특집논문 (Special Paper)
방송공학회논문지 제26권 제6호, 2021년 11월 (JBE Vol.26, No.6, November 2021)
https://doi.org/10.5909/JBE.2021.26.6.748
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

딥 러닝 기반의 눈 랜드마크 위치 검출이 통합된 시선 방향 벡터 추정 네트워크

주 희 영[a)], 고 민 수[a)], 송 혁[a)‡]

Deep Learning-based Gaze Direction Vector Estimation Network Integrated with Eye Landmark Localization

Heeyoung Joo[a)], Min-Soo Ko[a)], and Hyok Song[a)‡]

요 약

본 논문은 눈 랜드마크 위치 검출과 시선 방향 벡터 추정이 하나의 딥러닝 네트워크로 통합된 시선 추정 네트워크를 제안한다. 제안하는 네트워크는 Stacked Hourglass Network를 백본(Backbone) 구조로 이용하며, 크게 랜드마크 검출기, 특징 맵 추출기, 시선 방향 추정기라는 세 개의 부분(Part)으로 구성되어 있다. 랜드마크 검출기에서는 눈 랜드마크 50개 포인트의 좌표를 추정하며, 특징 맵 추출기에서는 시선 방향 추정을 위한 눈 이미지의 특징 맵을 생성한다. 그리고 시선 방향 추정기에서는 각 출력 결과를 조합하여 최종 시선 방향 벡터를 추정한다. 제안하는 네트워크는 UnityEyes 데이터셋을 통해 생성된 가상의 합성 눈 이미지와 랜드마크 좌표 데이터를 이용하여 학습하였으며, 성능 평가는 실제 사람의 눈 이미지로 구성된 MPIIGaze 데이터셋을 이용하였다. 실험을 통해 시선 추정 오차는 3.9°의 성능을 보였으며, 네트워크의 추정 속도는 42 FPS(Frame per second)로 측정되었다.

Abstract

In this paper, we propose a gaze estimation network in which eye landmark position detection and gaze direction vector estimation are integrated into one deep learning network. The proposed network uses the Stacked Hourglass Network as a backbone structure and is largely composed of three parts: a landmark detector, a feature map extractor, and a gaze direction estimator. The landmark detector estimates the coordinates of 50 eye landmarks, and the feature map extractor generates a feature map of the eye image for estimating the gaze direction. And the gaze direction estimator estimates the final gaze direction vector by combining each output result. The proposed network was trained using virtual synthetic eye images and landmark coordinate data generated through the UnityEyes dataset, and the MPIIGaze dataset consisting of real human eye images was used for performance evaluation. Through the experiment, the gaze estimation error showed a performance of 3.9, and the estimation speed of the network was 42 FPS (Frames per second).

Keyword : Gaze Estimation, Eye Landmark Localization, Eye Landmark Detection

Ⅰ. 서 론

2017년 포켓몬 고(Pokemon Go)의 폭발적 인기는 콘텐츠 시장에서 스마트폰 기반의 증강현실(Augmented Reality)을 활용한 애플리케이션이 지니는 가치가 무궁무진함을 보여주었다. 증강현실이란 실제 환경과 그래픽 가상 사물을 합성하여 가상의 사물이 마치 실제 환경에 존재하는 것처럼 정보를 만들어내는 컴퓨터 그래픽 기법[1] 이며, 가상현실(Virtual Reality)이란 실제가 아닌 환경을 인공으로 만들어내는 기법[2] 이다. 시장 조사기관 스태티스타(Statista)는 2024년 증강현실 및 가상현실 시장 규모는 3000억 달러에 이를 것으로 예측하고 있으며[3], 구글 사 (社)의 구글 글래스(Google Glass)[4]와 마이크로소프트(Microsoft) 사(社)의 홀로렌즈(HoloLens)[5]와 같은 웨어러블(Wearable) 기기는 증강현실 및 가상현실 기술의 대중화를 앞당기고 있다. 그림 1과 같은 웨어러블 기기들은 인간-컴퓨터 상호작용(Human-Computer Interaction)을 바탕으로 사용자의 움직임, 신호 등을 입력으로 받는데 이 때 사용자의 시선 정보는 매우 중요한 데이터라 할 수 있다. 구글 어스(Google Earth)는 HMD(Head Mounted Display) 장비를 착용한 사용자의 시선 정보와 조작에 따라 가상의 세계를 보여주는 소프트웨어로서 가상현실 기술에 시선 정보를 적극적으로 활용한 대표적인 기술이다[6]. 사용자의 시선 정보를 활용하면 개인의 시각적 관심을 파악할 수 있으므로 시선 추정 기술은 증강현실 및 가상현실 뿐 아니라 마케팅, 게임 분야에 활용되고 있다. 그림 2는 현대모비스 사(社)가 개발하고 있는 운전자 시선 추적을 이용한 부주의 교통사고 예방 기술과 관련된 것으로서[7], 이러한 개발 흐름은 시선 추정 기술의 활용 범위가 점점 확대되고 있음을 보여준다.

(a) Google Glass

(b) HoloLens

그림 1. 사용자의 시선 정보를 입력으로 갖는 웨어러블 기기의 예 (a) 구글 글래스 (b) 홀로렌즈
Fig. 1. Examples of wearable devices which take gaze information of users as an input
(a) Google Glass (b) HoloLens

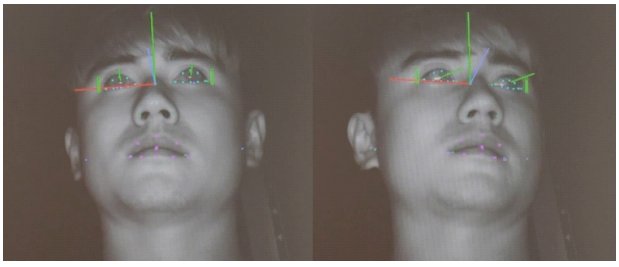

그림 2. 현대모비스 사(社)의 운전자 시선 추적 개발 내용
Fig. 2. Hyundai Mobis' research development about driver's eye tracking

시선 추정(Gaze Estimation)이란, 사용자의 시선 방향을 포함한 시선 정보를 추정하는 것을 말한다[8]. 그림 3[9]과 같이 사용자의 눈동자로부터 응시지점까지의 방향을 추정하고 이를 화살표나 직선을 이용하여 시각화 하여 그 결과를 나타낸다. 그동안의 시선 추정 기술은 기존 전통적인 패턴 분석 방법을 시작으로 발전하였으며 최근에는 인공지능 기술이 발전함에 따라 딥 러닝을 기반으로 한 시선 추정 기술

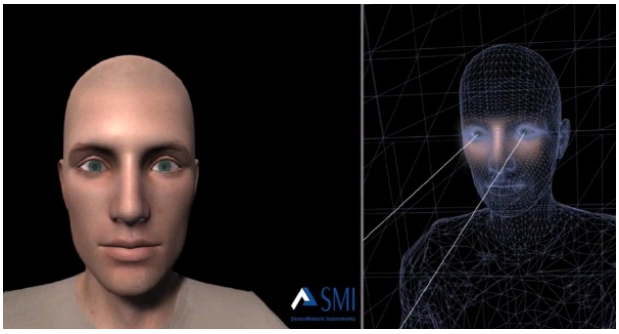

그림 3. 시선 추정 결과의 시각화 예시
Fig. 3. An example of the visualization of the gaze estimation

a) 한국전자기술연구원(Korea Electronics Technology Institute, KETI)
‡ Corresponding Author : 송혁(Hyok Song)
E-mail: hsong@keti.re.kr
Tel: +82-31-789-7000
ORCID: http://orcid.org/0000-0003-0376-9467
※ 이 논문의 연구 결과 중 일부는 한국방송·미디어공학회 "2021년 하계 학술대회에서 발표한 바 있음.
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-01982, Development of online exam fraud prevention and class concentration improvement technology).
· Manuscript received September 9, 2021; Revised November 8, 2021; Accepted November 16, 2021.

이 소개되고 있다. 딥 러닝 기반의 시선 추정 기법에서는 눈 랜드마크(Eye Landmark)의 위치 검출(Localization) 태스크(Task)가 시선 방향(Gaze Direction)을 추정하는 데 있어서 중요한 역할을 한다. 이전에 제안된 딥 러닝 기반의 시선 추정 기법들은 크게 두 단계로 구분되었다고 볼 수 있는데, 첫째 단계는 눈 랜드마크 위치 검출 단계이고 둘째 단계는 최종 시선 방향 벡터를 추정하는 단계이다. 각 단계마다의 태스크라 할 수 있는 랜드마크 위치 검출 태스크와 시선 방향 추정 태스크는 각기 다른 네트워크로부터 분리되어 학습된다는 특징이 있다. 본 논문은 서로 연관되어 있는 각 태스크가 서로 다른 구조의 네트워크로부터 학습되어 행해졌던 이전의 기법을 개선하여, 단 하나의 네트워크에서 두 가지의 태스크가 학습되어 시선 추정이 이루어지는 새로운(Novel) 네트워크를 제안한다. 본 논문은 눈 랜드마크 위치 검출과 시선 방향 벡터 성분 예측이 하나의 네트워크에서 학습이 가능한 통합 네트워크를 설계하기 위해 이라는 손실 함수(Loss Function)을 제안하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개한 후 3장에서 제안하는 기법에 대해 설명한다. 4장에서는 제안하는 바에 대한 실험과 성능 평가 결과를 소개한 후 5장에서 결론을 맺는다.

Ⅱ. 관련 연구

1. 시선 추정(Gaze Estimation)

시선 추정(Gaze Estimation) 연구는 특징-기반(Feature-based), 모델-기반(Model-based), 외형-기반(Appearance-based)이라는 세 가지의 큰 흐름으로 발전해왔다[10]. 각 흐름을 간략히 요약하면 먼저, 특징-기반 방법은 인간의 눈에 대해 수작업으로 얻은 특징 벡터(Handcrafted Feature Vector)를 시선 추정에 활용한다[10][11]. 통계적 머신러닝 기법을 기반으로 한 시선 추정 연구는 특징-기반 방법에 해당하며 이 방법론의 연구들은 조명(Illumination Condition), 머리의 위치(Head Position), 안경 착용 등으로 인한 폐색(Occlusion)과 같은 변화(Variation)에 강인(Robust)한 모델을 설계하기 위한 방향으로 발전하였다[12]. 모델-기반 방법은 사람의 안구를 두 개의 구가 교차하는 형태로 모델링하는 데에서 출발한다[13]. 3차원 공간에 안구를 배치하고 이 보다 크기가 작은 가상의 구가 안구와 적정 범위 내에서 교차하며 움직이는 것으로 눈동자의 움직임을 모델링(Modeling)하는 것이다. 이 때, 큰 구와 작은 구 사이의 교차되면서 생기는 단면은 홍채(Iris)에 대한 모델링에 해당한다. 외형 기반 방법은 눈 영역에서 피처(Feature)를 추출하여 동공의 위치를 검출하는 방법으로 딥 러닝 기반의 시선 추정 기법들이 여기에 속한다[10]. 딥 러닝 기반의 시선 추정 모델들은 위에서 언급한 외형적 변화들에 있어서 특징-기반 방법에 비해 강인하다는 장점이 있다. 딥 러닝 기반의 시선 추정 연구 중 특히, Stacked Hourglass Network를 백본(Backbone) 네트워크로서 적용한 연구에는 S. Park et al.의 연구[14]와 동일한 저자의 연구[22]가 있다. 연구[14]에서 제안하는 모델은 3개의 Hourglass Network를 쌓은 네트워크를 기반으로 눈 랜드마크를 검출한 후, SVR(Support Vector Regressor)[15]을 적용해서 최종 시선 방향을 얻는다. 연구[22]에서 제안하는 모델은 3개의 Hourglass Network모듈을 기반으로 불리언(Boolean) 타입의 중간 결과물(Intermediate)인 Gazemap을 생성하는 네트워크와 Gazemap으로부터 최종 출력인 시선 방향 벡터를 출력하는 DenseNet[23] 네트워크로 구성되어 있다. 이 연구에서는 눈 랜드마크에 대한 위치 검출 과정 없이, 저자들이 별도 정의한 중간 결과물 Gazemap을 얻기 위해 Stacked Hour glass Network를 적용하였으며 최종 시선 방향 벡터는 DenseNet을 기반으로 회귀한다. 한편, VGGNet[24]을 백본 네트워크로 기반한 Z.Xucong et al.의 시선추정 연구[18]가 있다. 이 연구에서 제안하는 네트워크는 입력 눈 이미지로부터 피처를 추출하는데, 눈 랜드마크 위치 정보는 피처 추출 과정에 주입되지 않는 대신, 얼굴 랜드마크 검출 정보로부터 계산된 머리 위치 정보(Head pose information)가 네트워크의 특정 레이어(Layer)에 주입되어 최종 시선 방향 벡터를 출력한다.

본 논문이 제안하는 모델은 눈 랜드마크의 검출과 시선 방향 벡터 추정이 분리된 네트워크에서가 아닌 하나의 통합 네트워크에서 한 번에 학습된다는 점에서 S. Park et al.이 제안한 두 개의 연구[14, 22]와 차별점이 있다. 또한, 눈 랜드마크 위치 정보를 네트워크의 피처 추출 과정에 직접

적으로 주입한다는 점에서 연구[18]과 차별점이 있다. 본 논문이 제안하는 기법은 모델-기반 방법에서 사용한 안구 모델을 기반으로 모델링 된 시선 방향 벡터를 딥러닝 네트워크를 기반으로 추론한다는 점에서 모델-기반과 외형-기반의 방법론을 모두 적용하였다고 볼 수 있다.

2. Stacked Hourglass Network

인간 포즈 추정 태스크를 위해 Newell A. et al.은 신체와 관련된 공간적 관계를 포착하기 위해 다양한 규모(Scale)에 걸쳐 특징을 처리하는 컨볼루션(Convolutional) 네트워크 아키텍처인 Stacked Hourglass Network를 제안하였다[16]. 모래시계(Hourglass)라는 네트워크의 이름에서 유추할 수 있듯이 이미지 피처는 다운샘플(Downsample)을 거듭하다가 업샘플(Upsample)과정을 다시 거치면서 이전의 피처들과 조합되는 구조를 갖는다. 이는 이미지 전반에 대한 피처와 국소(Local) 부분에 대한 피처를 모두 추출할 수 있다는 장점이 있다. Newell A. et al.이 논문에서 제시한 실험결과는 이러한 구조가 다양한 스케일로 피처를 추출할 수 있으므로 주요 신체 랜드마크의 위치 검출과 최종 포즈 추출 태스크에 적합한 구조임을 보여준다.

Ⅲ. 제안하는 기법

본 논문은 인간 포즈 추정 태스크를 위해 설계된 Hourglass Network를 백본 네트워크로 하여, 눈 랜드마크 픽셀 위치와 시선 방향을 추정할 수 있는 딥 러닝 네트워크를 제안한다. 제안하는 네트워크는 눈 영역 이미지를 입력으로 하여 눈 랜드마크 좌표와 함께 구면 좌표계의 표현법을 따르는 시선 방향 벡터를 추정한다. 추정 결과는 직교 좌표계에서의 벡터로 변환된 후 이를 동공에 해당하는 중심점을 시작점으로 하는 화살표 선으로 시각화하여 출력된다.

1. 안구 모델링을 이용한 시선 방향 추정 기법

인간의 시선 방향을 추정하기 위해 본 논문에서는 시선 방향을 구면 좌표계에서의 벡터 $[g]_s = (1, \theta, \phi)$로 설정하였다. 이러한 모델링은 그림 4가 나타내는 바와 같이 모델 기반 시선 추정 기법에서 사용하는 인간의 안구 모델링 방법을 따른 것이다. 이러한 모델링 기법에 따르면 인간의 동공은 3차원 공간 상의 구면 위에 위치하고 있으며 이 때 구면은 인간의 안구를, 인간의 눈동자 움직임은 이 구와 교차하

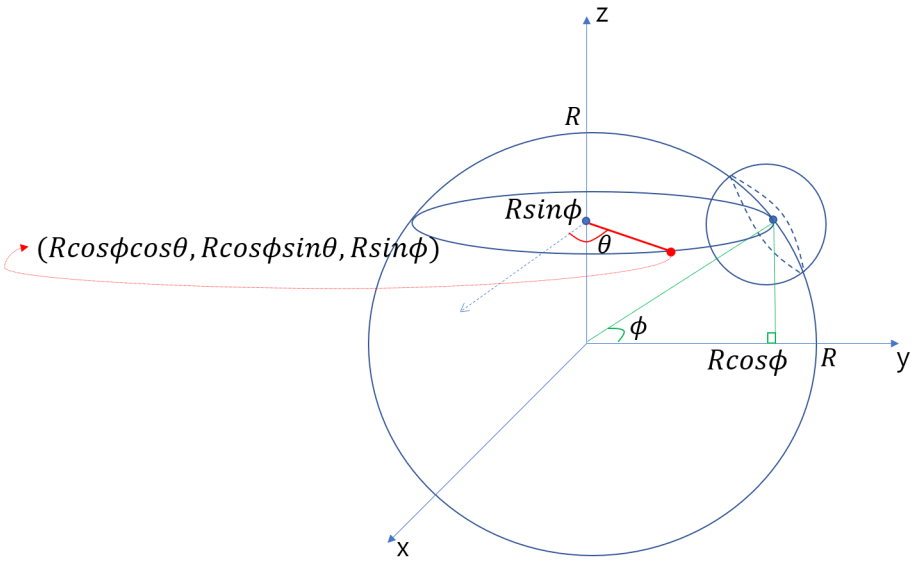

그림 4. 인간의 안구와 동공의 움직임 모델링
Fig. 4. The modeling of human eye ball and the movement of a pupil

며 움직이는 가상의 또 다른 작은 구로 모델링된다. 동공은 구면위에 위치하는 점으로 모델링 가능하므로 인간의 시선 방향을 구면 좌표계에서 $[\hat{g}]_s = (1, \hat{\theta}, \hat{\phi})$로 표현한다. 한편, 학습 데이터셋 UnityEyes[17]는 직교 좌표계(Cartesian Coordinate System)에서 벡터의 노름(Norm)이 1로 정규화된 시선 방향 벡터 $[g]_c = (x, y, z)$에 대한 어노테이션(Annotation)을 제공하므로 아래의 (식 1)을 사용하여 좌표 변환을 수행하였다.

$$tan\theta = \frac{y}{x},\ tan\phi = \frac{\sqrt{1-z^2}}{z} \tag{1}$$

2. 제안하는 네트워크 아키텍처

제안하는 네트워크는 재귀구조를 갖는 Hourglass Network를 3개 쌓은 Stacked Hourglass Network를 백본 네트워크로 갖는다. 흐름 구조상 그림 5가 나타내는 바와 같이, 랜드마크 검출기(Landmark Detector), 특징 맵 추출기(Feature Map Extractor), 그리고 시선 방향 벡터 추정기(Gaze Direction Vector Estimator)라는 세 개의 부분(Part)으로 구성되어 있다. 표 1은 그림 5에 나타난 Convs Layer의 상세에 해당한다. 이들은 표 1이 나타내는 바와 같이 컨볼루션 계층과(Convolutional Layers) 잔차 블록(Residual Block)으로 구성되어 있다. 먼저, 랜드마크 검출기는 다수의 컨볼루션 계층으로 구성되어 있으며 각 픽셀에 대하여 각 랜드마크가 위치할 확률을 행렬의 형태로 표현한 히트맵(Heatmap)과 함께 각 랜드마크의 픽셀 위치 좌표를 추정한다. 이는 전체 네트워크 구조상 랜드마크 검출기에 포함된다고 볼 수 있다. 특징 맵 추출기는 시선 방향 벡터 추정

표 1. 제안하는 네트워크의 Convs Layers 상세
Table 1. The details of Convs Layers in proposed network

Convs Layer 1	Conv2d $(7\times7, 64)$, stride 1, BatchNorm2d, Relu
	Residual $\begin{pmatrix}3\times3, 128\\3\times3, 128\end{pmatrix}$
	Maxpool2d (2×2)
	Residual $\begin{pmatrix}3\times3, 128\\3\times3, 50\end{pmatrix}$

Convs Layer 2	Conv2d $(7\times7, 64)$, stride 1, BatchNorm2d, Relu
	Residual $\begin{pmatrix}3\times3, 128\\3\times3, 128\end{pmatrix}$
	Maxpool2d (2×2)
	Residual $\begin{pmatrix}3\times3, 128\\3\times3, 50\end{pmatrix}$

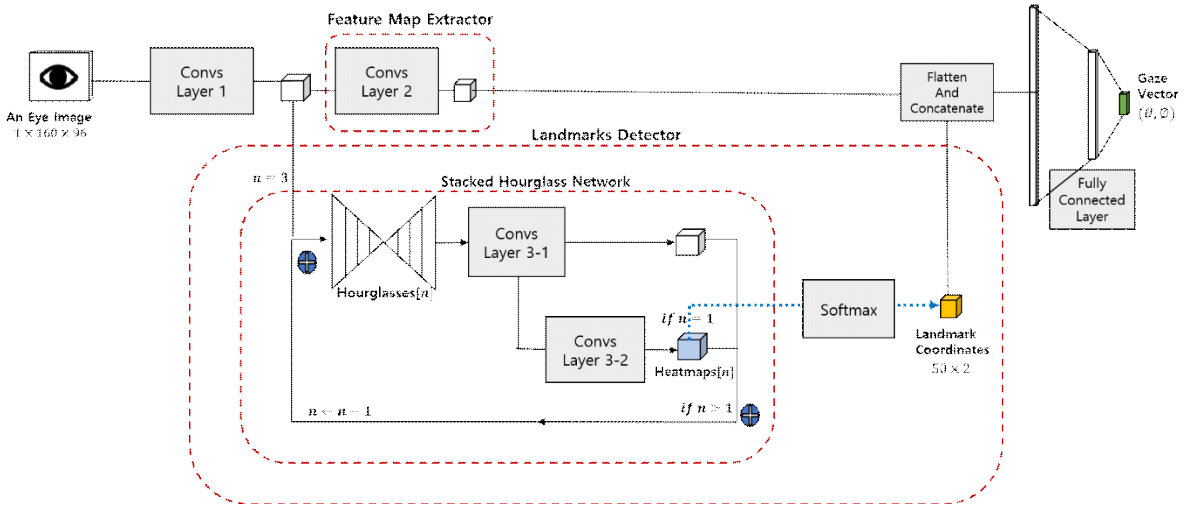

그림 5. 제안하는 네트워크 아키텍처
Fig. 5. The architecture of proposed network.

에 필요한 눈 이미지의 특징 맵(Feature Map)을 생성한다. 이는 하나의 눈 이미지에 대해 다수의 컨볼루션 계층을 통과하여 얻는다. 시선 방향 벡터 추정기는 랜드마크 검출기와 특징 맵 추출기로부터 생성된 각각의 출력 결과를 조합하여, 최종적으로 정규화 된 시선 방향 벡터를 추정한다. 구체적인 연산을 설명하면, 앞의 두 생성 결과에 픽셀-단위 덧셈(Pixel-wise Addition) 연산이 적용된 후에 Flatten과 Concatenate이 이루어진 후 완전연결계층(Fully Connected Layer)를 통과하여 최종적으로 두 개의 성분으로 구성된 시선 방향 벡터를 얻는다. 이 벡터는 단위 구면 에 대한 구면 좌표계 표현을 따른 것이므로 추정되는 벡터의 차원은 2차원이다. 제안하는 네트워크는 50개의 눈 랜드마크들의 위치 검출과 함께 2차원 시선 방향 벡터의 성분(Component)에 대한 회귀(Regression)가 단 하나의 네트워크에서 학습될 수 있는 구조를 갖는다. 이러한 구조는 신경망 네트워크 단위의 관점에서 볼 때, 시선 방향 추정에 필요한 모든 피처들이 하나의 스테이지(Stage)에서 학습된다는 특징이 있다. 한편, 각 추정 결과인 히트맵, 랜드마크 픽셀의 위치 좌표, 시선 방향 벡터 성분에 대한 손실(Loss)은 모두 평균 제곱 오차(Mean Square Error)로 계산한다. 최종 손실 함수는 이들의 선형 결합(Linear Combination)으로 설정하였다. 제안하는 네트워크에 적용된 손실함수는 다음 절에서 설명하기로 한다.

3. 손실 함수

본 논문은 눈 랜드마크 위치 검출과 시선 방향 벡터 추정에 대한 학습이 하나의 네트워크에서 가능하도록 하기 위해 L_{total}이라는 손실함수를 제안하였다. 이는 히트맵 예측, 랜드마크 예측, 시선 방향 추정이라는 서로 연관 있는 여러 개의 태스크(Task)을 결합하여 학습하는 Multi-task Learning을 위해 설계된 손실함수이다. 함수식은 다음과 같다.

$$L_{total} = \alpha L_{heatmap} + \beta L_{landmark} + \gamma L_{gaze} \quad (2)$$

이 때, L_{total}을 구성하는 각 항(Term)은 다음과 같다.

$$L_{heatmap} = \frac{1}{N}\sum_{i=1}^{N}\sum_{p} \| \hat{h}_i(p) - h_i(p) \|^2 \quad (3)$$

$$L_{landmarks} = \frac{1}{N}\sum_{i=1}^{N} \| \hat{w}_l - w_i \|^2 + \| \hat{h}_l - h_i \|^2 \quad (4)$$

$$L_{gaze} = \| \hat{\theta} - \theta \|^2 + \| \hat{\phi} - \phi \|^2 \quad (5)$$

각 항을 살펴보면 먼저, $L_{heatmap}$란 각 픽셀에 대하여 특정 랜드마크가 위치하는 확률에 대한 오차이며, $L_{landmark}$란 각 랜드마크의 픽셀 좌표에 대한 오차이다. L_{gaze}란 시선 방향 벡터 성분에 대한 오차이다. $L_{heatmap}$, $L_{landmark}$, L_{gaze}는 모두 평균 제곱 오차(Mean Square Error)를 적용하여 얻는다. 이 때, L_{total}은 세 가지 종류의 항(Term)인 오차, $L_{heatmap}$, $L_{landmark}$, L_{gaze}에 가중치를 둔 덧셈 연산을 적용하여 얻는다. 본 논문은 (식 2)의 가중치에 해당하는 α, β, γ를 각각 1, 1, 1000으로 설정하여 네트워크 학습을 진행하였다.

Ⅳ. 실 험

1. 데이터셋

본 논문에서 제안하는 네트워크를 학습하기 위해 사용한 데이터셋은 두 종류로서 UnityEyes[17]와 MPIIGaze[18]이다. UnityEyes는 네트워크의 학습을 위해, 그리고 MPIIGaze는 네트워크의 성능 평가를 위해 사용하였다. UnityEyes는 영국 Cambridge 대학에서 제작한 소프트웨어로서 가상의 눈 이미지와 이에 대한 레이블(Label)을 생성하는 프로그램이다. 본 논문은 UnityEyes를 사용하여 생성한 가상의 눈 이미지 30791장을 학습용 데이터로, 3849장을 검증용 데이터로 사용하였다. 그림 6[19]은 UnityEyes가 제공하는 눈 이미지에 대한 예시이다. 한편, MPIIGaze는 15명의 실험 참가자로부터 3개월동안 노트북을 사용하는 실제 일상을 수집한 데이터셋이다. 제약 조건이 없는(Unconstrained) 환경에서 실제 사람의 눈 이미지를 수집한 것이기 때문에 눈의 모양 및 조명이라는 환경 변화에 큰 편차가 보장된다. 이러한 특성으로 MPIIGaze는 제약 조건이 없는 모양 기반 시선 추정에 있어서 표준 벤치마크 데이터 셋으로 최근 몇 년동안의 시선 추정 분야에서 활용되어왔다[22]. 제안하는 네트워크는 실제 사람의 눈이 아닌 가상의 데이터인 UnityEyes로만 학습이 이루어졌기 때문에 MPIIGaze 데이터셋에 대

한 성능 평가 결과는 모델의 강인성(Robustness)을 보여준다고 할 수 있다. 본 논문은 37767장의 MPIIGaze 데이터셋을 성능 평가에 사용하였다. 그림 7[20]은 MPIIGaze 데이터셋의 예시이다.

그림 6. UnityEyes로부터 생성된 가상의 눈 이미지 예시
Fig. 6. Examples of virtual eye images generated from UnityEyes

그림 7. MPIIGaze 데이터셋 예시
Fig 7. Examples of MPIIGaze dataset

2. 구현 상세

네트워크 학습에 앞서, UnityEyes와 MPIIGaze로부터 네트워크 학습의 입력(Input)에 적합한 해상도의 눈 영역 이미지만을 추출하는 전 처리(Pre-processing)를 진행하였다. 데이터 증강(Augmentation)을 위해 이미지의 밝기(~0.1), 대비(~0.1), 채도(satuaration, ~0.05), 색상(hue, ~0.05)를 변경하는 Color Jitter를 적용하였으며 눈 랜드마크 픽셀 값 및 시선 방향에 대한 어노테이션을 고정시키기 위해 어파인 변환(Affine Transformation)은 적용하지 않았다. 또한, 인위적인 폐색을 적용하기 위해 3개 이하의 무작위(Random) 기울기를 갖는 선분을 추가하여 이미지에 변형을 주었다. 제안하는 네트워크 학습에 사용된 하이퍼-파라미터(Hyper-parameter)를 정리한 결과는 표 2와 같다. Adam (Adaptive Moment Estimator)[21] 최적화 기법을 사용하였으며 초기 학습률에 대해 매 25 에포크(Epoch)마다 배씩 감소시키며 이를 조정하였다. 실험 및 성능 평가에 사용된 PC의 사양은 표 3과 같다.

표 2. 제안하는 네트워크 학습에 사용된 하이퍼-파라미터 값
Table 2. Hyper-parameters used for training the proposed network

Hyper-parameter	Type/ Value
Optimizer	ADAM (Adaptive Moment Estimation)
Initial Learning Rate	4×10^{-4}
Learning Rate Scheduler	Step, Decrease a learning rate by every 25 epoch

표 3. 실험에 사용된 PC 사양
Table 3. Specifications of the PC used in the experiment and evaluation

	Specification
CPU	Intel® Core™ i9-10850K 3.60GHz
GPU	Nvidia® GeForce GTX TITAN XP
RAM	32GB
OS	Ubuntu 20.04 LTS (64bit)

3. 성능 평가

성능 평가를 위해 사용한 지표(Metric)는 예측 시선 방향 벡터 성분과 레이블 시선 방향 벡터 성분과의 평균 시선 추정 오차(Mean Gaze Estimation Error)이다. 성능 평가 결과 제안하는 모델의 평균 시선 추정 오차는 3.9°이고 추론 시간(Inference Time)은 42 FPS를 얻었다. 표 4는 제안하는 네트워크의 성능 평가 결과와 다른 모델과의 성능 비교 결과를 나타낸다. 성능 비교를 위해 사용한 모델은 S. Park et al.가 제안한 모델[14,22]과 GazeNet[18]이다. 모델[14,22]은 Stacked Hourglass Network를 백본 네트워크로 사용하였다는 것은 동일하지만, 눈 랜드마크 검출과 시선 방향 벡터 추정이 하나의 네트워크가 아닌 분리된 네트워크에서 이루어진다는 점이 본 논문과 다르다. 또한 본 논문에서 제안하는 모델은 GazeNet과 달리 눈 랜드마크 위치 정보를 네트워크의 피처 추출 과정에 직접적으로 주입한다는 점에서

차별점이 있다. 제안하는 모델은 모델 [14]와 모델 [22]에 비해 각각 15.2% (4.6° → 3.9°)와 13.3% (4.5°→ 3.9°) 개선되었으며, 모델 [18]에 비해 29% (5.5°→ 3.9°) 개선되었다. 즉, 제안하는 통합 네트워크 구조는 기존 모델들에 비해 개선된 구조임을 보여준다.

한편 제안하는 모델에 적용된 손실함수 (식 2)를 구성하는 세 종류의 항인 (식 3), (식 4), (식 5)이 성능에 어떠한 영향을 미치는지 파악하기 위한 Ablation study를 진행하였으며 표 5는 그 결과를 나타낸다. 표 5의 결과를 분석하면, 최종 목적 함수(Objective Function) L_{total}에 항 L_{gaze}를 덧붙임으로서 그렇지 않을 경우에 비해 87% (31°→ 3.9°) 더 높은 성능을 갖는 통합 모델을 구현할 수 있었다.

아래의 그림 8, 그림 9, 그림 10은 정성평가의 한 예시이다. 그림 8은 본 논문에서 제안하는 기법에 따른 실시간 시선 추정 프로그램의 출력을 캡처(Capture)한 예시이다. 그림 9는 MPIIGaze 데이터 셋의 추정 결과 중, 추정 오차가 추정 오차가 3°미만인 경우에 해당하며, 그림 10은 추정 오차가10°이상인 경우에 대한 예시이다. 그림 10을 살펴보면, 추정이 잘 이루어지지 않은 이유로는 원본 이미지의 훼손 또는 안경으로 인해 랜드마크가 제대로 검출되지 않은 경

표 4. 제안하는 네트워크의 성능 평가 결과
Table 4. The result of the performance of the proposed network

Model	The value of Mean Gaze Estimation Error (degrees)
S. Park et al. [14]	4.6
S. Park et al. [22]	4.5
GazeNet [18]	5.5
Ours	**3.9**

표 5. 제안하는 손실함수의 Ablation study 결과
Table 5. The result of the Ablation study on our proposed Loss fundction

Loss function	The value of Mean Gaze Estimation Error (degrees)
$L_{total} = L_{heatmap} + L_{landmark} + 1000 L_{gaze}$	3.9
$L_{total} = L_{heatmap} + L_{landmark}$	31

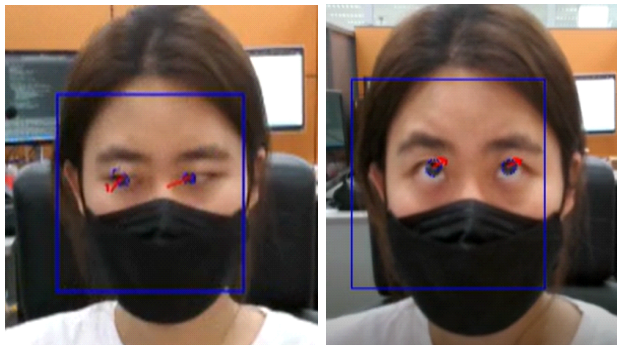

그림 8. 제안하는 기법에 따른 시선 추정 결과 예시
Fig. 8. Examples of the gaze estimation result based on proposed method

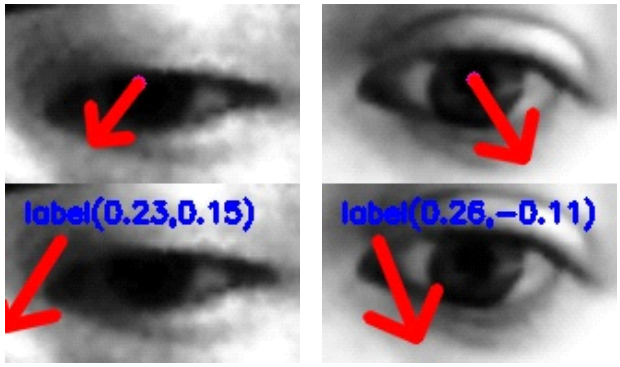

그림 9. MPIIGaze 데이터 셋의 추정 결과 중, 추정 오차가 추정 오차가 3° 미만인 경우 예시
Fig. 9. Among the estimation results of the MPIIGaze data set, an example where the estimation error is less than 3°

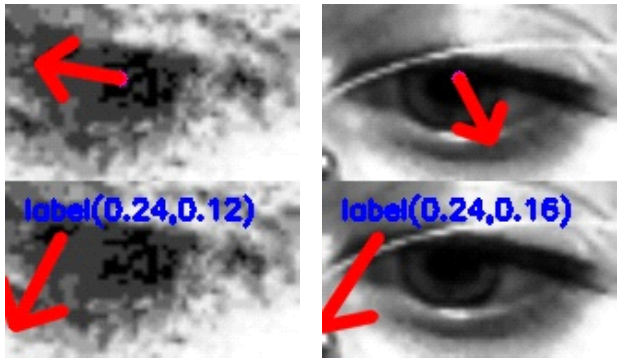

그림 10. MPIIGaze 데이터 셋의 추정 결과 중, 추정 오차가 추정 오차가 10° 이상인 경우 예시
Fig. 10. Among the estimation results of the MPIIGaze data set, an example where the estimation error is more than 10°

우로 분석할 수 있다. 가상의 눈 이미지만을 학습 데이터 셋으로 사용하여, 안경을 착용한 이미지가 학습 데이터에 포함되어 있지 않아 안경이 눈 랜드마크 지점을 일부 가리는 경우 추정 결과에 오차가 발생하는 것으로 분석된다.

V. 결 론

본 논문은 눈 랜드마크 위치 검출과 시선 방향 벡터를 하나의 네트워크로 학습하여 추정할 수 있는 딥 러닝 기반 네트워크를 제안한다. 제안하는 기법은 랜드마크 좌표를 추정하는 단계 까지만 딥 러닝 학습을 수행하였던 기존 연구와는 달리, 랜드마크 위치 검출과 시선 방향 벡터 추정을 하나의 네트워크로 한 번에 학습이 가능하다는 점에서 기존의 연구와 차별점이 있다. 이러한 통합 구조의 제안하는 네트워크는 기존 연구에 비해 최소 13.3%, 최대 29%의 성능 개선을 이끌었다. 제안하는 통합 네트워크의 구조가 갖는 장점은 다음과 같다. 첫째, 시선 추정에 필요한 서로 연관된 두 개의 태스크에 대한 학습이 분리되지 않으므로, 태스크끼리 갖는 상호작용 및 연관성을 부여하여 파라미터를 피팅(Fitting)시킬 수 있다. 둘째, 모델의 전체적인 구조가 단순 해졌으므로 모델이 갖는 복잡도가 감소하였다. 이는 입력 이미지에 적용되는 연산이 줄어들었음을 의미하므로 추론시간의 감소에 영향을 준다는 장점이 있다.

참 고 문 헌 (References)

[1] J. Carmigniani, and B.Furht, “Augmented Reality: An Overview,” Springer, New York, pp.3-46, 2011.

[2] R. Sherman, and Alan B. Craig, “Understanding Virtual Reality: Interface, Application, and Design, Second Edition,” Morgan Kaufmann Series in Computer Graphics, Massachusetts, pp.3-58, 2018.

[3] The Market prediction for the virtual, augmented, and mixed reality technology by Statista https://www.statista.com/statistics/591181/global-augmented-virtual-reality-market-size (accessed Sep. 3, 2021).

[4] Oliver J. Muensterer, Martin Lacher, Christoph Zoeller, Matthew Bronstein, and Joachim Kübler, “Google Glass in pediatric surgery: An exploratory study,” International Journal of Surgery, Vol.12, No.4, pp.281-289, 2014.

[5] M. Tepper, L. Rudy, A. Lefkowitz Aaron, A. Weimer, M. Marks, S. Stern, and S. Garfein, “Mixed Reality with HoloLens: Where Virtual Reality Meets Augmented Reality in the Operating Room,” Plastic and Reconstructive Surgery, Vol.140, No.5, pp.1066-1070, 2017.

[6] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore,“Google Earth Engine: Planetary-scale geospatial analysis for everyone,” Remote Sensing of Environment, Vol.202, pp.18-27, 2017.

[7] Hyundai mobis’ research development about driver’s eye tracking, http://www.epnc.co.kr/news/articleView.html?idxno=91211 (accessed Sep. 3, 2021).

[8] Anuradha Kar, and Peter Corcoran, “A review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms,” IEEE Access, Vol.5, pp.16495-16519, 2017.

[9] An example of the visualization of the gaze estimation, https://www.hankyung.com/it/article/201701051859v (accessed Sep. 3, 2021).

[10] Sunghyun Cho, “Introduction to eye-tracking technology,” The Magazine of the IEEK, Vol.45, pp.23-32, 2018.

[11] Laura Sesma, Arantxa Villanueva, and Rafael Cabeza, “Evaluation of pupil center-eye corner vector for gaze estimation using a web cam,” Proceedings of the Symposium on Eye Tracking Research and Applications, pp.217-220, 2012.

[12] A. Tsukada, M. Shino, M. Devyver, and T. Kanade, “Illumination-free gaze estimation method for first-person vision wearable device”, 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp.2084-2091, 2011.

[13] Christian Nitschke, Atsushi Nakazawa, and Haruo Takemura, “Display-camera calibration using eye reflections and geometry constraints”, Computer Vision and Image Understanding, Vol.115, No.6, pp.835-853, 2011.

[14] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, pp.1-10, 2018.

[15] Mariette Awad, Rahul Khanna, "Support vector regression," Efficient learning machines, pp.67-80, 2015.

[16] Alejandro Newell, Kaiyu Yang, Jia Deng, "Stacked hourglass networks for human pose estimation," European conference on computer vision, pp.483-499, 2016.

[17] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, Andreas Bulling, "Learning an appearance-based gaze estimator from one million synthesized images," In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, pp. 131-138. 2016.

[18] Xucong Zhang, Yusuke Sugano, Mario Fritz, Andreas Bulling, "MPIIGaze: Real-world dataset and deep appearance-based gaze estimation," IEEE transactions on pattern analysis and machine intelligence Vol.41, pp.162-175, 2017.

[19] Examples of UnityEyes dataset,

[20] https://www.cl.cam.ac.uk/research/rainbow/projects/unityeyes/tutorial.html#:~:text=UnityEyes%20is%20a%20tool%20for,for%20other%20eye%20tracking%20systems (accessed Sep. 3, 2021).

[21] Examples of MPIIGaze dataset, https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/gaze-based-human-computer-interaction/appearance-based-gaze-estimation-in-the-wild (accessed Sep. 3, 2021).
[22] Diederik P. Kingma, Jimmy Ba, "Adam: A method for stochastic optimization," arXiv preprint, 2014.
[23] Park, Seonwook, Adrian Spurr, and Otmar Hilliges, "Deep pictorial gaze estimation," Proceedings of the European Conference on Computer Vision (ECCV), pp. 721-738, 2018.
[24] Gao Huang, Zhuang Liu, Kilian Q. Weinberger, and Laurens van der Maaten, “Densely Connected Convolutional Networks,” arXiv preprint arXiv:1608.06993, 2016.
[25] Simonyan, Karen, and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

저 자 소 개

주 희 영

- 2014년 8월 : 이화여자대학교 수학교육과 학사
- 2021년 2월 : 이화여자대학교 컴퓨터공학과 석사
- 2020년 12월 ~ 현재 : 한국전자기술연구원 근무
- ORCID : http://orcid.org/0000-0003-4978-5649
- 주관심분야 : 컴퓨터 비전, 딥러닝

고 민 수

- 2010년 2월 : 광운대학교 전자공학과 학사
- 2012년 2월 : 광운대학교 전자공학과 석사
- 2016년 2월 : 광운대학교 전자공학과 박사
- 2015년 12월 ~ 현재 : 한국전자기술연구원 근무
- ORCID : http://orcid.org/0000-0003-0675-1756
- 주관심분야 : 영상 신호처리, 머신러닝

송 혁

- 2012년 : 광운대학교 전자공학과 공학박사
- 2000년 - 현재 : 한국전자기술연구원 수석연구원
- ORCID : http://orcid.org/0000-0003-0376-9467
- 주관심분야 : 영상분석, 영상인식, 보안시스템