멀티 프레임을 활용한 시선 추적모델에 관한 연구

이정민 [1], 한요섭 [2, 3]

[1] 숭실대학교 정보통신공학과

[2] 숭실대학교 전자정보공학부

[3] 숭실대학교 지능형반도체학과

peter8366@soongsil.ac.kr, yoseob.han@ssu.ac.kr

A Study on Gaze Estimation Models Utilizing Multi Frames

Jung Min Lee[1], Yoseob Han[2, 3]

School of Information and Communication Engineering, Soongsil University

School of Electronic Engineering, Soongsil University

Department of Intelligent Semiconductors, Soongsil University

요 약

본 논문은 기존의 단일 프레임 기반 시선 추정(single-frame gaze estimation)연구와 달리, 하나의 프레임의 시선 위치(gaze point)를 예측하기 위해 이전 여러 프레임의 정보를 함께 활용하는 멀티 프레임 기반 시선 추정(multi-frame gaze estimation) 방법을 제안하였다. 제안 모델은 시각적 정보 처리에 특화된 Video Swin Transformer(VST)를 백본으로 채택하여, 시선이 이산적인 점이 아닌 연속적인 흐름을 형성한다는 점을 반영해 이전 프레임들의 시계열 컨텍스트(temporal context)를 동시에 학습하도록 설계되었다. Gaze360 데이터셋을 대상으로 수행한 실험에서, 제안한 모델이 기존의 단일 프레임 모델 들에 비해 우수한 예측 성능(predictive performance)을 달성함을 확인하였다.

Ⅰ. 서 론

얼굴 이미지에서부터 사용자의 시선 위치를 정확하게 예측하는 것은 인간의 인지와 행동을 이해하는데 매우 중요한 역할을 한다. 이러한 시선 위치의 예측은 가상 현실(Virtual Reality), 의료 및 헬스케어(Healthcare & Medical Diagnostics), 운전자 보조 시스템(Driver Assistance Systems), 심리학 연구(Psychological Research)등과 같이 다양한 분야에서 필수적인 요소로 활용되고 있다. 이에 따라 시선 추적 방법은 크게 모델 기반 방법(model-based)과 외관 기반 방법(appearance-based)의 두 가지 방식으로 구분되어 다양한 연구가 이루어져 왔다.

모델 기반 방법(model-based)은 기하학적인 눈의 특성을 활용하여 시선 좌표를 예측한다. 이 방법은 동공 중심, 홍채 경계, 각막 반사 등의 해부학적 특징을 분석해 정확한 시선 예측을 수행한다. 반면 외관 기반 방법(appearance-based method)은 얼굴 이미지나 눈 영역의 이미지로만 머신러닝/딥러닝 알고리즘을 통해서 시선 좌표를 예측한다. 최근에는 CNN 기반의 appearance-based 방법을 사용하는 방법에 대해서 많은 연구가 진행되고 있다. [1]

최근 Transformer [2] 기반 모델이 자연어 처리 분야에서 뛰어난 성능을 보인 이후, 컴퓨터 비전 분야에서도 활발히 응용되고 있다. 특히 appearance 방식의 시선 추정 (gaze estimation)에서도 CNN 기반 모델을 대체하거나 보완하기 위해 transformer 구조를 적용하는 연구가 진행되어 왔다. Transformer 는 self-attention 메커니즘을 통해 입력 전체에 대한 전역적인 관계를 효과적으로 학습할 수 있어, 지역적인 특징에 집중하는 CNN 보다 저 정교한 시선 예측이 가능하다. 실제로, 여러 시선 추정 연구에서 transformer 기반 모델이 CNN 보다 우수한 예측 정확도를 기록하였다.

기존의 대부분의 시선 추정 연구는 단일 프레임 기반의 이미지를 사용하여 예측을 하였지만, 실제로 시선은 시간에 따라 연속적으로 변화하는 동적인 특성을 지닌다. 이러한 시간의 시계열 적인 흐름을 반영하기 위해 본 연구에서는 과거 프레임 정보를 함께 활용하는 멀티프레임 기반 시선 추정 모델을 제안한다. 이를 위해 시간적 정보를 처리할 수 있는 Video-Swin-Transformer(VST)를 활용하였다. [3] VST 는 기존에는 행동 인식과 같은 비디오 기반 작업에 최적화된 구조로 설계되었으며, 시간 축을 따라 프레임 간의 관계를

학습할 수 있는 장점을 지닌다. 본 연구에서는 이 구조를 시선 추적 문제에 맞게 변형하여, 하나의 타겟 프레임에 대한 시선 예측을 수행할 때 인접한 이전 프레임들을 함께 활용하는 모델을 제안한다.

Ⅱ. 본론

인간의 시선은 연속적인 움직임을 가지며, 이러한 연속성은 심리학적, 신경과학적 연구에서도 명확히 관찰된다. 즉, 시선은 하나의 독립적인 정적인 상태가 아니라, 이전 시점의 시선과 물리적인 연관성을 가지며 시간에 따라 부드럽게 변화하는 동적인 시계열 데이터로 볼 수 있다. 이러한 성질은 수학적으로는 다음과 같이 간단히 표현될 수 있으며, 시간 t 에서의 시선 벡터 $g_t \in \mathbb{R}^3$ 는 과거 시점들의 시선 정보를 기반으로 예측할 수 있다.

$$g_t \approx f(I_{t-k}, \dots, I_{t-1}, I_t)$$

이때, I_{t-1}은 $t-1$ 시점의 입력 프레임이며, $f()$는 시선 추정 모델을 의미한다. 이 수식은 단일 프레임 기반 접근과 달리 시간 정보를 보존해 더 풍부한 정보를 활용할 수 있음을 강조한다. CNN 은 국소적인 공간 정보를 처리하는 데 특화되어 있으나, 멀티프레임 간의 장기 의존성을 모델링 하는 데는 한계가 있다. 반면 Transformer 구조는, 특히 VST 를 활용하면 self-attention 을 통해 거리가 먼 프레임이어도 정보를 직접 연결할 수 있으므로 이러한 관계를 효율적으로 학습할 수 있다.

본 연구에서는 VST 의 시공간 토큰화 구조를 유지하되, 입력을 비디오 전체 길이 대신 8 프레임의 고정 윈도우로 정의하였다. 예측 대상 프레임 n을 중심으로 바로 앞 7 프레임을 함께 스택하여 $T = 8$의 시퀀스를 구성하고, 각 프레임을 2x4x4 크기의 3-D 패치로 분할하여 시공간 토큰을 생성한다. 이렇게 함으로써 self-attention 연산은 과거 프레임의 움직임 정보를 현재 예측에 자연스럽게 통합한다. 모델의 헤드 부분은 행동 분류용 softmax 계층을 제거하고, 전역 평균 풀링 뒤의 768 차원 특징을 256 차원으로 투영한 후 $\mathbb{R}^{256\to3}$ 회귀층으로 교체하였다. 이 회귀 헤드는 정규화된 3 차원 시선 벡터 (x,y,z)를 직접 예측한다. 학습 시에는 MSE 손실함수를 사용하였고, 성능 평가는 시선 추정 분야의 표준 지표인 Angular Error 를 사용하여 기존 연구와의 정량적 비교 가능성을 확보하였다.

Figure 1 은 위에서 서술한 시선 좌표를 예측하는 모델의 구조의 모식도를 보여준다. 입력으로는 8 개의 프레임이 시간 축의 방향으로 스택이 되어서 들어가게 되고 3D 패치 임베딩을 통해서 입력이 토큰화가 된다. 모델에 들어가서 Stage 1 부터 Stage 4 를 거치게 되는데 이렇게 해서 나온 임베딩 벡터를 최종적으로 회귀 헤드를 통과하여 시선 좌표의 형태로 출력이 된다.

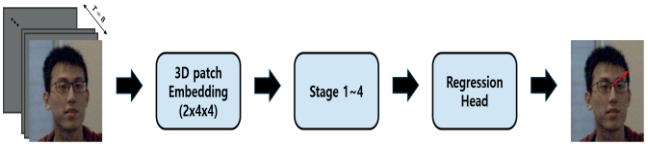

Figure 1 멀티프레임 얼굴 시퀀스를 모델에 입력하여 3 차원 시선 벡터를 예측하는 전체 모델 구조

Ⅲ. 결론

본 연구에서는 시선 추정에 멀티프레임 정보를 도입하고, 이를 VST 기반 구조로 구현함으로써 전역적 시공간 의존성을 효과적으로 학습할 수 있음을 실증하였다. 8 프레임 윈도우 입력, 3 차원 회귀헤드, 그리고 Angular Error 중심의 평가 프로토콜을 적용한 결과, 기존의 여러 연구에서 제안한 모델들보다 낮은 Angular Error 을 달성하였다. Table 1 을 보면 논문에서 제안한 모델을 이용하여 학습한 결과 Angular Error 가 10.39°가 나옴을 확인할 수 있다. 이러한 성능 우위는 시선이 시각적으로 연속적이라는 가정을 제안하는 모델의 구조를 통해 입증하는 것이다.

향후 연구 방향으로는 두가지의 확장 방향성을 계획하고 있다. 첫째로 얼굴 이미지와 별도로 양안(eye-patch) 크롭을 입력에 같이 넣어주어서, 눈 주위 세부 특징을 더욱 정교하게 묘사하는 방안이다. 이는 다양한 연구에서 사용하는 방법으로, 결국 시선은 눈의 영향을 많이 받으므로 더 자세한 특징을 학습할 수 있다. 두번째로는 순차적인 예측(sequential framework) 프레임워크를 도입하여, 이전 프레임에서의 예측된 시선 좌표를 다음 프레임 예측에 조건으로 활용하는 방법이다. 이 방법은 시선 궤적의 부드러운 흐름을 모델이 보존하도록 유도함으로써 장기 예측 안정성을 높일 것으로 보인다.

Method	Gaze360	RT-Gene	CA-Net	GazeTR-Hybrid	Proposed
Angular Error	11.04°	12.26°	11.20°	10.62°	10.39°

Table 2 최신 기법과의 비교. 제안된 방법이 Gaze360 데이터셋에서 최고 성능을 달성함.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 - 대학 ICT 연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2020-II201602)

참 고 문 헌

[1] Cheng, Yihua, et al. "Appearance-based gaze estimation with deep learning: A review and benchmark." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[3] Liu, Ze, et al. "Video swin transformer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.