시선추적-뇌파 기반의 비디오 요약 생성 방안 연구*

Video Summarization Using Eye Tracking and Electroencephalogram (EEG) Data

김 현 희 (Hyun-Hee Kim)**
김 용 호 (Yong-Ho Kim)***

목 차

초 록

본 연구는 시선 및 뇌파 정보를 이용하여 오디오-비주얼(audio-visual, AV) 시맨틱스 기반의 동영상 요약 방법들을 개발하고 평가해 보았다. 이를 위해서 27명의 대학생들을 대상으로 시선추적과 뇌파 실험을 수행하였다. 평가 결과, 뇌파와 동공크기 데이터를 함께 사용한 방법의 평균 재현율(0.73)이 뇌파 또는 동공크기 데이터만을 사용한 방법의 평균 재현율(뇌파: 0.50, 동공크기: 0.68)보다 높게 나타났다. 또한 AV 시맨틱스 기반의 개인화된 동영상 요약의 평균 재현율(0.57)이 AV 시맨틱스 기반의 일반적인 동영상 요약의 평균 재현율(0.69)보다 낮게 나타난 원인들을 분석하였다. 끝으로, AV 시맨틱스 기반 동영상 요약 방법과 텍스트 시맨틱스 기반 동영상 요약 방법 간의 차이 및 특성도 비교분석해 보았다.

ABSTRACT

This study developed and evaluated audio-visual (AV) semantics-based video summarization methods using eye tracking and electroencephalography (EEG) data. For this study, twenty-seven university students participated in eye tracking and EEG experiments. The evaluation results showed that the average recall rate (0.73) of using both EEG and pupil diameter data for the construction of a video summary was higher than that (0.50) of using EEG data or that (0.68) of using pupil diameter data. In addition, this study reported that the reasons why the average recall (0.57) of the AV semantics-based personalized video summaries was lower than that (0.69) of the AV semantics-based generic video summaries. The differences and characteristics between the AV semantics-based video summarization methods and the text semantics-based video summarization methods were compared and analyzed.

키워드: 뇌파, 시선추적, 비디오 요약, 오디오-비주얼 시맨틱스, 텍스트 시맨틱스
EEG, Eye Tracking, Video Summarization, Audio-visual Semantics, Text Semantics

* 이 논문은 2020년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A01040945).
** 명지대학교 문헌정보학과 명예 교수(kimhh@mju.ac.kr / ISNI 0000 0004 6508 2465) (제1저자)
*** 부경대학교 신문방송학과 명예 교수(yh1228kim@gmail.com / ISNI 0000 0004 6481 7343) (교신저자)
논문접수일자: 2022년 1월 17일 최초심사일자: 2022년 2월 12일 게재확정일자: 2022년 2월 22일
한국문헌정보학회지, 56(1): 95-117, 2022. http://dx.doi.org/10.4275/KSLIS.2022.56.1.095

1. 서 론

1.1 연구 배경과 목적

최근 증가하고 있는 소셜 미디어 플랫폼, 인터넷 검색엔진, 디지털 도서관 내의 방대한 동영상 콘텐츠를 검색하고 이를 계속적으로 활용하기 위해서는 자동적으로 개인화된 동영상 요약을 생성할 수 있는 방안이 요구된다. 이러한 개인화된 동영상 요약은 이용자들이 이미 시청한 동영상 콘텐츠를 나중에 다시 보거나 재활용할 수 있도록 하기 위한 정보를 제공한다(Moon et al., 2019).

영상물 주제와 밀접하게 관련된 쇼트들(shots)을 추출하여 영상물 요약을 생성하기 위해서 시청자가 비디오를 시청할 때 비디오 쇼트로부터 의미 정보를 어떻게 수용하는지 조사해 보는 것이 필요하다. 비디오 쇼트는 비주얼 컨텐츠와 오디오 콘텐츠를 포함하고 있다. 비주얼 콘텐츠는 비주얼 요소(예, 객체)와 캡션 텍스트(예, 자막)를 포함하며(Kaur & Neeru, 2015), 오디오 콘텐츠는 스피치, 대화 및 내레이션(narration) 등과 같은 입말 텍스트(spoken text)를 포함한다. 좀 더 구체적으로 시청자가 어떻게 동영상 콘텐츠에 있는 의미 있는 비주얼 요소, 갭션 텍스트 및 입말 텍스트를 파악하고 이들 간의 상호관계(예, 비주얼 요소 간의 관계, 비주얼 요소와 입말 텍스트 간의 관계)를 분석하는지 알아야 한다. 이와 같은 오디오-비주얼(audio-visual, AV) 시맨틱스에 기반한 비디오 요약을 구성하기 위해서 신경과학적 방법인 시선 및 뇌파 반응을 이용할 수 있다.

최근 연구들은 시선추적 또는 뇌파 정보를 미디어 요약과 정보검색에 활용하는 방안을 제안하고 있다. Bhattacharya et al.(2020)은 텍스트 문헌의 시선 흐름 패턴이 주제 관련성에 따라서 달라진다는 사실을 이용하여 문헌을 판독할 때의 시선 흐름(scan path)을 이미지로 변환한 후 이러한 이미지들을 합성곱 신경망(convolutional neural network, CNN)으로 분류하여 문헌의 주제 관련성을 예측하였다. 한편 Kim과 Kim(2019a, 2019b)은 뇌파(electroencephalography, EEG)의 사건관련유발전위(Event Related Potentials, ERP)의 P600, N400 컴포넌트들을 이용하여 비디오 요약을 생성할 수 있다고 제안하였다. 구체적으로, 피험자가 비디오를 시청할 때 주제관련 쇼트는 자극 이후 600ms 근처에서 양전위 최대값이 나타나는 P600 효과를 보이고, 주제무관 쇼트는 400ms 근처에서 음전위 최소값이 나타나는 N400 효과를 보인다고 보고하였다.

뇌파 또는 시선추적 정보만을 사용한 경우에는 제한점들이 있다. 시선 정보는 시청자가 관심을 두는 객체들을 파악할 수 있게 하고 시청자의 시선 흐름을 분석하여 그 대상에 대한 반응을 측정할 수 있다. 그러나 시선 정보는 시청자가 특정 대상에 초점을 맞춰서 어떤 인지 작업을 하는지 정확히 알아낼 수 없다. 시선 고정의 횟수와 동공크기의 변화만으로 대상에 집중하여 사고를 하고 있다고 볼 수 없다. 다시 말해서 시각적 인지만으로는 인지된 정보가 이용자의 작업 기억에 머무르며 장기 기억과 통합되는 인지과정이 일어나는지에 대해서 정확히 파악하기 어렵다는 것이다(안형모, 2013).

한편 뇌파 정보는 이용자의 인지과정과 집중상태를 알려주지만 동영상에 대한 뇌파 반응이 개인별로 또는 측정 환경에 따라서 차이가 날

수 있다. 또한 각 동영상에서 비주얼 컨텐츠가 동시에 제시되는 경우 시청자가 어떤 비주얼 요소 또는 어떤 캡션 텍스트를 살펴보는지 더 나아가 이들 간의 관계를 어떻게 파악하는지에 대한 정보를 수집할 수 없다.

시선추적 및 뇌파 정보를 함께 사용한다면 이러한 제한점을 개선시키는 효과가 있을 것이다. 시선 정보는 시청자가 동영상을 시청하는 동안 동공의 변화, 시선 흐름 등을 분석하여 뇌파 측정 방법의 제한점을 보완할 수 있게 한다. 한편 뇌파 정보는 인지과정에 대한 정보를 제공하여 시선 정보의 부족한 점을 보완할 수 있게 할 수 있을 것이다.

본 연구의 목적은 시선 및 뇌파 정보를 이용하여 AV 시맨틱 기반의 동영상 요약물을 제작하는 방법을 제안하고 평가한 후 연구 결과를 개인별 맞춤형인 개인화된(personalized) 미디어 요약과 디지털 도서관의 메타데이터인 일반적인(generic) 미디어 요약의 구축에 활용하는 데에 있다.

이를 위해서 첫째, 뇌파와 시선추적의 패턴을 이용하여 주제관련 쇼트들을 추출할 수 있는 모형들을 개발하였다. 둘째, 개발된 모형들을 기반으로 하여 동영상 요약을 구성하였다. 셋째, 구성된 동영상 요약을 평가하기 위해서 전문가에 의해서 구성된 동영상 요약(ground truth method, 이하 '표준 동영상 요약'이라 함)과 비교하여 재현율을 측정해 보았다. 넷째, AV 시맨틱스에 기반한 일반적인 동영상 요약과 개인화된 동영상 요약 간에 어떤 차이가 있는지 알아보았다. 마지막으로, AV 시맨틱스에 기반하여 구성된 동영상 요약 방법과 동영상의 오디오 콘텐츠로부터 추출한 단어 또는 구절의 의미와 중요성을 이끌어내는 텍스트 시맨틱스에 기반한 동영상 요약 방법 간에 어떤 차이가 있는지 비교분석해 보았다.

1.2 용어 정의

(1) 뇌파(electroencephalography, EEG): 대뇌 피질의 활동을 반영하는 두피의 표면에서 전기 패턴을 측정한 것으로, 일반적으로 "뇌파"라고 한다(Hughes & John, 1999). EEG는 사건관련유발전위(event related potentials, ERP)와 푸리에 변환을 통해 얻어진 각 주파수 영역을 기준 축으로 각 주파수별 정량적인 수치로 표현된 정량화 뇌파(Quantitative EEG, QEEG)로 구분된다.

(2) ERP: 인간의 두뇌 외피에 전극을 부착하여 외부의 자극 또는 내부의 심리적 과정과 관련되어 나타나는 뇌의 반응으로 일정한 시간 동안의 전기적 활동을 의미한다. ERP의 유형으로는 P3b, N400, P600 등이 있다.

(3) N400: 자극 제시 이후 400ms 근처(400ms~600ms)에서 전전두엽, 전두엽, 중앙엽, 두정엽, 측두엽 등에서 음전위 최저 뇌파값이 나타나는 ERP 컴포넌트이다.

(4) P600: 자극 제시 이후 600ms 근처(500ms~700ms)에서 전전두엽, 전두엽, 중앙엽, 두정엽, 측두엽 등에서 양전위 최고 뇌파값이 나타나는 ERP 컴포넌트이다.

(5) ET P3b: 본 연구에서 제안한 용어로 자극 제시 이후 300ms 근처(200ms~500ms)에서 최고 동공크기값을 나타낼 것으로 가정한 시선추적(Eye Tracking, ET) 컴포넌트이다.

(6) 잠재의미분석(latent semantic analysis,

lsa): 문헌과 단어들간의 표면적인 관계가 아니라 문헌 내에 존재하는 단어들에 내포된 숨겨진 주제별로 그룹들을 만들며 그룹화된 단어와 문장 간의 관계를 분석하여 문장에 점수를 부여하는 방법이다(정영미, 2005).

(7) 페이지랭크 중심성(PageRank centrality): 구글의 검색 알고리즘에 쓰이는 네트워크 분석 방법으로 해당 노드(예, 단어)의 연결정도 중심성으로부터 발생하는 영향력과 관계된 노드들의 영향력을 고려하여 중심성을 결정하는 방법이다.

1.3 연구 방법

첫째, 뇌파 모형은 저자들의 기존 뇌파 연구들의 실험 데이터를 분석하여 구현하였다.

둘째, 시각 모형을 구현하고 시각 및 뇌파 모형의 유용성을 평가하기 위해서 27명의 피험자들을 이용하여 시선추적 및 뇌파 실험을 수행하였다.

셋째, 뇌파 및 시선 모형을 이용하여 AV 시맨틱 기반 동영상 요약들(개인화된 요약, 일반적인 요약)과 텍스트 시맨틱스 기반 동영상 요약들(일반적인 요약)을 구성하여 평가해 보았다.

2. 선행 연구와 연구문제

2.1 선행 연구

뇌파 또는 시선추적이 정보 검색과 미디어 요약에 활용된 선행 연구들을 기술한 후 뇌파와 시선추적 정보를 함께 적용한 연구들에 대해서 기술한다.

N400 컴포넌트는 의미가 일치하지 않은 언어 정보에 반응하는 ERP 유형으로 알려져 있다(van Berkum, Hagoort, & Brown, 1999). 한편 P600 컴포넌트는 새로운 자극이 유입되었을 때 나타나며, 담화 구조의 유지 및 갱신(Schumacher & Hung, 2012), 담화의 내적인 재조직과 통합(Wang & Schumacher, 2013) 등과 관련되어 있는 ERP 유형이다. Kim과 Kim (2019a, 2019b)은 언어 분야에 활발히 활용되고 있는 P600/N400 컴포넌트들을 동영상 요약에 적용하였다. 저자들은 P600 컴포넌트는 주제와 관련된 비디오 쇼트의 맥락 갱신에 적용될 수 있고, N400 컴포넌트는 주제와 무관한 비디오 쇼트의 의미적 불일치에 적용될 수 있다고 보고하였다.

시선추적 정보를 적합성과 관련시키는 연구들을 살펴보면 Ajanki et al. (2009)은 암묵적 적합성 피드백 시스템에서 추가적인 질의어들을 선택하기 위해서 시선추적 정보를 이용하였는데 회귀와 첫 번째 시선 고정이 가장 유용한 요인들이었다. Oliveira, Aula 및 Russell (2009)의 연구에 의하면 텍스트 및 이미지 웹 탐색 결과의 적합성을 결정하는 데에 동공크기의 변화가 가장 중요한 요인으로 나타났다. Katti et al.(2011)은 스토리보드 또는 비디오 요약을 생성하기 위해 시청자의 관심과 참여를 파악하기 위한 방법으로 동공 크기를 사용했다. Gwizdka와 Zhang (2015)은 피험자들이 관심을 갖는 웹페이지들을 방문할 때 관심을 갖지 않는 웹페이지들을 볼 때 보다 동공크기가 확대된다는 것을 발견했다.

Bhattacharya와 Gwizdka(2018)는 시선추적 장치를 이용하여 피험자들의 지식 변화의 차이가 텍스트 읽기와 관련된 시선 흐름과 관련된다고 가정하였다. 연구 결과, 지식 변화의 차이가 큰 피험자들과 크지 않는 피험자들을 비교해 볼 때 텍스트를 읽을 때 시선 고정 횟수와 시간에서 차이가 나타났다. Bhattacharya et al.(2020)은 피험자가 주제와 관련된, 부분적으로 관련된 또는 무관한 텍스트 문헌(뉴스)을 판독할 때 서로 다른 시선 흐름 패턴을 보인다고 가정하고, 이러한 시선 흐름 패턴을 이미지로 변환하였다. 합성곱 신경망(CNN)을 시선 흐름 이미지에 적용하여 최대 80% 정확도로 문헌의 관련성을 예측할 수 있었다.

Syn과 Yoon(2021)은 시선추적과 인지 결과를 이용하여 대학생들의 개인 및 건강 관련 특성이 Facebook 건강 정보의 읽기 행태, 인지적 결과와 어떤 관련이 있는지 살펴보았다. 저자들은 인지 결과와 시선고정 횟수와는 명확한 패턴은 없었으나 페이스북 게시물 읽기와 관심 영역(area of interests)과는 명확한 패턴이 있다고 기술하였다. 윤정원과 신수연(2021)은 시선추적과 회상 인식 테스트를 이용하여, Facebook 건강 정보 게시물 형식이 대학생들의 페이스북 게시물 시선주시패턴과 인지 테스트 결과에 미치는 영향을 분석하였다. 저자들은 이용자들이 정보를 포함하고 있는 영역에 주의를 집중하며, 정보를 포함하지 않는 사진보다는 본문의 내용에 먼저 시선을 집중하는 등의 결과를 보고하였다.

Gwizdka et al.(2017)은 탐색과정에서 텍스트 읽기 및 적합성 판정의 과정을 뇌파와 시선추적 기술을 이용하여 분석하였다. 저자들은 탐색 과정에서 다양한 적합성 수준의 텍스트를 평가하는데 이용된 인지 과정에서의 차이와 이러한 차이를 발견할 수 있는 근거를 발견했다. 또한 Golenia et al.(2018)은 EEG와 시선추적의 조합을 통해 이미지 검색에서 암묵적 적합성 피드백의 가능성을 조사하였다. 저자들은 EEG와 시선추적 정보를 결합했을 때의 정확도(85.9%)가 단일 측정방식의 정확도(시선 추적: 81.0%, EEG: 76.9%)보다 높게 나타났다고 보고하였다. Bezugam et al.(2021)은 정량화 뇌파(QEEG)와 시선추적(시선 도약, 시선 고정) 데이터를 이용하여 감시 비디오(surveillance video)의 키프레임들을 추출하는 비디오 요약 방안을 제안하였다. 표준 동영상 요약과 비교한 결과, 제안된 방식에 의한 동영상 요약의 정확률이 한 방법만 사용한 경우의 정확률보다 더 높게 나타났다.

앞에서 살펴본 선행 연구들은 비디오 요약을 구성하는 데에 ERP의 P600/N400 컴포넌트들의 중요성을 제안하고 있지만 실제 데이터에 이들 컴포넌트를 적용한 구체적인 방안을 제시하지 않고 있다. 또한 동공 크기 데이터는 뇌파 데이터와 함께 적합성 판정에 효율적으로 응용되고 있는 것을 확인할 수 있었다. 따라서 본 연구는 뇌파와 동공크기 데이터를 활용하여 주제와 관련된 비디오 쇼트들을 추출하여 개인화된 비디오 요약과 일반적인 비디오 요약을 구성할 수 있는 구체적인 방안들을 제안하고자 한다.

2.2 연구문제

1) 연구문제 1: 뇌파와 시선추적 데이터를 함께 사용하면 뇌파나 시선추적 데이터만을

사용한 경우보다 표준 동영상 요약과의 재현율(일치도)을 향상시킬 수 있는가?

2) 연구문제 2: 오디오-비주얼(AV) 시맨틱 기반의 일반적인 동영상 요약과 개인화된 동영상 요약 간에 어떤 차이가 있는가?

3) 연구문제 3: 오디오-비주얼(AV) 시맨틱 기반 동영상 요약 방법과 텍스트 시맨틱스 기반 동영상 요약 방법 간에 어떤 차이가 있는가?

3. 이론적 배경

3.1 오디오-비주얼(AV) 시맨틱스와 텍스트 시맨틱스

AV 시맨틱스는 비주얼 요소와 캡션 텍스트로 구성된 AV 콘텐츠의 의미와 중요성을 파악하는 분야이다(Foley & Kwan, 2015). 본 연구는 AV 시맨틱스에 기반한 비디오 요약을 구성하기 위해서 뇌파와 시선추적을 이용하였다. 한편 텍스트 시맨틱스는 텍스트 콘텐츠에 있는 단어와 구절의 의미와 중요성을 파악하는 분야이다. 본 연구에서 텍스트 콘텐츠는 오디오 트랙에서 추출한 내레이션, 스피치 및 대화의 글말 텍스트(written text)라 할 수 있는 트랜스크립(transcript)을 지칭하며, 텍스트 시맨틱스에 기반한 비디오 요약을 구성하기 위해서 잠재의미분석, 페이지랭크 중심성을 이용하였다.

3.2 오디오-비주얼(AV) 시맨틱스 기반 모형

3.2.1 뇌파 모형

P600/N400 컴포넌트에서 중요한 채널들을 선정하기 위해서 저자들이 2016년~2019년 동안 수행한 뇌파 실험들의 데이터(66 사례, 〈표 1〉 참조)를 이용하였다(김현희, 김용호, 2016; 김현희, 김용호, 2019; Kim & Kim, 2019b). 이후 SPSS 23의 인공 신경망(Artificial Neural Network, ANN)을 활용하여 상기 P600/N400 컴포넌트 분석에 의해서 선정된 채널들의 가중치를 구한 후 P600_uV 공식(1)과 N400_uV 공식(2)를 이용하여 뇌파 모형들을 구성하였다.

1) P600 컴포넌트를 활용한 모형

(1) 채널 선정: 본페로니 교정($p=0.0017$) 기준에 의해서 주제관련 양전위 최고점의 평균값과 주제무관 양전위 최고점의 평균값 간에 유의미한 차이가 나는 채널은 총 12개(cp4, c4, tp8, t8, cz, fc4, ft8, fcz, fz, f4, f8, fp2)로 나타났다.

(2) P600_uV 모형: 실험 데이터를 인공 신경망(SPSS 사용)으로 분석하여 상기 12개 채널의 가중치를 구한 후 공식(1)을 구성하였다.

〈표 1〉 실험 데이터

	피험자 수	실험 데이터	사례 수(총 66)
2016년	26명	비디오 6개(통합)	26
2019년 1차	22명	비디오 4개(통합)	22
2019년 2차	18명	비디오 1개	18

이를 위해서 P600 컴포넌트에서 실험 데이터를 인공 신경망을 이용하여 분류해 보았다. 인공 신경망에 의한 여러 분류 결과들 중에서 학습 및 검정 데이터의 평균 분류 정확도가 높은 경우 즉 77.0%(학습 데이터), 90%(검정 데이터)를 선정하였다. 이때 총 132 사례(66 사례 X 2 [주제관련, 주제무관]) 중 학습으로 120개가 사용되었고, 12개가 검정에 사용되었다. 이후, 아래와 같이 상기 학습 및 검정 데이터의 분류 결과를 산출할 때 사용된 12개 채널의 가중치에 기초하여 공식(1)을 구성하였다.

P600_uV = FC4 X 0.145 + F8 X 0.131 + T8 X 0.119 + FCz X 0.088 + C4 X 0.075 + FT8 X 0.074 + F4 X 0.07 + Fz X 0.069 + TP8 X 0.064 + CP4 X 0.06 + FP2 X 0.059 + Cz X 0.043 …… (1)

2) N400 컴포넌트를 활용한 모형

(1) 채널 선정: 본페로니 교정 기준에 의해서 주제관련 음전위 최저점의 평균값과 주제무관 음전위 최저점의 평균값 간에 통계적으로 유의미한 차이가 나는 채널은 총 15개(p4, CP4, C4, TP8, T8, CPZ, CZ, FC4, FT8, FCZ, FZ, F4, F8, FC3, fp2)로 나타났다.

(2) N400_uV 모형: N400 컴포넌트에서 채널들의 가중치를 산출하기 위해서 실험 데이터를 인공 신경망을 이용하여 분류해 보았다. 인공 신경망에 의한 여러 분류 결과들 중에서 학습 및 검정 데이터의 평균 분류 정확도가 높은 경우 즉 74.2%(학습 데이터), 80%(검정 데이터)를 선정하였다. 상기 P600_uV 모형과 동일한 방식으로 구성한 공식(2)는 다음과 같다.

N400_uV = T8 X 0.103 + FC4 X 0.088 + FT8 X 0.083 + F4 X 0.071 + F8 X 0.069 + Cz X 0.068 + FC3 X 0.067 + TP8 X 0.064 + Fz X 0.059 + P4 X 0.058 + C4 X 0.058 + CPz X 0.058 + FP2 X 0.055 + FCz X 0.054 + CP4 X 0.047 …… (2)

3.2.2 시각 모형

시선 데이터 분석을 위한 이론적인 틀을 개발하기 위해서 시선추적 실험을 수행하였다(자세한 내용은 “4.시선추적 및 뇌파 실험” 참조). 주제관련 쇼트들의 평균 동공크기와 주제무관 쇼트들의 평균 동공크기 간에 차이를 분석한 결과, 유의미한 차이가 나타나 동공 크기를 시각 모형의 기본 데이터로 사용하기로 결정하였다.

1) 시선 데이터 분석

피험자가 주제관련 쇼트, 주제부분관련 쇼트, 주제무관 쇼트를 볼 때 동공크기의 차이를 측정하기 위해서 동공크기의 기저값으로 각 피험자가 비디오를 시청하기 바로 전에 보여주는 회색 스크린을 보는 0.2초 동안의 평균 동공크기를 사용하였다. 이후 각 쇼트를 보는 처음 3초 동안의 동공크기를 0.1초 단위로 분석하였다. 각 0.1초 단위의 동공크기는 이 기저값과의 차이로 계산한 표준값을 이용하였다.

예를 들어서 피험자의 기저값이 2.82mm이고 처음 0.1초 동안의 동공크기의 평균값이 3.18mm이면 처음 0.1초 동안의 표준값은 0.36mm(=3.18-2.82)이 된다. 이와 동일한 방식으로 나머지 부분(0.2초 ~ 3초)의 표준값을 구한다. 〈그림 1〉은 비디오 쇼트를 보는 처음 3초 동안의 9명의 동공크기를 각각 구한 후 이들의 평균

동공크기값을 이용하여 그래프로 나타낸 것이다. x축은 시간(초)을 나타내고 y축은 동공크기(mm)의 표준값을 나타낸다. 동공크기의 변화도 뇌파의 변화와 같이 처음 1초 동안 주제무관과 주제관련 간에 차이를 보이는 것으로 나타났다. 즉, 주제관련을 나타내는 실선(파란색)은 강한 P3b의 효과를 보인 반면 주제무관을 나타내는 점선(연두색)은 낮은 P3b 효과를 보인 후 강한 N400 효과를 나타내는 것으로 보인다. 주제에 부분적으로 관련되어 있다고 판정된 파선(빨간색)은 대략적으로 주제관련과 주제무관 간의 사이에 위치되어 있다. 주제관련(평균 동공크기: 0.40)과 주제무관(평균 동공크기: 0.30) 간에 유의미한 차이가 나타났고(p=0.032), 주제부분관련의 평균 동공크기는 0.34이다.

2) ET P3b

주제관련 쇼트를 제시한 이후 300ms 근처에서 최고 동공크기값을 나타낼 것으로 가정하였다. 이에 따라서 뇌파 분석에 사용해 오고 있는 ERP의 P3b 컴포넌트를 시선추적에 적용하여 ET P3b(200ms~500ms)로 지칭하고 이를 각 쇼트의 주제관련과 주제무관 간의 동공크기 차이를 분석하기 위한 컴포넌트로 사용하였다.

4. 시선추적 및 뇌파 실험

4.1 피험자, 실험 자료 및 기기

뇌파는 성별과 나이에 따라서 차이가 있다고 알려졌다(Evans, Cui, & Starr, 1995). 따라서 시각 모형 구성과 시각 및 뇌파 모형의 유용성을 평가하기 위해서 피험자들은 20대의 오른손잡이 남자로 제한하여 대학교의 학부생 27명의 피험자들(1차 실험: 9명, 2차 실험: 18명)을 모집하고, 시선추적 및 뇌파 검사를 수행하였다(IRB 승인번호: MJU-2020-06-006-02). 실험 자료는 2개의 다큐멘터리 비디오들이다. 비

〈그림 1〉 시선 데이터

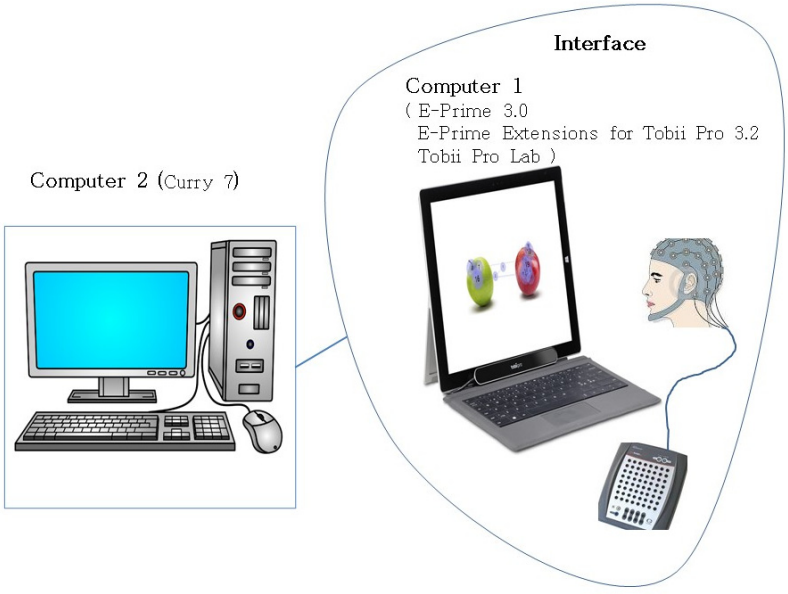
디오 1은 19개 쇼트로 구성된 김구의 회중시계로 재생시간은 1분 46초이며, 비디오 2는 8개 쇼트로 구성된 신라의 금관으로 재생시간은 58초이다.

실험 장비는 뉴로 스캔의 뇌파기기(Neuroscan Synamp Amplifier)와 토비테크놀로지의 시선추적 장치(Tobii Pro Nano)를 이용하였다. 피험자에게 캡(30채널 Quick Cap)를 장착시킨 후 동영상을 제시하고 시선추적과 뇌파 정보를 동시에 수집하기 위해서 컴퓨터 1에 인스톨된 E-Prime 3.0과 E-Prime Extensions for Tobii Pro 3.2를 이용하였다(〈그림 2〉 참조). 피험자는 컴퓨터 1을 인터페이스로 하여 동영상을 시청하며, 시청하면서 발생되는 뇌파 데이터는 컴퓨터 2의 Curry 7 소프트웨어에 의해서 수집되어 분석되었다. 한편 시선 데이터는 컴퓨터 1에 저장된 Tobii Pro Lab 소프트웨어에 의해서 수집되어 분석되었다.

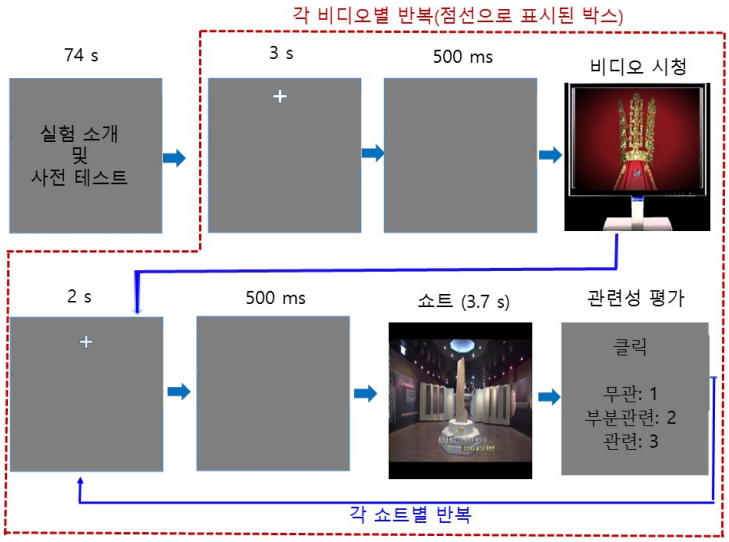
4.2 실험 절차

피험자들에게 실험 절차를 설명하였다. 이후 피험자의 눈의 위치를 교정하는 캘리브레이션(calibration) 과정을 거친 후 세 개의 쇼트로 구성된 짧은 비디오를 사용하여 사전 테스트를 수행하여 피험자들에게 실험 절차를 숙지시켰다. 이어서 피험자에게 전체 동영상을 시청하게 한 후 동영상의 각 쇼트의 처음 3.7초 동안만 제시한 후 현재 쇼트를 단서로 하여 방금 전 시청한 동영상의 주제와의 관련도(관련: 3, 부분관련: 2, 무관: 1)를 평가하도록 하였다(〈그림 3〉 참조).

뇌파와 시선추적은 비디오를 시청하는 동안 측정하고 주제 관련도는 비디오 시청후에 판정



〈그림 2〉 실험 시스템



〈그림 3〉 실험 절차

하기 때문에 이들 간에 차이가 생길 수 있다. 예를 들어서 피험자가 특정 비디오 쇼트를 시청할 때는 주제와 관련된 쇼트라고 생각했지만 비디오 시청후에는 주제와 무관하다고 판단할 수 있다. 이러한 차이를 최소화하기 위해서 주제 관련성 판정시 해당 쇼트가 주제와 관련되어 있는지, 비디오를 시청할 때도 같은 생각이었는지 기술하도록 하여 이 두 정보가 일치하는 경우만 분석하는 방법이 있으나 본 연구에서는 이러한 정보를 이용하지는 않았다.

4.3 뇌파 및 동공크기 측정과 분석

비디오 내의 해당 쇼트의 시작점을 찾아서 -200ms에서 시작점까지의 평균 뇌파값을 뇌파 기저값으로 하고, 비디오의 제시 이전에 보여주는 회색 스크린을 보는 200ms 동안의 평균 동공크기를 동공크기 기저값으로 사용하였다. 시작점에서부터 1,000ms에 대한 뇌파 및 동공크기 데이터를 각 동영상 쇼트를 시청할 때 나타나는 뇌파 및 시각 반응으로 보고 분석하였다. 각 피험자의 뇌파와 동공 크기를 세밀하게 분석하기 전에 뇌파 및 시각 패턴을 분석해 보았다.

4.3.1 뇌파 및 시각 패턴

피험자가 주제에 적합한 쇼트를 시청할 때 최고 뇌파값/동공크기값 대신에 최저 뇌파값/동공크기값, 중간 뇌파값/동공크기값을 나타내기도 하였다. 이와 같이 일반적인 뇌파 및 시각 패턴과 다르게 나타나는 피험자의 데이터를 분석하기 위해서 다음과 같은 작업을 수행했다. 30개 채널 중 13개 채널들(P600_uV/N400_uV 모형에서 함께 사용한 12개 채널에 Pz를 추가함)을 분석하였고 분석 대상 비디오는 김구의 회중시계이다. 13개 채널 중에 7개 이상이 주제 관련 뇌파값이 주제무관 뇌파값 보다 높으면 A 유형으로 정하였고 6개 이하이면 B유형으로 정하였다. 피험자 27명 중 13명(48%)만이 P600/

N400 컴포넌트에서 모두 A유형으로 나타났다. 7명(26%)은 P600에서 B유형, N400에서 A유형을 나타냈고, 3명(11%)은 P600에서 A유형, N400에서 B유형을 나타냈다. 나머지 4명(15%)은 P600/N400 모두에서 B유형으로 나타났다. 한편 피험자 27명 중 21명(78%)은 주제관련의 동공크기가 주제무관의 동공크기보다 큰 A유형을 보였고, 나머지 6명(22%)은 주제관련의 동공크기가 주제무관의 동공크기보다 작은 B유형을 보였다.

4.3.2 분석 결과

뇌파의 경우, 총 에폭 729개(피험자당 27개 에폭들[쇼트들])는 주제관련 에폭 219개, 주제부분관련 에폭 304개, 주제무관 에폭 206개로 구분되었다. 일반적인 동영상 요약을 구성하기 위해서 주제부분관련 에폭들은 애매성 때문에 제외시키고 425개의 에폭들만 이용하였다. 이러한 425개 에폭 데이터를 입력데이터로 하여 CURRY 7.0 프로그램과 SPSS 23을 이용하여 각 피험자의 비디오 쇼트별로 P600_uV값과 N00_uV값을 계산하였다. 개인화된 동영상 구성을 위해서 에폭 729개를 모두 사용하였고, 절차는 일반적인 동영상 요약과 동일하게 처리하였다.

동공크기의 경우, 데이터는 Tobii Pro Lab 프로그램에 의해서 수집되었고, 수집된 데이터와 엑셀을 이용하여 각 비디오 쇼트의 동공크기(ET_P3b)를 측정하였다. 구체적으로 동공크기의 기저값으로 각 피험자가 비디오를 시청하기 바로 전에 보여주는 회색 스크린을 보는 0.2초 동안의 평균 동공크기를 사용하였다(〈그림 2〉 참조). 이후 각 쇼트를 보는 처음 1초 동안의 동공크기를 0.1초 단위로 각각 분석하였다. 각 0.1초 단위의 동공크기(0.1 동안 동공크기의 평균값)는 기저값과의 차이로 계산한 표준값을 이용하였고, 이 표준값을 이용하여 각 피험자의 비디오 쇼트별로 ET_P3b값(각 쇼트의 시작점에서 0.2초~0.5초 사이의 최고 동공크기값)을 계산하였다.

5. 동영상 요약 방법

5.1 표준 동영상 요약

본 연구에서 제안한 방식들의 유용성을 평가하기 위해서 기준이 되는 표준 동영상 요약을 구성하였다. 표준 동영상 요약의 구성을 위해서 세 명의 미디어 전문가들에게 가장 주제를 잘 표현하는 쇼트들을 선정하도록 요청하였다. 신라의 금관 비디오에서는 세 개의 쇼트들을 추출하도록 하였고, 김구의 회중시계 비디오에서는 네 개의 쇼트들을 추출하여 공통된 쇼트들을 사용하였으며, 의견 일치가 안된 경우에는 다시 상의하여 최종 쇼트들을 선정하였다.

5.2 오디오-비주얼(AV) 시맨틱 기반 동영상 요약

5.2.1 개인화된 동영상 요약

동영상 요약의 구성을 위해서 주제관련, 주제부분관련, 주제무관 데이터를 사용하여 개별 피험자의 각 쇼트에 대한 뇌파값 및 동공크기값을 측정하였다.

(1) 사전 뇌파 측정(사전 테스트)을 수행한

다. 피험자에게 동영상을 시청하게 한 후 동영상의 각 쇼트를 제시하여 현재 쇼트를 단서로 하여 방금 전 시청한 동영상의 주제와의 관련도(관련, 부분관련, 무관)를 평가하도록 한다. 주제 관련성 판정시 해당 쇼트가 주제와 관련되어 있는지, 비디오를 시청할 때도 같은 생각이었는지 기술하도록 하여 이 두 정보가 일치하는 경우의 비디오 쇼트의 뇌파값(P600_uV, N400_uV)과 동공크기값(ET_P3b)을 학습 데이터로 저장한다.

(2) 피험자에게 요약을 원하는 비디오를 시청하게 하여 각 비디오 쇼트의 뇌파값/동공크기값을 측정하여 저장한다. 구체적으로 공식(1)의 뇌파값(P600_uV), 공식(2)의 뇌파값(N400_uV) 및 동공크기값(ET_P3b)을 각각 계산하며 분류의 검정 데이터로 사용한다.

(3) 판별 분석이나 합성곱 신경망(CNN) 분류 방법을 사용하여 상기의 학습 및 검정 데이터를 입력 데이터로 하여 P600_uV 기반, N400_uV 기반 및 ET_P3b 기반 동영상 요약물을 구성한다.

(4) 뇌파와 동공크기 데이터를 결합하기 위해서 다음과 같은 절차를 이용한다. 첫째, P600_uV/ET_P3b의 경우, P600_uV 기반 동영상 요약물과 ET_P3b 기반 동영상 요약물에 동시에 출현하는 쇼트들을 먼저 선정한다. 선정된 쇼트들이 충분치 않을 경우 동공크기(ET_P3b) 값들의 평균(임계치)을 구하고 앞에서 선정되지 않은 ET_P3b의 임계치 이상의 쇼트(들)에서 ET_P3b값이 높은 것을 우선적으로 하여 선정한다. 둘째, N400_uV/ET_P3b 기반 동영상 요약도 앞의 P600_uV/ET_P3b 기반 동영상 요약과 같은 방식으로 추출한다.

5.2.2 일반적인 동영상 요약

동영상 요약을 위해서 주제관련 및 주제무관 데이터(주제부분관련 데이터 제외)를 사용하여 27명 피험자들의 각 쇼트에 대한 평균 뇌파값, 평균 동공크기값을 사용하였다. 이때, 주제무관일 때 최저값 평균(P600_uV값[7.37], N400_uV값[-18.03], ET_P3b값[0.30]), 주제관련일 때 최고값 평균(P600_uV값[10.48], N400_uV값[-12.91], ET_P3b값[0.40])으로 주제관련일 때의 뇌파/동공크기값이 주제무관일 때의 뇌파/동공크기값보다 높게 나타났다.

(1) 비디오 쇼트별로 피험자들의 평균 뇌파값(P600_uV, N400_uV)과 평균 동공크기값(ET_P3b)을 계산하여 저장한다.

(2) 비디오 쇼트들을 뇌파값 또는 동공크기값의 내림차순으로 정렬한 후 원하는 수만큼의 쇼트들을 추출하여 P600_uV 기반, N400_uV 기반 및 ET_P3b 기반 동영상 요약물을 각각 구성한다.

(3) 뇌파와 동공크기 데이터를 결합하는 절차는 개인화된 동영상 요약과 같은 방식을 사용한다.

5.3 텍스트 시맨틱스 기반 동영상 요약

잠재의미분석과 페이지랭크 중심성을 이용하여 일반적인 동영상 요약을 구성한다. 두 비디오의 오디오 트랙에서 추출한 내레이션, 스피치 및 대화의 트랜스크립을 작성한 후 이로부터 불용어와 기능어를 제외한 키워드들을 추출하여 입력 데이터로 사용한다. 잠재의미분석 기반 요약을 구성하기 위해서 위의 키워드들을 입력 데이터로 하여 MATLAB R2014a의 특

이값 분해(SVD) 기능을 이용하여 각 비디오 쇼트의 가중치를 계산한다.

한편 페이지랭크 중심성 기반 요약을 위해서 키워드들을 입력 데이터로 하여 노드엑셀(NodeXL)을 이용하여 각 쇼트의 가중치를 구하였다(Hansen, Shneiderman, & Smith, 2010). 이후 각 비디오의 쇼트들을 가중치의 내림차순으로 정렬한 후 원하는 수만큼의 쇼트들을 추출하여 동영상 요약을 구성한다. 개인의 관심 주제에 관한 정보를 이용하면 개인화된 텍스트 시맨틱스 기반 동영상 요약 구성도 가능하나 본 연구에서는 다루지 않았다.

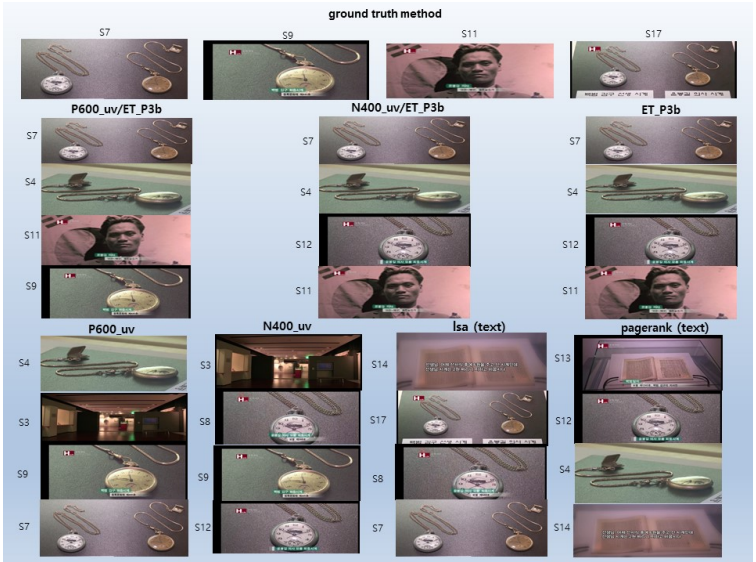
5.4 동영상 요약 구성

5.4.1 일반적인 동영상 요약 구성

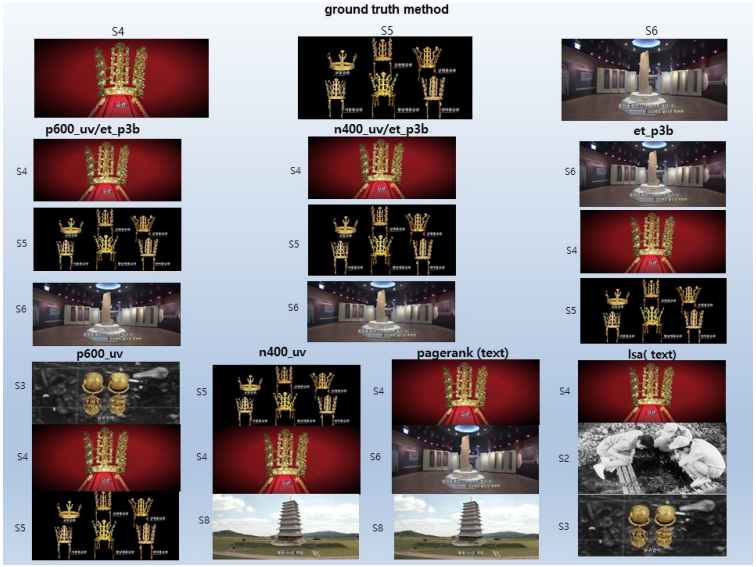
상기에서 기술한 AV 시맨틱스 기반의 일반적인 동영상 요약 방법을 구체적인 예를 들어서 설명하면 다음과 같다(〈그림 4〉, 〈그림 5〉 참조). 비디오 1의 첫번째 쇼트(s1)의 평균 P600_uV 값을 구하기 위하여 27명 각각의 해당 쇼트의 P600_uV값을 모두 합하여 이 합계를 27로 나눠서 평균을 구하였다. 이러한 방식으로 나머지 18개 쇼트의 평균을 구한 다음 이 값들을 내림차순으로 정렬하여 처음 네 개의 쇼트들(s4, s3, s9, s7)로 P600_uV 기반의 동영상 요약을 구성하였다.

동일한 방식으로 N400_uV 기반, ET_P3b 기반의 동영상 요약을 구성하였다. 이 값들을 이용하여 P600_uV/ET_P3b 기반, N400_uV/ET_P3b 기반의 동영상 요약을 구성하였다("5.2.2 일반적인 동영상 요약" 참조). 같은 방식으로 비디오 2의 일반적인 동영상 요약들도 구성하였다.

한편 텍스트 시맨틱스 기반의 잠재의미분석, 페이지랭크 중심성을 이용하여 일반적인 동영상 요약을 구성하였다. 구체적으로 이 두 방법



〈그림 4〉 일반적인 요약(김구의 회중시계)



〈그림 5〉 일반적인 요약(신라의 금관)

들에 의해서 비디오 쇼트의 가중치를 계산하였다. 이후 각 비디오의 쇼트들을 가중치의 내림차순으로 정렬하여 원하는 수만큼의 쇼트들을 추출하여 요약을 구성하였다("5.3 텍스트 시맨틱스 기반 동영상 요약" 참조).

5.4.2 개인화된 동영상 요약 구성

본 연구에서는 상기에서 기술한 개인화된 동영상 요약 방법과는 다르게 사전 테스트는 본 실험을 피험자에게 이해시키기 위해서 진행시켰다. 따라서 본 실험에서 수집한 각 비디오의 뇌파 및 동공크기 데이터를 학습 데이터로 사용하여 판별분석을 수행하였다. 이때 교차 타당화(cross-validation) 방법에 의해서 분류되었다. 교차 타당화 방법은 하나의 사례(쇼트)를 분류할 때 그 사례를 제외한 나머지 쇼트들을 학습 데이터로 이용하여 해당 쇼트(검정 데이터)가 분류되는 방식이다. 예를 들어서, 한 피험자의 신라의 금관 비디오의 첫번째 쇼트를 분류하기 위해서 나머지 7개 쇼트들의 뇌파값/동공크기값들이 학습 데이터로 사용되어 분류되었다. 이때 원하는 쇼트보다 더 많은 쇼트들이 예측소속집단(3: 주제관련)으로 분류될 경우 쇼트의 소속집단확률값이 높은 것을 우선하여 선정하였다.

6. 동영상 요약 평가와 연구문제 검증

6.1 상관관계 분석

평가에 앞서서 AV 시맨틱스 기반의 동영상 요약 방법들(뇌파[P600_uV, N400_uV], 동공크기[ET_P3b]), 텍스트 시맨틱스 기반의 동영상 요약 방법들 (잠재의미분석[lsa], 페이지랭크[PageRank])), 관련성(rel), 비주얼 콘텐츠 간의 상관관계를 분석하였다(〈표 2〉 참

〈표 2〉 상관관계 분석 결과

		lsa	rel	N400_uV	P600_uV	ET_P3b
lsa	Pearson Correlation (Sig.)	1	**.489**** **(0.003)**	0.297 (0.083)	0.097 (0.578)	0.074 (0.673)
rel	Pearson Correlation (Sig.)	**.489**** **(0.003)**	1	0.31 (0.07)	**.359*** **(0.034)**	0.149 (0.393)
N400_uV	Pearson Correlation (Sig.)	0.297 (0.083)	0.31 (0.07)	1	**.513**** **(0.002)**	0.117 (0.505)
P600_uV	Pearson Correlation (Sig.)	0.097 (0.578)	**.359*** **(0.034)**	**.513**** **(0.002)**	1	0.207 (0.233)
ET_P3b	Pearson Correlation (Sig.)	0.074 (0.673)	0.149 (0.393)	0.117 (0.505)	0.207 (0.233)	1
PageRank	Pearson Correlation (Sig.)	**.362*** **(0.033)**	**.469**** **(0.004)**	**.357*** **(0.035)**	0.054 (0.756)	0.316 (0.065)
caption	Pearson Correlation (Sig.)	0.203 (0.242)	0.204 (0.24)	0.16 (0.359)	0.219 (0.207)	0.081 (0.645)
person	Pearson Correlation (Sig.)	-0.276 (0.109)	-0.191 (0.272)	-0.3 (0.08)	-0.13 (0.455)	0.034 (0.846)
object	Pearson Correlation (Sig.)	0.29 (0.092)	**.740**** **(0)**	0.262 (0.128)	**.483**** **(0.003)**	**.345*** **(0.042)**
bg	Pearson Correlation (Sig.)	-0.124 (0.477)	**-.556**** **(0.001)**	-0.173 (0.32)	**-.513**** **(0.002)**	**-.405*** **(0.016)**
		PageRank	caption	person	object	bg
lsa	Pearson Correlation (Sig.)	**.362*** **(.033)**	.203 (.242)	-.276 (.109)	.290 (.092)	-.124 (.477)
rel	Pearson Correlation (Sig.)	**.469**** **(.004)**	.204 (.240)	-.191 (.272)	**.740**** **(.000)**	**-.556**** **(.001)**
N400_uV	Pearson Correlation (Sig.)	**.357*** **(.035)**	.160 (.359)	-.300 (.080)	.262 (.128)	-.173 (.320)
P600_uV	Pearson Correlation (Sig.)	.054 (.756)	.219 (.207)	-.130 (.455)	**.483**** **(.003)**	**-.513**** **(.002)**
ET_P3b	Pearson Correlation (Sig.)	.316 (.065)	.081 (.645)	.034 (.846)	**.345*** **(.042)**	**-.405*** **(.016)**
PageRank	Pearson Correlation (Sig.)	1	.150 (.389)	-.110 (.530)	**.371*** **(.028)**	-.296 (.084)
caption	Pearson Correlation (Sig.)	.150 (.389)	1	-.052 (.767)	.264 (.125)	-.290 (.091)
person	Pearson Correlation (Sig.)	-.110 (.530)	-.052 (.767)	1	**-.349*** **(.040)**	-.276 (.108)
object	Pearson Correlation (Sig.)	**.371*** **(.028)**	.264 (.125)	**-.349*** **(.040)**	1	**-.747**** **(.000)**
bg	Pearson Correlation (Sig.)	-.296 (.084)	-.290 (.091)	-.276 (.108)	**-.747**** **(.000)**	1

* $p < 0.05$, ** $p < 0.01$

조). 비주얼 콘텐츠는 잠재기가 가장 긴 P600(500~700ms)을 고려하여 각 쇼트의 처음 700ms 동안에 제시되는 콘텐츠를 세 개의 비주얼 요소 즉 인물(person), 객체(object), 배경(background, bg) 그리고 캡션 텍스트(caption)로 구분한 것이다. 상관관계 분석 결과를 요약하면 다음과 같다.

첫째, AV 시맨틱스 기반의 뇌파(P600_uV, N400_uV)에 의한 방법들과 동공크기(ET_P3b)에 의한 방법 간의 관계를 살펴보니 P600_uV와 N400_uV 간에 정적 상관관계가 있었으나 뇌파와 동공크기 간에는 직접적인 관련이 없는 것으로 나타났다. 다만 뇌파(P600_uV)와 동공크기(ET_P3b)는 0.05 유의수준에서 두 개의 비주얼 요소 즉, 객체(정적 상관관계)와 배경(부적 상관관계)과 관계가 있는 것으로 나타났다. 이는 만약 쇼트가 객체를 포함하고 있다면 해당 쇼트가 높은 P600_uV값 또는 높은 ET_P3b값을 가질 가능성이 있다는 것을 의미한다. 다른 한편 쇼트가 배경 쇼트라면 낮은 P600_uV값 또는 낮은 ET_P3b값을 가질 가능성이 있다.

둘째, 각 쇼트에 대한 관련성 판정은 0.01 유의수준에서 텍스트 시맨틱스(lsa, PageRank)와 정적 상관관계가 있는 것으로 나타났다. 예를 들어서, 주제 관련성이 높은 쇼트의 경우 잠재의미분석값이나 페이지랭크값이 높게 나타난다는 것을 의미한다. 또한 객체가 나타나는 쇼트의 주제 관련도가 높아지는 반면 배경 쇼트는 주제 관련도가 낮아진다.

셋째, AV 시맨틱스 기반 방법과 텍스트 시맨틱스 기반 방법 간의 상관관계를 분석한 결과, 유의수준 0.05에서 페이지랭크와 N400_uV 간에 정적 상관관계가 있는 것으로 나타났다. 즉, 쇼트의 페이지랭크값이 높으면 해당 쇼트의 N400_uV값이 높아진다는 것을 의미한다.

6.2 동영상 요약 평가

AV 시맨틱스 기반 방법들에 의한 일반적인 요약에서 뇌파 P600_uV와 동공크기 ET_P3b를 함께 사용한 경우의 동영상 요약의 재현율(0.88)이 가장 높게 나왔다. 동공크기에 기반한 동영상 요약의 재현율(0.75)이 뇌파 P600_uV(0.59) 또는 뇌파 N400_uV(0.46)에 기반한 동영상 요약의 재현율보다 각각 더 높게 나타났다.

〈표 3〉 재현율 평가

방법	재현율		
	일반적인 요약	개인화된 요약	평균
P600_uV/ET_P3b	0.88	0.64	0.76
N400_uV/ET_P3b	0.75	0.63	0.69
ET_P3b	0.75	0.61	0.68
P600_uV	0.59	0.45	0.52
N400_uV	0.46	0.50	0.48
잠재의미분석(las)	0.42	-	-
페이지랭크(pagerank)	0.34	-	-

한편 개인화된 요약에서는 뇌파 P600_uV와 동공크기 ET_P3b를 함께 사용한 경우의 동영상 요약의 재현율(0.64)이 가장 높게 나왔다. 동공크기에 기반한 동영상 요약의 재현율(0.61)이 뇌파 P600_uV(0.45) 또는 뇌파 N400_uV(0.50)에 기반한 동영상 요약의 재현율보다 더 높게 나타났다.

6.3 연구문제 검증

다음은 세 가지 연구문제의 검증 결과를 기술하며, 자세한 내용은 논의에서 다룬다.

6.3.1 뇌파, 시선추적 및 뇌파/시선추적의 재현율 비교

일반적인 요약과 개인화된 요약에서 P600 뇌파/시선추적 방법(0.76)이나 N400 뇌파/시선추적 방법(0.69)을 사용한 경우의 평균 재현율이 시선추적(0.68), P600(0.52), N400(0.48) 방법만을 사용한 경우의 평균 재현율보다 높게 나타났다. 따라서 뇌파와 시선추적 데이터를 함께 사용하면 뇌파 또는 시선추적 데이터만을 사용한 경우보다 재현율을 향상시킬 수 있다는 것을 확인하였다. 시선추적 기반 방식의 평균 재현율(0.68)이 뇌파 기반 방식의 재현율(P600_uV: 0.52 N400_uV: 0.48)보다 높게 나타났다.

6.3.2 오디오-비주얼(AV) 시맨틱스 기반의 일반적인 동영상 요약과 개인화된 동영상 요약 간의 비교

AV 시맨틱스 기반 방법들에 의한 개인화된 요약의 평균 재현율(0.57)이 일반적인 동영상 요약의 평균 재현율(0.69)보다 낮게 나타났다(<표 3> 참조). 또한 P600_uV(일반: 0.59, 개인화: 0.45), N400_uV(일반: 0.46, 개인화: 0.50) 방식에 의해서 구성한 동영상의 재현율은 모두 0.60 미만으로 시선추적의 재현율(일반: 0.75, 개인화: 0.61)보다 낮게 나타났다.

6.3.3 오디오-비주얼(AV) 시맨틱스 기반 동영상 요약 방법과 텍스트 시맨틱스 기반 동영상 요약 방법 간의 비교

AV 시맨틱스 기반 방법들에 의한 일반적인 동영상 요약들의 평균 재현율(0.69)이 텍스트 시맨틱스 기반 방법들에 의한 일반적인 동영상 요약들의 평균 재현율(0.38)보다 높게 나타났다.

7. 논의와 결론

본 연구는 AV 시맨틱스에 기반한 시선 및 뇌파 반응을 활용한 동영상 요약의 구성 방법을 제안한 후 각 방법을 표준 동영상 요약과 비교하여 평가해 보았다. 다음은 세 가지 연구문제들의 분석 결과와 개선점, 연구 결과의 활용 및 제한점에 대해서 살펴본다.

7.1 뇌파, 동공크기 및 뇌파/동공크기의 재현율 비교

일반적인・개인화된 요약에서 뇌파와 동공크기 데이터를 함께 사용한 방법의 평균 재현율(0.73)(P600 뇌파/동공크기: 0.76, N400 뇌파/동공크기: 0.69)이 뇌파 기반 방법의 평균 재현율(0.50)(P600: 0.52, N400: 0.48) 또는 동공크기 기반 방법의 평균 재현율(0.68)보다 높

게 나타났다.

신경과학적 방법으로 데이터를 수집하는 것이 필요하다면 이 두 방법들을 함께 사용하는 것이 어느 한 쪽의 데이터 손실이 발생했을 때를 대비할 수 있다. 본 연구에서 실험을 하는 동안 피험자의 움직임, 주위 환경의 영향 등 다양한 이유로 시선추적의 데이터 손실이 생각보다 많았고, 같은 이유로 뇌파 신호도 왜곡되는 경우가 있었다. 이와 같이 이 두 종류의 데이터가 완벽하게 수집되지 않은 경우 한 종류의 데이터라도 사용해야 하는 상황이 발생할 수 있을 것이다.

앞에서 언급한 것처럼 동공크기 데이터의 평균 재현율(0.68)이 뇌파 데이터의 평균 재현율(0.50)보다 높게 나타났다. 이는 ET_P3b 효과가 관심있는 대상이나 주제를 인지할 때 일반적으로 동공크기가 커지는 경향이 있으며 시선반응에 있어서 개인차가 크지 않기 때문으로 볼 수 있다.

한편 뇌파 반응은 개인별로 또는 검사 환경에 따라서 차이가 날 수 있다. 따라서, 관심있는 대상이나 주제를 인지할 때 강한 P600 효과를 나타내고, 관심이 없는 대상이나 주제의 경우는 강한 N400 효과를 나타내는 것 이외에 여러 요인들 즉, 주제 분석과 통합, 주제 관련도 판정, 사전 지식 등이 뇌파에 영향을 미치는 것이 아닌지 생각된다. 구체적으로 피험자가 각 비디오 쇼트를 보면서 주제를 파악하고 이전 쇼트의 내용과 통합하는 과정에서 핵심 주제를 파악하고 중요하지 않는 것은 걸러내는 과정을 거치게 된다. 이러한 과정에서 피험자들은 주제에 관련된 쇼트를 시청할 때 뇌파값이 높아지는 경우도 있고, 때로는 주제, 개념, 상황이 명확하게 파악되지 않을 경우 인지부하로 인해 뇌파값이 높아지기도 한다.

예를 들어서, 일부 피험자들은 비디오 주제에 부분적으로 관련되어 있는 다소 애매한 쇼트를 시청할 때 뇌파값이 최고가 되었다. 이 경우 각 쇼트의 뇌파값을 내림차순으로 정렬하여 선정한다면 피험자가 주제관련이라고 판정한 쇼트의 뇌파값이 중간 영역에 있게 되어 동영상 요약을 위해 선정되지 않을 수 있다. 더 나아가 13개 채널 모두가 주제관련의 뇌파값이 주제무관 뇌파값보다 높은 피험자들이 있는 반면 어떤 피험자들은 일부 채널에서만 그렇고 나머지 채널에서는 반대로 주제관련의 뇌파값이 주제무관 뇌파값보다 낮아지기도 한다. 따라서 좀 더 다양하고 정확한 개인별 뇌파 패턴을 가려내기 위해서 각 개인의 13개 채널들의 P600/N400 뇌파 신호를 입력 데이터로 사용해서 합성곱 신경망(CNN)으로 분석해 보는 것이 후속 연구로 필요해 보인다.

7.2 오디오-비주얼(AV) 시맨틱스 기반의 일반적인 동영상 요약과 개인화된 동영상 요약 간의 비교

개인화된 요약의 평균 재현율(0.57)이 일반적인 동영상 요약의 평균 재현율(0.69)보다 낮게 나타났다. 이는 개인에 따라서 비디오 쇼트에 대한 주제 관련성에 대한 판정이 다를 수 있기 때문일 수 있다. 또 다른 이유를 살펴보면, 일반적인 요약을 구성하기 위해서 27명 피험자들의 각 비디오 쇼트의 뇌파 및 동공크기의 평균을 계산할 때 주제부분관련 데이터를 제외시켰다. 그 결과, 세 가지 컴포넌트(P600_uV, N400_uV,

ET_P3b)에서 주제관련의 평균(P600_uV값[10.48], N400_uV값[-12.91], ET_P3b값[0.40])이 주제무관의 평균(P600_uV값[7.37], N400_uV값[-18.03], ET_P3b값[0.30])보다 낮게 나타났다. 따라서 피험자가 주제관련의 쇼트를 시청할 때 평균 뇌파값/동공크기값이 가장 높아서 피험자가 주제관련이라고 판정한 쇼트가 일반적인 동영상 요약에 포함될 가능성이 높아졌기 때문일 수도 있다.

7.3 오디오-비주얼(AV) 시맨틱스 기반 동영상 요약 방법과 텍스트 시맨틱스 기반 동영상 요약 방법 간의 비교

텍스트 시맨틱스 기반 방법들의 동영상 요약들의 평균 재현율(0.38)이 AV 시맨틱스 기반 방법들에 의한 동영상 요약들의 평균 재현율(0.69)보다 낮게 나타났다. 이러한 결과가 나오는 이유는 실험에 사용된 두 개의 비디오들에서 비주얼 콘텐츠가 주제 관련성을 판정하는데 중요한 역할을 했을 것으로 생각된다. 또 다른 이유로는 본 연구에서 사용한 텍스트 시맨틱스 기반 방법들은 오디오 콘텐츠의 의미를 파악하기 위해 오디오 콘텐츠에서 키워드들을 추출하여 분석한 것이기 때문에 이용자와의 상호 작용에서 의미 격차(conceptual gap) 문제가 발생할 수 있기 때문이다. 한편, 실험 비디오가 다큐멘터리이기 때문에 이 두 방법들에 의해서 추출된 비디오 요약 간에 중복이 있을 것으로 예측하였다. 예측한 대로 신라의 금관과 김구의 회중시계 비디오 요약에서 중복이 있었으나 비디오별로 차이도 있었다. 이를 자세히 설명하면 다음과 같다.

첫째, 신라의 금관 비디오의 경우 쇼트 4(s4)는 두 방법들에 의해서 구성된 비디오 요약들에 모두 포함되어 있다(〈그림 5〉 참조). 쇼트 4는 비디오의 해설자(narrator)가 "금관은 신라인들이 꽃피운 고도한 황금 문화의 상징이자 당대 정치상황을 보여주는 유물이다"라는 입말 텍스트를 제시하면서 금관 이미지를 보여주는 장면이다. 따라서 쇼트 4는 비디오 주제와 관련된 '금관'과 '신라'라는 단어가 해당 쇼트의 트랜스크립에 포함되어 있어서 높은 가중치를 갖게 되었을 것이고 이로 인하여 텍스트 시맨틱스 기법들에 의해서 구성된 비디오 요약에 포함된 것으로 보인다. 한편 쇼트 4는 주제와 관련된 금관(객체) 이미지와 화면 아래에 캡션 텍스트로 '금관'이라는 단어를 보여주고 있으며 이로 인하여 AV 시맨틱스 기법들에 의해서 구성된 비디오 요약에 포함된 것으로 보인다.

둘째, 김구의 회중시계 비디오의 경우 페이지랭크 중심성 기법에 의해서만 선정된 쇼트 13과 쇼트 14 그리고 잠재의미분석에 의해서 선정된 쇼트 14가 있다. 쇼트 13은 김구의 백범 일지를 소개하고, 쇼트 14는 백범 일지에 나오는 김구와 윤봉길과의 대화 내용을 화면의 자막으로 처리한 쇼트이다. AV 시맨틱스 기반 기법들에 의해서 구성된 요약들에는 쇼트 13과 쇼트 14가 포함되지 않았는데 이는 이 쇼트들에 비디오 주제와 관련된 객체(예, 회중시계) 또는 인물(예, 김구)과 같은 비주얼 요소가 포함되지 않아서 뇌파와 동공 반응이 크지 않았다고 생각해 볼 수 있다.

비디오 요약을 위해서 본 연구에서 제안한 AV 시맨틱스 기반 방법만 사용할 경우, 비디

오 쇼트가 주제와 관련된 시맨틱 AV 정보를 포함하고 있다고 해서 그 쇼트가 반드시 주제와 관련된 시맨틱 텍스트 정보를 포함하고 있다고 가정할 수 없다는 사실은 여전히 중요한 쟁점이 된다.

일반적인 동영상 요약의 재현율을 높이기 위해서 정적 상관관계가 있는 페이지랭크와 N400_uV를 결합하였을 때 재현율이 높아지지 않고 오히려 낮아졌다(N400_uV: 0.46 -〉 페이지랭크/N400_uV: 0.34).

7.4 응용 영역

본 연구에서 제안한 동영상 요약 방법들에 의하면 각 비디오 쇼트의 시작점에서 최대 700ms 동안의 비주얼 콘텐츠에 대한 뇌파 및 시선 반응이 오디오 콘텐츠에 대한 반응에 비하여 매우 중요한 부분을 차지할 것이라고 예측된다. 실제 실험에 사용된 비디오들을 분석한 결과, 시작점에서 700ms 동안에 하나 또는 복수개의 음성 단어들 또는 심지어는 음성 단어가 전혀 제시되지 않기도 하였다. 이에 따라서 본 연구에서 제안한 AV 시맨틱스 기반 동영상 요약 방법들은 비주얼 콘텐츠 중심의 비디오에 적용할 경우 더 효율적일 것으로 생각된다.

구체적으로, 본 연구 결과는 이용자들이 이미 시청한 비디오를 재활용할 수 있게 하는 개인화된 비디오 요약 구성에 유용하게 사용될 수 있다. 또한 수집된 다양한 개인화된 비디오 요약들(빅데이터)을 분석한 후 많은 시청자들에 의해서 중요하다고 측정된 부분(예, 비디오 쇼트)을 추출하여 일반적인 동영상 요약으로 구성할 수 있을 것이다. 이러한 일반적인 동영상 요약은 디지털 도서관 시스템의 메타데이터로 활용되어 멀티미디어의 내용 기반 검색과 브라우징을 지원할 수 있을 것이다. 또한, 본 연구 결과는 각 시청자의 관심 분야에 맞는 동영상 자료를 추천하는 데에도 활용될 수 있을 것으로 생각된다.

디지털 도서관에서 적합성 판정은 이용자들이 정보를 찾는데 기준이 되는 매우 중요한 개념이다. 최근 암묵적인 적합성 판정에 대한 관심이 높아지고 있고 이를 디지털 도서관에 적용하려고 하는 연구들이 진행되고 있다. 본 연구 결과는 디지털 도서관 시스템의 멀티미디어 자료를 검색하는 과정에서 암묵적 적합성 판단의 기준으로 적용되어 차세대 시선추적-뇌파 기반 인터페이스(EBI, Eye-Brain Interface)의 구현을 위한 이론적인 틀로 이용될 수 있을 것이다.

7.5 제한점

실험의 피험자들이 20대 남학생들로 구성되어 있어서 연구 결과를 다른 연령이나 성별에 일반화하는데 제한이 있을 수 있다. 또한 본 연구의 실험 데이터로 사용한 다큐멘터리 비디오와 다른 장르(예, 뉴스, 영화)의 비디오를 실험에 사용한다면 그 결과가 달라질 수 있을 것이다.

참 고 문 헌

[1] 김현희, 김용호 (2016). 뇌파측정기술 (EEG)에 기초한 멀티미디어 자료의 주제 적합성에 관한 연구. 한국문헌정보학회지, 50(3), 361-381.

[2] 김현희, 김용호 (2019). 뇌파측정기술을 활용한 언어 기반 사운드 요약의 생성 방안 연구. 정보관리학회지, 36(3), 131-148.

[3] 안형모 (2013). 텍스트 기반 학습에서 학습매체의 크기와 종류가 학습 집중도에 미치는 영향. 국내석사학위논문, 한국교원대학교 대학원.

[4] 윤정원, 신수연 (2021). 페이스북 건강정보 게시물 형식이 시각적 주의와 인지결과에 미치는 영향. 한국문헌정보학회지, 55(3), 219-237.

[5] 정영미. (2012). 정보검색연구 (증보판). 서울: 연세대학교 출판문화원.

[6] Ajanki, A., Hardoon, D. R., Kaski, S., Puolamäki, K., & Shawe-Taylor, J. (2009). Can eyes reveal interest? Implicit queries from gaze patterns. User Modeling and User-Adapted Interaction, 19(4), 307-339.

[7] Bezugam, S. S., Majumdar, S., Ralekar, C., & Gandhi, T. K. (2021). Efficient video summarization framework using EEG and eye-tracking signals. arXiv preprint arXiv: 2101.11249.

[8] Bhattacharya, N. & Gwizdka, J. (2018). Relating eye-tracking measures with changes in knowledge on search tasks. arXiv preprint arXiv: 1805.02399.

[9] Bhattacharya, N., Rakshit, S., Gwizdka, J., & Kogut, P. (2020). Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (pp. 223-233).

[10] Evans, W. J., Cui, L., & Starr, A. (1995). Olfactory event-related potentials in normal human subjects: Effects of age and gender. Electroencephalography and Clinical Neurophysiology, 95(4), 293-301.

[11] Foley, J. J. & Kwan, P. (2015). Feature extraction in content-based image retrieval. In Encyclopedia of Information Science and Technology, Third Edition (pp. 5897-5905). IGI Global.

[12] Golenia, J. E., Wenzel, M. A., Bogojeski, M., & Blankertz, B. (2018). Implicit relevance feedback from electroencephalography and eye tracking in image search. Journal of Neural Engineering, 15(2), 026002.

[13] Gwizdka, J., Hosseini, R., Cole, M., & Wang, S. (2017). Temporal dynamics of eye-tracking and EEG during reading and relevance decisions. Journal of the Association for Information

Science and Technology, 68(10), 2299-2312.

[14] Gwizdka, J. & Zhang, Y. (2015). Differences in eye-tracking measures between visits and revisits to relevant and irrelevant web pages. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 811-814). ACM.

[15] Hansen, D., Shneiderman, B., & Smith, M. A. (2010). Analyzing social media networks with NodeXL: Insights from a connected world. MA: Morgan Kaufmann.

[16] Hughes, J. R. & John, E. R. (1999). Conventional and quantitative electroencephalography in psychiatry. The Journal of neuropsychiatry and clinical neurosciences, 11(2), 190-208.

[17] Katti, H., Yadati, K., Kankanhalli, M., & Tat-Seng, C. (2011). Affective video summarization and story board generation using pupillary dilation and eye gaze. In 2011 IEEE International Symposium on Multimedia (pp. 319-326). IEEE.

[18] Kaur, T. & Neeru, N. (2015). Text extraction from natural scene using PCA. International Journal of Computer Science Engineering & Technology, 5(7), 272-277.

[19] Kim, H. H. & Kim, Y. H. (2019a). Video summarization using event related potential responses to shot boundaries in real time video watching. Journal of the Association for Information Science and Technology, 70(2), 164-175.

[20] Kim, H. H. & Kim, Y. H. (2019b). ERP/MMR algorithm for classifying topic relevant and topic irrelevant visual shots of documentary videos. Journal of the Association for Information Science and Technology, 70(9), 931-941.

[21] Moon, J., Kwon, Y., Park, J., & Yoon, W. C. (2019). Detecting user attention to video segments using interval EEG features. Expert Systems with Applications, 115, 578-592. https://doi.org/10.1016/j.eswa.2018.08.016

[22] Oliveira, F. T., Aula, A., & Russell, D. M. (2009). Discriminating the relevance of web search results with measures of pupil size. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2209-2212). ACM.

[23] Schumacher, P. B. & Hung, Y.-C. (2012). Positional influences on information packaging: Insights from topological fields in German. Journal of Memory and Language, 67(2), 295-310.

[24] Shi, Z. F., Zhou, C., Zheng, W. L., & Lu, B. L. (2017). Attention evaluation with eye tracking glasses for EEG-based emotion recognition. In 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 86-89). IEEE.

[25] Syn, S. Y. & Yoon, J. (2021). Investigation on reading behaviors and cognitive outcomes of Facebook health information. Online Information Review, 45(6), 1097-1115.

[26] van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. Journal of Cognitive Neuroscience, 11(6), 657-671.

[27] Wang, L. & Schumacher, P. B. (2013). New is not always costly: evidence from online processing of topic and contrast in Japanese. Frontiers in Psychology, 4, 363.

[28] Zhao, L. M., Li, X. W., Zheng, W. L., & Lu, B. L. (2018). Active feedback framework with scan-path clustering for deep affective models. In International Conference on Neural Information Processing (pp. 330-340). Springer, Cham.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kim, H. & Kim, Y. (2016). Understanding topical relevance of multimedia based on EEG techniques. Journal of Korean Library and Information Science Society, 50(3), 361-381.

[2] Kim, H. & Kim, Y. (2019). Towards the generation of language-based sound summaries using electroencephalogram measurements. Journal of the Korean Society for Information Management, 36(3), 131-148.

[3] Ahn, H. M. (2013). The Effect of lEarning Media Size and Type to Learning Concentration in Text based Learning. Master's thesis, Korea National University of Education.

[4] Yoon, J. & Syn, S. Y. (2021). How do formats of health related Facebook posts effect on eye movements and cognitive outcomes?. Journal of the Korean Society for Library and Information Science, 55(3), 219-237.

[5] Chung, Y. M. (2012). Research in Information Retrieval. Seoul: Yonsei University Press.