위해, 위와 같이 각기 다른 구조를 가지는 알고리즘 ViT, ST, NAT에 대한 얼굴 유형 분류 성능 비교 분석을 진행한다.

III. Experiment

1. Dataset

ViT 기반 알고리즘의 얼굴 연령 분류 성능 비교 분석실험을 위한 Facial Age Dataset[4]은 데이터 공개 플랫폼 Kaggle에서 이용하였다. 원본 Dataset의 경우 1세부터 93세까지 93개의 폴더와, 95세부터 110세 범위에는 6개의 폴더로 총 99개의 폴더 내에 9,778개의 이미지로 구성된다. 하지만, 원본 Dataset 이미지 내부에 워터 마크가 존재하여 얼굴에 폐색(Occulsion)이 발생하거나, 1세 폴더에 노인 사진이 포함되어있는 오라벨링된 경우와 같이 얼굴 특징 학습에 방해 가능성이 있는 요인을 다수 확인하였다. 또한, 1세 이미지의 경우 1,112장으로 구성되지만 13세 이미지의 경우 75장으로 구성되는 등 특정 클래스에 대한 과적합 요인 가능성이 있는 클래스 불균형 문제를 확인하였다.

따라서, 본 논문에서는 실험 진행을 위해 10살 단위로 클래스를 축소 통합하여 클래스 불균형 문제를 해결하였으며, 폐색이 발생하지 않은 이미지를 선별하여 원본 Dataset 내에 존재하는 문제점을 해결하였다. Dataset 구성은 Table 1과 같다.

Table 1. Dataset

	Train	Valid	Test
01-10	240	30	30
11-20	240	30	30
21-30	240	30	30
31-40	240	30	30
41-50	240	30	30
51-60	240	30	30
61-70	240	30	30
71-80	240	30	30
81+	240	30	30

2. Experiment Result

얼굴 연령 분류 실험을 위한 하이퍼파라미터 설정으로는 최고 Epoch을 200으로 설정하였으며, 과적합 방지를 위해 Loss가 5회

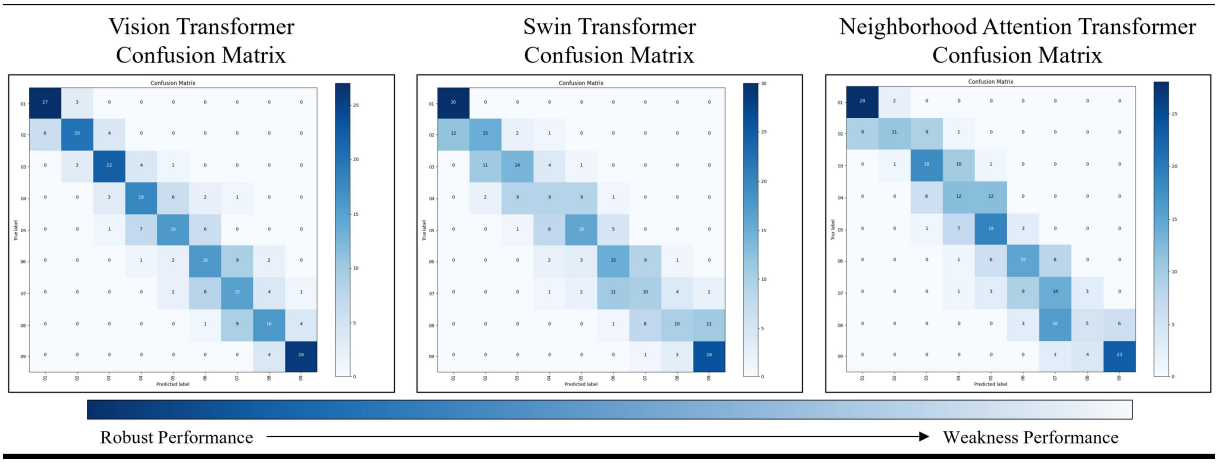

Fig. 2. ViT Base Models' Confusion Matrix Result

이상 감소하지 않는다면 학습을 종료하는 Early Stop을 적용하였으며, 실험 결과는 아래 Table 2와 Fig. 2와 같다.

Table 2. ViT Base Models' Test Result

	Accuracy	Precision	Recall	F1-Score
ViT	**0.6519**	**0.6548**	**0.6519**	**0.6519**
ST	0.5370	0.5221	0.5370	0.5215
NAT	0.5370	0.5513	0.5370	0.5259

Table 2를 확인해보면 정확도, 정밀도, 재현율, F1-Score 측면에서 모두 ViT가 뛰어난 분류 성능을 보이는 것을 확인할 수 있다. 하지만, Fig. 2에서 확인할 수 있는 것처럼 1세에서 10세 및 80세 이상 이미지에 대해서는 모든 모델이 강건한 성능을 보였지만, ST와 NAT의 경우 10세 이상 80세 이하 Dataset에서 강건하지 못한 성능을 확인할 수 있다. ViT의 경우 10세 이상 80세 이하 Dataset에서 타 알고리즘 보다는 강건한 성능을 보이지만 여전히 많은 오분류가 존재함을 확인할 수 있다. 이러한 결과를 통해 Facial Age Dataset에서의 얼굴 연령 분류의 경우 계층적 구조를 가지는 Tansformer 보다 ViT가 더 효과적임을 확인하였다.

IV. Conclusions & Future Work

본 논문에서는 얼굴 연령 분류에 가장 적합한 알고리즘을 도출하기 위해 ViT, ST, NAT 세 가지 모델에 대해 실험을 진행하였다. 실험 결과 ViT가 얼굴 연령 분류 실험에 가장 적합한 모델임을 확인하였지만 정확도가 높지 않다는 단점을 확인하였다. 실험 분석 결과 10세 이상 80세 이하 연령의 데이터셋에서 다수의 오분류 샘플을 도출하는 것을 확인하였다. 추후 2D 얼굴 이미지에 대해 3D로 변환하는 Depth Estimation 기법 등의 적용을 통해 보다 정확한 얼굴 연령 분류 기법을 꾸준히 연구할 예정이다.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2014-3-00077).

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021. DOI : 10.48550/arXiv.2010.11929

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022, October 2021, DOI : 10.48550/arXiv.2103.14030

[3] A. Hassani, S. Walton, J. Li, S. Li and H. Shi, "Neighborhood Attention Transformer," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6185-6194, June 2023 DOI :10.48550/arXiv.2204.07143

[4] https://www.kaggle.com/datasets/frabbisw/facial-age/data