JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 3, pp. 463-472, Mar. 2023

Check for updates

사용자의 시선 및 감정을 활용한 AI 기반 영상 콘텐츠 평가

곽 수 찬[1] · 김 지 윤[1] · 박 태 정[2*]
[1]덕성여자대학교 사이버보안 전공 학부과정
[2]덕성여자대학교 사이버보안 전공 교수

AI-Based Video Qualification Using the User's Gaze and Emotion

Soochan Kwak[1] · Jiyun Kim[1] · Taejung Park[2*]

[1]Bachelor's Course, Department of Cybersecurity, Duksung Women's University, Seoul 01369, Korea
[2]Professor, Department of Cybersecurity, Duksung Women's University, Seoul 01369, Korea

[요 약]

영상 콘텐츠 산업이 성장함에 따라 영상 콘텐츠의 제작과 개선을 위해 콘텐츠를 객관적으로 평가할 수 있는 도구에 대한 수요가 증가하고 있다. 본 논문에서는 시선 추적 기술을 이용하여 제작된 영상 콘텐츠의 제작 의도에 부합되는지 보다 객관적으로 평가하기 위한 방법론을 제안한다. 지금까지 영상 콘텐츠가 제작 의도에 부합되는지의 여부를 평가하기 위해서는 사용자의 주관적인 리뷰를 기반으로 한 텍스트 분석을 통해 이루어지는 것이 일반적이었다. 기존 방식은 영상의 목적 적합도를 객관적으로 평가하기 어려웠고 많은 시간 및 비용이 투입되어야 했으며 충분한 양질의 결과 확보가 어렵다는 한계점이 존재한다. 이러한 문제점을 보완하고자 본 논문에서는 전용 하드웨어 장비 없이 일반적인 RGB 카메라만을 사용하여 사용자의 시선과 감정을 인식하고 분석함으로써 사용자가 영상 콘텐츠에 얼마나 집중하고 있는지 객관적으로 측정할 수 있는 세 가지 평가 지표를 제안한다. 또한 이 평가 지표를 활용한 영상 분석 사용자 실험과 결과를 제시함으로써 제안하는 방법론의 타당성을 검증한다.

[Abstract]

As the video content industry grows, demand for tools that can objectively evaluate content for video production has increased. In this study, we propose a methodology using eye-tracking technology to more objectively evaluate whether the video content produced meets the production intention. Thus far, it was standard practice to employ text analysis based on users' subjective reviews to determine whether the video content met the production intention. Existing methods have limitations in that it is challenging to objectively evaluate the fitness for purpose of the video, an excessive amount of time and money is needed, and it is difficult to secure sufficient high-quality results. The three indicators in this study can compensate for these issues by recognizing and analyzing the user's gaze and emotion using a standard RGB camera without dedicated hardware, allowing us to determine how much the user is focusing on the video content. In addition, the validity of the proposed methodology is verified by presenting video analysis user experiments and results using these evaluation metrics.

색인어 : AI 기반 영상 콘텐츠 평가, 시선 추적, Saliency Map, 감정 인식, 사용자 집중도

Keyword : AI-based video content qualification, Gaze estimation, Saliency map, Emotion detection, User concentration

http://dx.doi.org/10.9728/dcs.2023.24.3.463

Received 15 February 2023; **Revised** 06 March 2023
Accepted 15 March 2023

***Corresponding Author; Taejung Park**

Tel: +82-2-901-8339
E-mail: tjpark@duksung.ac.kr

Ⅰ. 서 론

1-1 연구 배경

최근 미디어 플랫폼의 발전과 함께 1인 미디어 시장이 활성화되면서 개인도 손쉽게 영상을 제작하여 업로드할 수 있는 환경이 마련되었다. 1인 미디어가 트렌드로 자리를 잡게 되면서 사용자들에게 단순한 취미나 정보 제공의 목적이 아닌 수익 창출 수단으로까지 진화하고 있다. 이러한 시장의 흐름에 따라 영상 제작자들은 사용자의 피드백이 영상의 개선점과 나아갈 방향성을 제시한다는 측면에서 영상에 대한 사용자의 피드백을 바탕으로 그들의 요구사항을 보다 객관적으로 충족시키는 영상을 보다 체계적으로 제작해야 하는 필요에 직면하고 있다.

사용자의 피드백은 '영상의 품질'을 향상시키는데 매우 중요한 역할을 해 왔다[1]. 본 논문에서 언급하는 '영상의 품질'이란 영상의 화질을 의미하는 것이 아닌 해당 영상의 콘텐츠가 얼마나 사용자의 반응과 집중도를 이끌어내는 가를 의미한다.

이러한 측면에서 지금까지 영상 콘텐츠의 품질 평가는 주로 사용자가 해당 영상을 시청한 후 작성한 리뷰를 토대로 평가되어 왔다. 그러나 이러한 리뷰는 보다 객관적이고 엄밀한 평가 기준을 설정하기 쉽지 않으며 리뷰에 참여한 사용자의 주관적 견해나 개인적인 판단에 따라 평가가 이루어지는 것이 대부분이었다. 그 결과, 평가 결과물을 객관적인 지표로 활용할 때에는 분명한 한계가 있다.

본 논문에서는 이러한 한계를 극복하기 위해서 사용자가 영상 콘텐츠를 시청할 때에 수집한 시선 좌표 정보와 얼굴 사진을 토대로 한 사용자의 감정 판단 정보를 이용한다. 본 논문에서 다루는 영상 콘텐츠는 제작자 또는 기획자가 영상 내에 특정한 대상에 시청자가 관심을 가지고 보기를 원하는 광고, 개인 제작 동영상, 영화 내 특정 장면 등을 대상으로 한다. 이러한 측정 결과물을 보다 객관적으로 측정할 수 있도록 사용자의 시선과 감정을 기반으로 사용자의 피드백을 수치화하는 3가지 지표를 제안하고 실험 결과를 통해 그 타당성을 증명한다.

1-2 연구의 구성

본 연구는 영상 품질을 평가하기 위한 객관적이고 정량화된 데이터를 추출하는 새로운 방식을 제안하는 연구로 해당 논문의 구성은 다음과 같다. 1장은 서론으로, 연구를 시작한 배경과 기존 영상 품질을 평가하는 방식에 대한 한계점 설명과 더불어 연구의 목적을 설명한다. 2장은 연구 방법과 실험 대상, 실험 장치 및 실험 진행 과정에 대해 설명한다. 3장은 본론으로, 사용자의 객관적인 피드백을 얻기 위해 필요한 측정 지표들에 대해 제안한다. 영상에 감정적으로 반응하는 시청자의 비율, 영상에 집중해야 하는 영역에 잘 집중하고 있는 시청자의 비율, 영상에 시선이 있는 시청자의 비율을 측정하고 나타낸다. 4장은 결론으로 해당 연구가 적용될 분야에 대한 기대와 본 연구가 가지는 한계점에 대해 서술한다.

1-3 관련연구

1) 텍스트 분석 연구

가장 보편적인 영상 품질 평가 방법은 특정 영상에 대한 사용자의 댓글을 분석하는 방식이라고 볼 수 있다. 댓글을 통한 영상 품질 평가 방식은 2010년대부터 다양한 방법으로 측정되어왔고 댓글의 영향력이 점차 커지면서 관련 연구들이 많이 진행되었다[2]. 이러한 텍스트 분석 연구에서는 인터넷 용어를 포함한 다양한 댓글들을 분석하며 비속어, 은어, 줄임말, 이모티콘 등이 분석의 대상이 된다[3]. 이러한 연구에서는 인터넷 용어와 표준어와의 유사도를 기반으로 감성 분석 시스템을 통해 텍스트 분석을 진행한 결과를 영상 품질 개선을 위해 사용한다.

그러나 텍스트 분석 방식으로 영상의 품질을 판단할 경우 사용자의 주관적인 평가가 포함될 수도 있기 때문에 객관성을 확보하기가 어려운 측면이 있다. 또한 대다수의 댓글이 영상에 대한 전체적인 리뷰가 아닌 특정 관심 부분에 대한 사용자의 평가로 한정된다는 문제도 무시할 수 없다. 따라서 영상을 시청하는 사용자의 피드백을 각 시간대 별로 모두 알 수 없으므로 영상 전반에 대한 피드백을 얻기 어렵다. 게다가 사용자의 의도적인 거짓 리뷰로 인해 평가의 신뢰성이 저하되는 문제점도 발생할 수 있다[4].

2) 유튜브의 시청 지속 시간의 주요 순간 측정

유튜브에서는 그림 1에서 제시한 것처럼 시간대 별 시청자

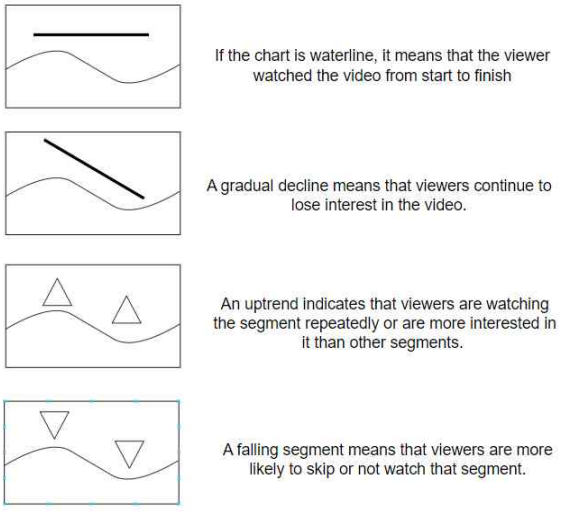


그림 1. 시청 지속 시간에 따른 4가지 그래프 모양
Fig. 1. Four graph shapes by duration of view

표 1. fixation과 saccade
Table 1. fixation and saccade

Eye Tracking	Description
fixation	Gaze points are continuously stamped at similar positions.
saccade	A property whose gaze points move quickly to a distant target rather than a similar location when compared to the previous gaze points

의 영상 및 영상 구간 별에 대한 관심도를 나타낸 4가지 유형의 그래프로 시간대 별 시청자 수를 그래프로 시각화하여 시청 지속 시간을 제시한다[5]. 이 시각화 방식에서는 그래프의 형태에 따라 동영상에서 시청자들이 많은 관심을 보였던 부분과 그렇지 못했던 부분을 확인할 수 있으며 제작자 입장에서는 이러한 정보를 영상 개선에 활용할 수 있다. 그러나 해당 측정 기술에서는 시청자가 영상을 재생하기만 하고 집중하고 있지 않아도 영상 시청 상태로 간주되어 시청자 통계에 포함되는 한계가 있기 때문에 결과의 신뢰성을 저하시킨다. 다시 말해서 특정 영상의 시간대 별 시청자 수가 높다고 해서 해당 부분에 대해 시청자들이 많은 관심을 가졌다고 단정하기 어려울 수 있다.

3) 시선 추적 기술과 응용

시선 추적 기술이란 특수 장비나 일반적인 RGB 카메라를 통해 사람의 눈동자를 관찰하여 비전 기술 또는 인공지능 기술을 기반으로 현재 바라 보고 있는 지점의 좌표를 파악하는 기술을 의미한다[6].

시선 추적 기술 구현 기술은 크게 모델 기반 시선 추적(model based gaze estimation)과 외형 기반 시선 추적(appearance based gaze estimation) 방식 두 가지로 나눌 수 있다[7]. 모델 기반 시선 추적 기술은 사용자의 안구의 기하학 정보를 분석해서 현재 바라보고 있는 지점을 계산하는 방식으로 보통 적외선 조명 및 적외선 카메라, 특수 전용 장비 등이 필요하다[8]. 이에 비해 외형 기반 시선 추적 기술은 일반적인 RGB 카메라를 이용해서 사용자의 얼굴을 촬영하고 인공지능에 기초하여 현재 바라보고 있는 지점을 추정한다[9].

최근 시선 추적 연구는 카메라를 바라보는 사용자 한 명의 시선 정보 추적을 넘어서 카메라에 촬영된 여러 명의 시선 정보까지 획득하는 연구로 진행되고 있다[10].

이러한 시선 추적 관련 연구는 마케팅, 광고, 심리학 등 다양한 분야에서 수행되고 있다. 인터넷 광고 효과 측정을 하기 위해 소비자의 시선을 활용해서 쇼핑몰 사이트 배너 광고를 중심으로 시각적인 요소의 가중치를 객관적으로 분석할 수 있는 방법이 제안됐으며[11] 모바일 커머스 관련 연구에서도 시선 추적을 접목시켜 사용자의 시선이 상품 유형에 따라 집중하는지 혹은 상품 내용에 따라 집중하는지에 대해 사용자의 시선을 분석함으로써 상품 구매 유도 전략을 파악하였다[12].

그러나 기존 연구에서는 대부분 안경 형태 등의 고가의 하드웨어 기반 시선 추적 장비를 이용하여 자연스러운 사용 환경에서의 사용자의 선택을 조사하기에는 한계가 있었다.

본 연구에서는 이러한 한계를 극복하고 보다 보편적인 연구 도구로 활용할 수 있도록 기존의 고가의 하드웨어를 사용하지 않고 웹캠 또는 휴대전화나 태블릿 등의 모바일 기기에 설치된 일반 RGB 카메라를 이용해서 시선 추적을 수행할 수 있는 VisualCamp사의 SeeSo SDK[13]를 이용하여 시선 추적을 수행하였다. 이 SeeSo SDK는 별도의 하드웨어가 필요하지 않으며 일반적인 스마트폰, 태블릿, PC 웹 환경에서 바로 응용할 수 있기 때문에 본 연구에서 제안하는 동영상 콘텐츠 품질 평가 방법론과 평가 지표를 보다 다양한 응용 분야에 대해 사용자의 일상적인 사용을 방해하지 않는 환경에서 적용할 수 있다는 장점이 있다.

또한 기존의 연구는 주로 인터넷, 쇼핑몰 등 정적인 이미지를 대상으로 한 시선 추적 연구에 국한되었으며 본 논문의 주제인 시선 추적을 통한 영상 콘텐츠 품질 평가를 위한 연구는 아직 활성화되지 않은 것으로 파악된다.

II. 본론

2-1 시스템 구성

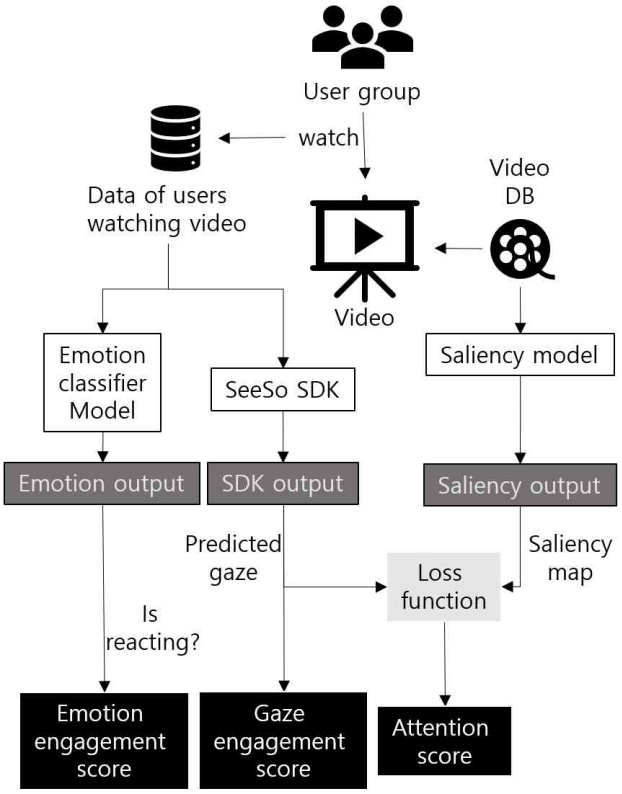


그림 2. 전체 시스템 구성도
Fig. 2. Overall system flowchart

그림 2에서는 본 연구를 위해 구현된 시스템을 제시한다. 가장 왼쪽에 제시된 사용자 그룹은 비디오 데이터베이스에서 제시된 비디오를 시청한다. 사용자가 비디오를 시청하는 동안 사용자의 얼굴 이미지와 기타 관련 정보가 SeeSo SDK와 인공지능 기반 감정 분류 모델(Emotion Classifier Model)[14]로 전달되고 비디오 정보는 Saliency Model[15],[16]로 전달되어 정보량이 많은 (즉, 중요한) 비디오 영역을 Saliency Output으로 출력한다. 사용자의 얼굴 이미지를 전달 받은 SeeSo SDK는 이미 사전에 학습된 정보를 바탕으로 현재 사용자가 화면에서 바라보는 지점의 2차원 좌표를 SDK Output으로 제공한다. 이 정보는 바로 사용자가 주어진 대상을 주의 깊게 살펴 보고 있는지를 평가하는 시선 반응 점수(Gaze Engagement Score) 계산에 활용되고 Saliency Map과 시선 좌표 사이의 차이(loss function)를 계산해서 집중 점수(Attention Score)를 계산한다. 또한 감정 분류 모델(Emotion Classifier Model)은 사용자가 현재 콘텐츠에 감정적으로 반응하는지의 여부를 계산해서 감정 반응 점수(Emotion Engagement Score)를 계산한다. 이 과정에서 사용자의 얼굴 이미지는 저장되지 않으며 시선 정보만 저장된다.

2-2 안구 운동 - Fixation과 Saccade

사람의 안구는 사물을 바라 볼 때 fixation과 saccade로 대표되는 움직임을 보인다[17]. 이러한 안구의 움직임은 시선 정보 추적에도 그대로 나타나며 시선 좌표가 연속적으로 가까운 위치에 클러스터를 형성하는 패턴을 fixation이라고 한다. fixation이 비교적 긴 시간 동안 형성될 경우, 해당 시선 좌표가 위치하는 곳에 있는 대상이 사용자의 관심과 흥미를 유발한다고 해석할 수 있다[18]. 또한 빠른 안구의 움직임과 함께 시선 좌표가 직전의 좌표와 비교했을 때 가까운 위치가 아닌 멀리 떨어진 대상으로 빠르게 이동하는 운동을 saccade라고 정의한다. 즉, 안구의 빠른 움직임으로 인한 시선의 순간적인 이동을 의미하며 또 다른 관심 대상을 찾아 시선이 이동하고 있다고 해석할 수 있다(표 1).

2-3 시선 데이터 전처리

시선 데이터에는 시선이 형성된 x 좌표, y 좌표 뿐만 아니라 시간(timestamp) 정보도 함께 저장된다. 만약 한 시선 지점이 fixation이라고 판단되면 해당 시선 좌표는 fixation 기간(fixation으로 판단된 첫 번째 시선부터 마지막 fixation 시선까지의 시간)에 속하는 모든 시선의 좌표값들을 fixation이 처음 시작된 시선의 좌표값으로 변환한다. 반면 saccade로 판정되는 경우, 이동이 진행 중인 좌표값은 무시한다. 그림 3의 왼쪽 그림에서 saccade로 판정된 시점 하나는 중간 과정으로 간주되고 무시되기 때문에 결국 fixation에서 fixation으로의 시선 도약 방향이 결정된다.

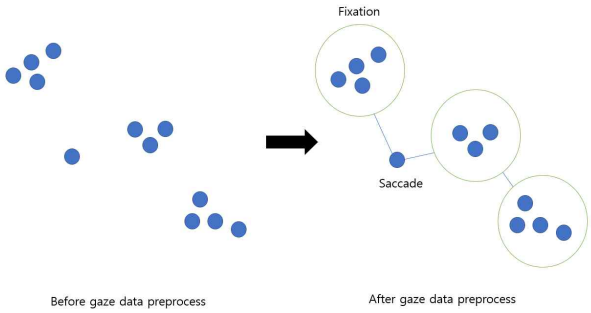


그림 3. fixation과 saccade
Fig. 3. fixation and saccade

2-4 실험 대상, 실험 장치 및 과정

본 논문에서 제안하는 방법론을 검증하기 위해 실시한 실험에서는 일반 직장인 12명과 대학 재학 중인 20대 초 중반 8명으로 총 20명이 참가하였다. 실험 참가자는 남자 10명(50%), 여자 10명(50%)이며 연령 대는 20대가 주를 이루었다. 시선추적 장비는 VisualCamp사의 SeeSo SDK[13]와 10.5인치 아이패드를 이용하여 데이터를 수집하였다. 사용자의 시선은 평소 아이패드를 바라보는 시선과 유사하게 하기 위해서 아이패드를 아래로 바라보는 방향으로 설정하였고 시선이 화면 가운데에 위치하도록 하였다. 영상은 길이와 주제가 다른 것으로 10개를 선별해 시청하도록 하였지만 해당 논문에서 제시하는 실험 사례는 하나의 영상만을 활용하여 분석을 하였다.

2-5 영상 품질 측정 및 분석

특정한 영상물이 얼마나 시청자들의 관심을 끌도록 제작되었는지 평가하는 작업은 개인적인 취향 등과 같은 주관적인 요소들이 작용하고 주변 소음, 주변의 다른 영상 매체 등 여러 주변적인 요소들이 개입되기 때문에 간단하지는 않다. 그러나 본 논문에서는 목적에 부합되도록 훌륭하게 제작된 영상물은 1) 시청자들이 영상 또는 영상이 재생되는 화면 이외의 다른 곳에 시선을 옮기지 않고 2) 영상에서 제작자의 의도가 나타난 중요한 지점들을 바라보며 3) 영상의 내용에 대한 감정적인 반응을 보일 것이라고 가정하고 이러한 세 가지 가정 하에 시선 추적 기술을 응용하여 보다 객관적으로 영상 콘텐츠의 품질을 통계적인 측면에서 평가하는 지표를 제안한다.

표 2에서는 본 논문에서 제안하는 측정 지표를 정리한다. 시선 반응 점수(Gaze Engagement Score)는 시선 추적을 통해 비디오 재생 중 비디오 또는 비디오가 재생되는 화면을 바라보는 시청자들의 비율로 계산된다. 집중 점수(Attention Score)는 Saliency Map을 통해 추출한 비디오 화면 내의 중요한 위치 또는 영상 제작자가 중요하다고 판단한 위치를 보는 시청자들의 비율로 계산된다. 마지막으로 감정 반응 점수(Emotion Engagement Score)는 인공지능을 이용해서 현

표 2. 분석에 사용된 측정 지표(Metrics)
Table 2. Metrics used in the analysis

Metrics	Description
Gaze Engagement Score	Percentage of viewers with eyes on the video
Attention Score	Percentage of viewers who focus well on areas where they need to focus on the video
Emotion Engagement Score	Percentage of viewers who react emotionally to the video

재 비디오 콘텐츠에 따라 감정적인 반응이 얼굴에 나타난 시청자의 비율을 계산한다.

이 세 가지 지표는 영상 제작자의 의도가 얼마나 충분히 반영되고 있는지를 평가하기 위한 것이다. 일반적인 영상 제작자의 의도는 1) 시청자들이 다른 주변 사물에 주의를 뺏기지 않고 영상이 재생되는 화면을 주목하고 2) 영상 내에서도 제작자나 기획자가 중요하다고 판단하는 사물에 집중하며 3) 그로 인한 사용자 감정 상태의 변화가 일어나도록 하려는 것으로 볼 수 있다. 이 세 가지 측정 지표는 독립적인 요소들이 아니라 순차적으로 발생하며 이전 단계가 이루어져야 다음 단계가 발생하는 과정이다. 예를 들어 특정한 영상이 1) 단계까지 충족되었으나 2). 3) 단계까지 도달하지 못했다면 제작자는 2) 단계에서 의도한 중요한 대상에 집중하도록 하지 못하는 요소들을 분석하여 영상을 수정할 수 있을 것이다.

1) 시선 반응 점수 (Gaze Engagement Score)

그림 4에서는 시선 반응 점수를 구하는 순서도를 제시한다. 시선 반응 점수(Gaze Engagement Score)는 영상이 재생되는 화면 또는 영상 내부에 사용자의 시선 추적 2차원 좌표값(fixation)이 포함되는 시청자의 비율을 나타낸다. fixation으로 구분된 시선 좌표가 스크린 내 또는 영상 영역 내에 존재하면 시청자가 해당 장면을 보고 있다고 판단한다. 반면에 시선 좌표가 스크린 내에 존재하지 않거나 연속된 두 시선이 saccade로 판정되는 경우에는 해당 장면을 보고 있지 않다고 판단한다. 이때 시선 좌표가 스크린 내에 존재하는지에 대한 판단은 fixation 클러스터에 처음 속한 시선 좌표의 중심점(그림 3)을 기준으로 한다.

시선 좌표를 시각화한 클러스터의 중심이 화면 밖에 존재할 경우 그림 5와 같이 해당 좌표를 (0, 0)으로 변경한다. 따라서 구현한 시스템에서 시선 좌표 (0, 0)은 시청자의 시선이 화면이 아닌 다른 곳을 보고 있다는 의미로 해석된다.

통계 정보 구축을 위해 초 단위를 기준으로 시청자가 화면을 보고 있는지에 대한 여부를 판단하기 위해 사용자가 시청한 영상의 초당 프레임인 24fps를 기준으로 하여 절반 이상의 프레임 길이 동안 화면 내에 시선 좌표가 존재한다면 해당 시간 내에는 시청자가 영상을 시청했다고 판단한다.

그림 6은 총 20명의 사용자에 대해 실험한 시선 반응 점수이다. 이 그래프를 통해 이 영상물은 시작(0초)부터 3초간은 90% 이상의 사용자들의 시선이 영상물로 향했음을 볼 수 있고 그 후 4초부터 5초 사이에는 40%~60%의 사용자만이 화면을 보고 있음을 알 수 있다.

이러한 시선 반응 점수를 통해 콘텐츠 제작자는 시선 반응 점수가 낮은 지점에서의 콘텐츠가 시청자의 관심을 저해하고 있는 문제를 확인하고 개선하거나 온라인 교육 콘텐츠 등에서 주변 여건 개선, 또는 학습자가 놓친 부분의 반복 강화 학습 등을 수행할 수 있을 것이다.

2) 집중 점수 (Attention Score)

집중 점수(Attention Score)는 영상에 집중해야 하는 영역에 잘 집중하고 있는 시청자의 비율을 나타내는 점수다. 집중 점수는 시선 추적 기술을 이용해서 시청자의 시선 좌표를 계산하고 이 좌표가 중요 지점에 잘 일치하는지 파악함으로

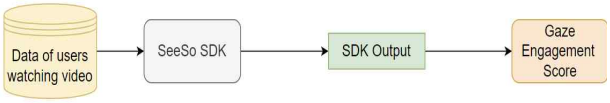


그림 4. 시선 반응 점수 계산 순서도
Fig. 4. Flow chart for the process of calculating Gaze Engagement Score

그림 5. 스크린 밖에 존재하는 시선
Fig. 5. Gaze out of the screen

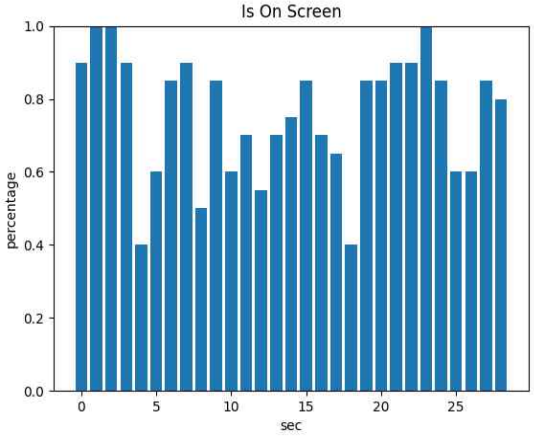


그림 6. 20명의 사용자에 대한 시선 점수 그래프
Fig. 6. Graph of Gaze Engagement Score for 20 users

써 시청자의 집중도를 판단한다. 집중 점수를 구하는 과정에 대한 순서도는 그림 7과 같다. 해당 실험은 사람이 움직이는 물체에 보다 많은 관심을 가지고 시선을 고정한다는 가정을 바탕으로 진행하였으며, 집중 점수를 계산하기 위해 Saliency Model을 사용했다[16]. 이 논문의 구현에서 사용한 Saliency Model은 영상 속 움직이는 패턴을 감지하여 관심 영역을 추출하는 모델이다. Saliency Model을 통해 나온 출력값(그림 8)을 히트맵으로 표현하여 시청자가 관심이 있만한 영역을 시각화하였다. 이는 시청자의 관심 영역을 파악하는 기준이 되어 영상의 품질을 높일 수 있는 자료로 활용된다. 본 논문에서는 Saliency Map을 이용했지만 제작 의도에서 시청자들이 반드시 봐야 할 부분을 Saliency Map 대신 직접 지정할 수도 있다.

집중 점수(Attention Score)를 계산하기 위해 영상을 시청하는 사용자의 시선 좌표와 사용자가 시청한 영상이 필요하다. 사용자가 시청한 영상을 Saliency Model의 입력값으로 하여 나온 출력값(그림 8)과 영상에 대한 사용자의 시선(그림 9)을 평균제곱오차(MSE)를 이용하여 loss 값을 계산한다. 기준값(threshold)을 2000으로 지정하고 이를 기준으로 사용자가 영상에 집중을 하고 있는지에 대한 여부를 확인한다. 그림 10에서는 3명의 시청자에 대한 Saliency Map의 출력값과 시선 좌표의 평균제곱오차(MSE) 계산 결과인 loss 값을 각각 제시한다. loss 값이 기준값보다 작으면 해당 프레임에 대한 시청자들의 집중도가 높고, 기준값보다 높으면 해당 프레임에 대해서 집중도가 낮은 것으로 판단한다.

그림 11에서는 Saliency Model의 출력값을 히트맵으로 생성한 이미지와 사용자의 시선 좌표를 실제 비디오 영상[19]에 합성한 이미지를 제시한다. 사용된 영상은 광고 영상으로 영상의 길이는 31초이다. 그림 11은 영상을 시청한 20명의 사용자의 시선 좌표를 시각화하였고, fixation의 기간이 길어질수록 시선을 시각화한 원의 크기가 커지도록 하였다. fixation의 기간이 길다는 것은 해당 좌표 주변을 오랫동안 응시했다는 것을 의미하기 때문에 시선을 시각화한 원의 크기를 크게 함으로써 사용자의 관심 영역을 명확하게 파악할 수 있다. 만약 특정 영역에 대다수의 사용자의 시선 고정 지

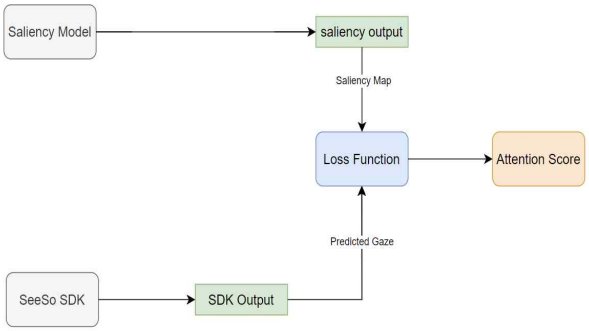


그림 7. 집중 점수를 구하는 과정에 대한 순서도
Fig. 7. Flow chart for the process of calculating Attention Score

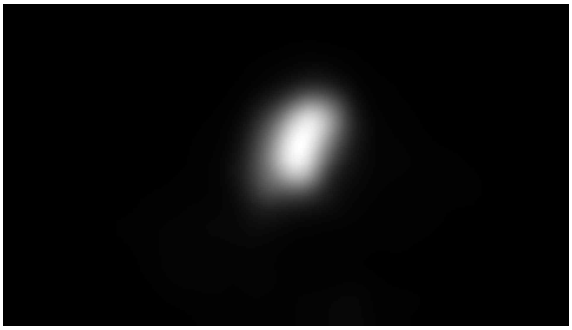


그림 8. Saliency Model의 출력값
Fig. 8. Output value of Saliency Model

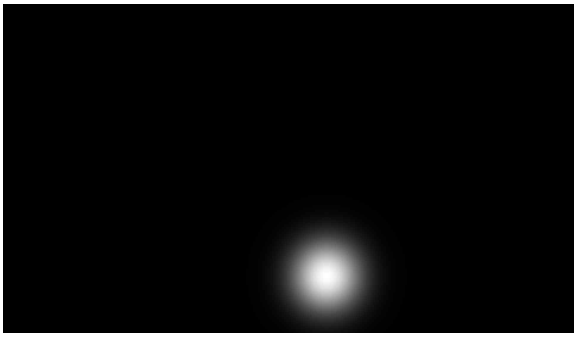


그림 9. 시선 좌표 시각화
Fig. 9. Visualization of gaze coordinates

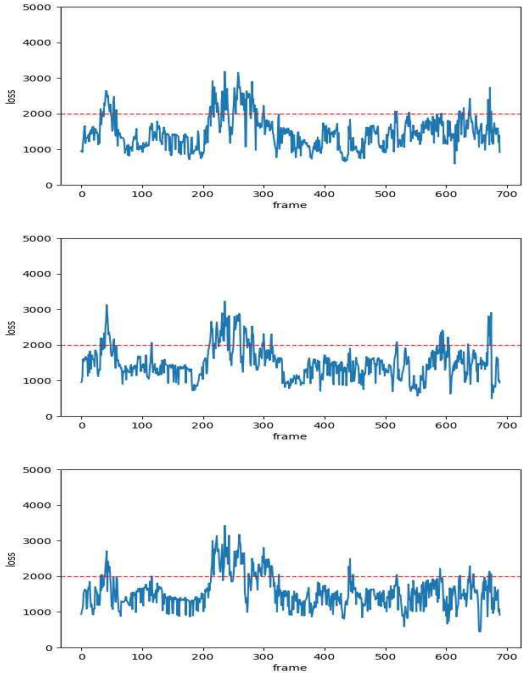


그림 10. Saliency Model의 출력값과 사용자의 시선 좌표의 loss를 나타내는 그래프
Fig. 10. Graph showing the difference between the output value of the Salience Model and the user's gaze coordinates

그림 11. Saliency Model의 출력값을 히트맵으로 생성한 이미지에 시청자 20명의 시선을 합친 이미지

Fig. 11. An image that combines the gaze of 20 users with an image generated by a heat map of the output value of the Saliency Model

속 시간이 길다면 그 영역에 사용자의 관심을 끄는 내용이 존재한다고 판단할 수 있다[20].

그림 11 상단 이미지를 살펴보면 20명의 사용자의 시선(원)이 히트맵 주위로 퍼져있는 것을 볼 수 있다. 이는 사용자들이 광고 글귀를 중점으로 집중하고 있다고 판단된다. 반면에 그림 11 하단 이미지의 경우 사용자의 시선 분포가 히트맵이나 광고 모델에 집중된 것이 아니라 흩어져 있는 것을 볼 수 있다. 사용자의 시선이 히트맵보다는 히트맵 주위의 환경 또는 다른 영역에 더 집중 분포되어 있다는 사실을 통해 사용자의 시선 분포를 히트맵 주위를 응시할 수 있도록 피드백을 반영하여 영상을 수정할 수도 있다.

앞서 언급한 대로 시청자가 특정 영역을 반드시 시청하도록 하려는 제작자의 의도가 있다면 Saliency Map 대신 임의의 heat map을 형성한 다음 시청자가 얼마나 제작자의 의도대로 시청하고 있는지 평가할 수 있다. 이러한 보다 객관적인 평가를 토대로 제작자는 제작 의도에 부합되는 영상으로 수정해 나갈 수 있을 것이다.

3) 감정 반응 점수(Emotion Engagement Score)

감정 반응 점수(Emotion Engagement Score)는 영상에 감정적으로 반응하는 시청자들의 비율을 나타내는 점수이다. 그림 12에서는 감정 반응 점수 계산 방식을 제시한다. 이 과정에서는 인공지능에 기초한 감정 분류 모델(Emotion Classifier Model)을 사용한다[21].

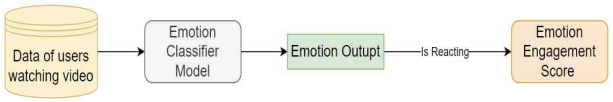


그림 12. 감정 반응 점수를 구하는 과정에 대한 순서도

Fig. 12. Flow chart for the process of calculating Emotion Engagement Score

```
label: neutral
probability: [
        'neutral': 0.28702497482299805,
        'happy': 0.10719223320484161,
        'sad': 0.044058218598365784,
        'surprise': 0.23802989721298218,
        'fear': 0.02736002951860428,
        'disgust': 0.023419566452503204,
        'anger': 0.022784020751714706,
        'contempt': 0.2501310408115387
]
```

그림 13. 감정 추론 모델의 출력값

Fig. 13. Output value of the Emotion Classifier Model

감정 분류 모델은 사람 얼굴 이미지로 감정을 추론하는 인공지능 모델이다. 이 논문에서 수행한 실험에서는 사용자가 비디오를 시청하고 있는 모습을 촬영한 영상의 모든 프레임을 모델 입력값으로 적용한다. 모델 입력 후 출력값은 그림 13과 같이 'neutral', 'happy', 'sad', 'surprise', 'fear', 'disgust', 'anger', 'contempt' 총 8가지의 감정에 대한 각각의 확률값으로 표시되며 보통 이 중 가장 높은 확률을 가진 감정이 현재 감정으로 판정된다.

제안하는 방식에서는 특정한 감정에 대한 평가 없이 사용자의 영상에 대한 반응 여부만을 알기 위해 모델을 통해 출력된 확률에서 감정의 변화가 있는 상황만 판단한다. 이러한 목적을 위해 neutral을 제외한 나머지 7개 감정의 확률값을 모두 합한 값을 시청자가 영상에 반응하는 확률 P_{react}로 계산한다. 즉,

$$P_{react} = \sum_{e \in Z} P_e = 1 - P_{\text{neutral}} \quad (1)$$

$Z=\{$'happy', 'sad', 'surprise', 'fear', 'disgust', 'anger', 'contempt'$\}$

이렇게 특정 감정을 세부적으로 평가하지 않고 단순하게 감정의 변화만을 포착하는 이유는 아직까지 얼굴 표정만으로 감정을 인식하는 인공지능 기술의 정밀도(accuracy)가 SOTA(State-Of-The-Art)를 달성한 기술마저도 66% 정도

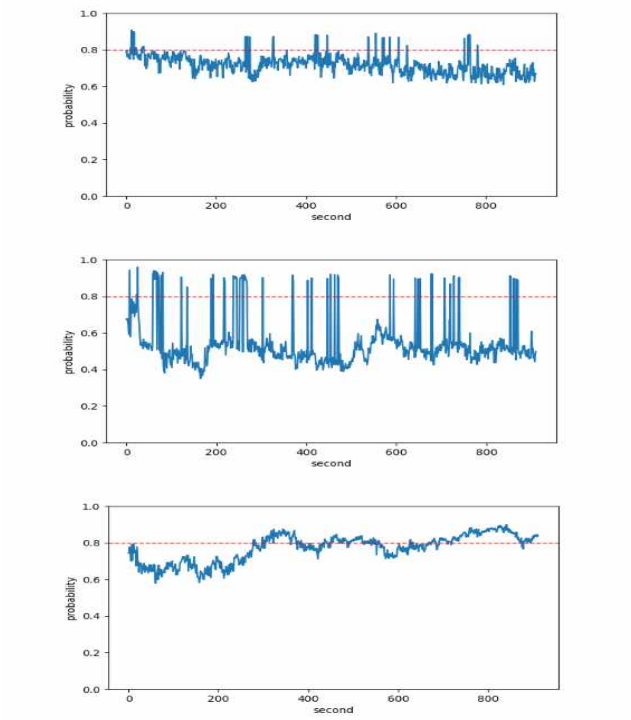


그림 14. 사용자 3명의 감정 반응 확률을 나타낸 그래프
Fig. 14. Graph showing the probability of three users' emotional reaction

수준에 머무르고 있기 때문이다[22]. 따라서 본 연구에서는 보다 정밀한 평가를 위해 구체적인 감정을 구분하지 않고 감정의 변화만을 파악하기 위해서 식 (1)에서 제시한 확률 계산 방식을 적용한다. 향후 감정 인식 인공지능 기술이 발전한다면 영상 콘텐츠로 인한 보다 정교한 감정 평가가 가능할 것이다. 그림 14에서는 3명의 시청자가 영상에 대해 반응하는 확률(reacting probability) 결과값을 제시한다. 각 시청자의 감정 반응에서 볼 수 있듯이 감정 반응 점수는 사용자들의 개인적인 성격이나 취향 등에 따라서 비교적 큰 편차를 보임을 확인할 수 있다.

본 논문의 실험에서는 기준값(threshold)을 0.6으로 하여 각각의 frame에 대한 영상에 대해 반응하는 확률(reacting probability)이 기준값보다 높다면 해당 프레임에서 반응하고, 작다면 해당 프레임에서 반응하지 않는다고 판단하였다. 그림 15에서는 동일한 영상을 시청한 20명의 데이터를 이용하여 감정 반응 점수를 계산하고 반응 여부를 표현한다.

또한 초 단위를 기준으로 시청자가 영상에 반응하는지에 대한 여부를 판단하기 위해 사용자가 영상을 시청하는 모습을 촬영한 영상의 초당 프레임인 30fps를 기준으로 하여 절반 이상의 프레임에서 시청자가 영상에 반응한다면 해당 초에는 시청자가 영상으로 인한 감정적인 반응을 유발했다고 판단한다. 감정 반응 점수는 다른 두 개의 지표에 비해서 아직 초기 단계라고 할 수 있는 얼굴 이미지 기반 감정 인식 기술[22], [23]의 영향을 많이 받아서 성능 상의 한계가 있고,

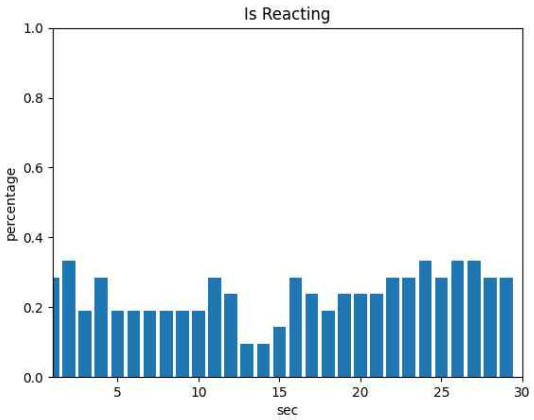


그림 15. 20명의 사용자에 대한 감정 반응 점수 그래프
Fig. 15. Graph of Emotion Engagement Score for 20 users

감정 반응이 반드시 현재 재생 중인 영상으로 인한 것이라고 보기 힘든 측면도 있기 때문에 독립적으로 의미있는 지표로 활용하기에는 한계가 있다. 그러나 나머지 두 지표와 함께 적용한다면 영상 콘텐츠 품질을 향상시킬 수 있는 보다 객관적인 도구로 활용 가능하리라 생각한다.

Ⅲ. 결론

미디어 플랫폼의 발전과 방대하고 다양한 영상 콘텐츠가 제작되는 시대의 흐름에 따라 제작자들은 사용자의 관점에서 영상 콘텐츠의 품질을 평가할 수 있는 객관적인 방법이 필수적이다. 기존 방식대로 사람의 주관적인 견해만을 바탕으로 영상 품질을 평가하게 되면 많은 시간적인 비용 소요 및 영상 품질에 대한 낮은 신뢰도를 얻게 된다. 따라서 본 논문에서는 사용자의 시선 및 감정 분석을 이용한 보다 객관적인 영상 품질 평가 방법을 제안한다.

시선을 이용한 평가 방식은 본 논문 외에도 현재 마케팅, 광고, 심리학 등 다양한 연구 분야에서 활용되고 연구되고 있다[11],[24]. 기존에는 시선을 추적하기 위해 시선 추적 장비를 추가적으로 부착하는 형태를 통해 시선을 분석하였으나 본 연구에서는 시선 추적 장비를 추가적으로 사용하지 않고 일반 RGB 카메라만으로 시선 추적이 가능한 VisualCamp사의 SeeSo SDF를 활용해 기존 장비에 있는 카메라만으로 시선을 추적할 수 있도록 하였다. 이는 사용자의 동의만 있다면 손쉽게 사용자의 시선 데이터를 확보하고 분석할 수 있다는 것을 의미한다.

시선 데이터는 객관적이기 때문에 정확도가 떨어지는 사용자의 주관적인 견해를 보완할 수 있다. 본 논문에서 제시한 방식을 통해 사용자의 시선이 집중적으로 분포되어 있는 부분 또는 어느 시점에서 이탈자가 발생하는 지에 대한 분석 결과를 통해 영상 제작자는 영상 품질을 보완하여 사용자가 영

상에 끝까지 집중할 수 있도록 할 수 있을 것으로 기대한다. 동시에 영상에 대한 사용자의 감정분석을 통해 영상 제작자의 의도대로 사용자가 반응하는지 확인하기 위한 객관적인 도구로서도 활용 가치가 높을 것으로 기대한다.

영상에 대한 사용자의 반응을 통계적인 방식으로 객관적으로 검증하기 위해서는 여러 사용자들의 데이터들이 필수적이다. 그러나 기존의 시선 추적 방식에서는 고가의 전용 장비가 필수적이었기 때문에 참여하는 사용자 수가 제한적이고 특정 지역에 국한될 수 밖에 없는 한계가 있다. 이와는 달리, 제안하는 방식은 일반적인 스마트폰이나 태블릿, 웹캠이 설치된 일반 PC에서 소프트웨어만 설치하면 전세계를 대상으로 한 번에 대규모 사용자 실험을 수행할 수 있기 때문에 시간적, 공간적 제약에서 자유롭다. 따라서 통계적 타당성 확보를 위해 필수적인 대규모 실험 데이터 확보가 가능하다는 장점이 있다.

감사의 글

본 연구는 한국연구재단을 통해 과학기술정보통신부의 기초연구사업으로부터 지원받아 수행되었습니다(과제번호-2021R1A4A502890711).

참고문헌

[1] A. R. Park and J. N. Lee, “The Influence of Beauty Influencer’s Characteristics on Makeup Behavior and Color Cosmetics Purchase Intentionin Young Female Consumers aged 20-30s,” *Journal of the Korean Applied Science and Technology*, Vol. 38, No. 4, pp. 1093-1106, August 2021. https://doi.org/10.12925/jkocs.2021.38.4.1093

[2] J. Lee and M. Han “The Impact of Individual Political Inclinations on the Reliability of News Comments and Perception of Social Influence,” *The Journal of Society for e-Business Studies*, Vol. 17, No. 1, pp. 173-187, February 2012. http://dx.doi.org/10.7838/jsebs.2012.17.1.173

[3] H. Shin, S. Lee, G. Son, H. Kim, and Y. Kim, “Integrated Verbal and Nonverbal Sentiment Analysis System for Evaluating Reliability of Video Contents,” *KIPS Transactions on Software and Data Engineering*, Vol. 10, No. 4, pp. 153-160. April 2021. https://doi.org/10.3745/KTSDE.2021.10.4.153

[4] S. Kim and S. Yang. “A Comparative Analysis between General Comments and Social Comments on an Online News Site,” *The Journal of the Korea Contents Association*, Vol. 15, No. 4, pp. 391-406, April 2015. http://dx.doi.org/10.5392/JKCA.2015.15.04.391

[5] YouTube Help Center. Measuring Key Moments in Viewing Duration [Internet]. Available: https://support.google.com/ youtube/answer/9314415?hl=ko

[6] S. Cho, “An introduction to eye tracking technology,” *Electronic Engineering Journal*, Vol. 45, No. 8, pp. 23-32, 2018.

[7] E. Lindén, J. Sjöstrand, and A. Proutiere, “Learning to Personalize in Appearance-Based Gaze Tracking,” in *Proceeding of IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Seoul, pp. 1140-1148, September 2019. https://doi.org/10.1109/ICCVW.2019.00145

[8] Z. Wu, S. Rajendran, T. V. As, V. Badrinarayanan, and A. Rabinovich, “EyeNet: A Multi-Task Deep Network for Off-Axis Eye Gaze Estimation,” in *Proceeding of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW),* Seoul, pp. 3683–3687, September 2019. https://doi.org/10.1109/ICCVW.2019.00455

[9] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, “Eye Tracking for Everyone,” in *Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, pp. 2176-2184, June 2016. https://doi.org/10.1109/CVPR.2016.239

[10] M. Zhang, Y. Liu, and F. Lu, “GazeOnce: Real-Time Multi-Person Gaze Estimation,” in *Proceeding of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, pp. 4187-4196, June 2022. https://doi.org/10.1109/CVPR52688.2022.00416

[11] H. Kim, An Analysis of visual effects of online banner advertisement with application of eye tracking, Master’s thesis, Seoul National University Graduate School of Industry, 2010.

[12] J. Park, H. Kim, and H. Kwon. “An Eye-tracking Study of Virtualgraph Contents in Mobile Commerce,” *Journal of Digital Contents Society*, Vol. 22, No. 10, pp. 1653-1659, October 2021. https://doi.org/10.9728/dcs.2021.22.10.1653

[13] Seeso SDK, VisualCamp [Internet]. Available: https://visual.camp/

[14] EmotionNet, NVIDIA [Internet]. Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/emotionnet

[15] C. Guo and L. Zhang, “A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression,” *IEEE Transactions on Image Processing*, Vol. 19, No. 1, pp.

185-198, January 2010.
https://doi.org/10.1109/TIP.2009.2030969

[16] K. Min and J. J. Corso, “TASED-Net: Temporally -Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection,” in *Proceeding of the IEEE/CVF International Conference on Computer Vision*, Seoul, pp. 2394-2403, October 2019.
https://doi.org/10.1109/ICCV.2019.00248

[17] J. S. Kim, “Physiology of Eye Movements,” *Annals of Clinical Neurophysiology*, Vol. 1, No. 2, pp. 173-181, 1999.

[18] E. Seo, “Mobile Eye Tracker and for Use of the Same for Revitalizing Studies on Eye Tracking,” *The Journal of the Korea Contents Association,* Vol. 16, No. 12, pp. 10-18, December, 2016.
https://doi.org/10.5392/jkca.2016.16.12.010

[19] Mustard Official Channel. [Internet]. Available: https://youtu.be/pI14GjQNAgc

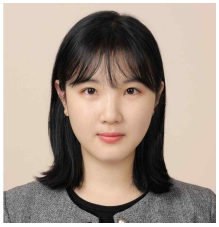
[20] I. H. Yang, S. M. Lim, and Y. H. Kim, “Investigation of Eye-Tracking on Learning Task Perceiving Process of Elementary Students with Different Motivation System on Science Learning,” *Journal of Korean Elementary Science Education*, Vol. 34, No. 1, pp. 86-94, February 2015. https://doi.org/10.15267/KESES.2015.34.1.086

[21] Z. Wen, W. Lin, T. Wang, and G. Xu, “Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition,” arXiv, November 2022. https://doi.org/10.48550/arXiv.2109.07270

[22] Facial Expression Recognition on AffectNet (SOTA) [Internet]. Available: https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet

[23] A. V. Savchenko, L. V. Savchenko, and I. Makarov, “Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network,” *IEEE Transactions on Affective Computing*, Vol. 13, No. 4, pp. 2132-2143, October 2022. https://doi.org/10.1109/TAFFC.2022.3188390

[24] Z. Y. Wang and J. Y. Cho, “Older Adults' Response to Color Visibility in Indoor Residential Environment Using Eye-Tracking Technology,” *Sensors*, Vol. 22, No. 22, pp. 8766, November 2022. https://doi.org/10.3390/s22228766



곽수찬(Soochan Kwak)

2020년～현 재: 덕성여자대학교 사이버보안전공 학사 과정

※관심분야 : Machine Learning, Deep Learning

김지윤(Jiyun Kim)

2020년～현 재: 덕성여자대학교 사이버보안전공 학사 과정

※관심분야 : Machine Learning, Deep Learning

박태정(Taejung Park)

1997년 : 서울대 전기공학부 (공학사)
1999년 : 서울대 전기공학부 대학원 (공학석사, 반도체 물리 전공)
2006년 : 서울대 전기컴퓨터공학부 대학원 (공학박사, 컴퓨터 그래픽스 전공)
2006년～2013년: 고려대학교 연구교수
2013년～2017년: 덕성여자대학교 정보미디어대학 디지털미디어학과 조교수
2018년～현 재: 덕성여자대학교 공과대학 사이버보안/IT미디어공학과 부교수
※관심분야 : 컴퓨터그래픽스, 인공지능, 수치해석, 3차원 모델링