Journal of KIIT. Vol. 19, No. 6, pp. 29-37, Jun. 30, 2021, pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.6.29

숏폼 콘텐츠 자막 자동 추출 시스템

조준영*, 김장원**[1], 온병원**[2], 정동원**[3]

Subtitle Automatic Extraction System for Short-form Contents

Junyoung Jo*, Jangwon Gim**[1], Byung-Won On**[2], and Dongwon Jeong**[3]

이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임
(NRF-2019R1F1A1060752)

요 약

최근 유튜브와 모바일 플랫폼 환경에서 짧고 강렬한 콘텐츠를 뜻하는 숏폼(short-form) 콘텐츠가 급증하고 있다. 콘텐츠의 홍수 속에서 핵심 내용 파악이 용이하여 시간 투자 대비 높은 만족감을 얻을 수 있기 때문이다. 숏폼 콘텐츠에서 핵심 내용의 간결하고 효율적인 표현을 위해 이미지(사물, 사람)와 한글 자모의 조합으로 구성된 특수 자막을 사용한다. 그러므로 일반적인 형태의 자막이 아닌 특수한 형태로 표현되는 특수 자막을 추출하는 방법이 필요하다. 본 논문에서는 숏폼 콘텐츠들에 존재하는 자막 자동식별 시스템을 제안한다. 숏폼 콘텐츠에 포함된 자막 데이터를 기계 학습하여 특수 및 일반자막의 식별 및 추출 결과 기존 방법보다 추출 정확도(95%)가 우수함을 보였다. 따라서 다양한 자막 형태가 포함된 숏폼 콘텐츠에 대한 사용자 추천 시스템 및 콘텐츠 큐레이션 서비스에 다양하게 활용할 수 있을 것으로 보인다.

Abstract

Recently, short-form content, which means short and intense content, is rapidly increasing in YouTube and mobile platform environments. Short-form contents use special subtitles and general subtitles composed of a combination of images (objects, people) and Hangul letters to deliver key content to viewers in a short time and give high satisfaction compared to time investment. Therefore, a particular subtitle extraction method is needed in short-form contents. In this paper, we propose an automatic identification system for subtitles appearing in short-form content. Through the algorithm for identifying the shape of the subtitles of the proposed system, the subtitle extraction accuracy (95%) included in the short-form contents was shown to be excellent. Therefore, it is expected that the proposed system can apply to the user recommendation system and content curation service for short-form content in the future.

Keywords
shortform content, caption extraction, optical character recognition, neural network

* 군산대학교 소프트웨어융합공학과 학부생
- ORCID: https://orcid.org/0000-0002-9944-2401
** 군산대학교 소프트웨어융합공학과 교수(**[1,2]교신저자)
- ORCID[1]: https://orcid.org/0000-0002-4480-7944
- ORCID[2]: https://orcid.org/0000-0001-6929-3188
- ORCID[3]: https://orcid.org/0000-0001-9881-5336

· Received: May 12, 2021, Revised: Jun. 24, 2021, Accepted: Jun. 27, 2021
· Corresponding Author: Jangwon Gim and Byung-Won On
Dept. of Software Science and Engineering, Kunsan National University, 558, Daehak-ro, Gunsan, Jeollabuk-do, Korea
Tel.: +82-63-469-8916, Email: {jwgim, bwon}@kunsan.ac.kr

Ⅰ. 서 론

최근 소셜네트워크서비스(SNS)에서의 사용자들의 의사소통 방식은 정형 텍스트에서 이미지 중심으로 변화하고 있으며 모바일 플랫폼의 활성화와 다양한 메신저 서비스의 발전으로 이러한 의사소통 방식은 더욱 가속화되고 있다. 즉, 장문의 텍스트 보다는 상황에 적절한 이모티콘, 동영상 콘텐츠에 포함된 특정 장면 이미지를 의사소통에 사용하는 것이 모바일 의사소통에서 중요한 요소가 되었다[1]. 국내에서 운영되는 대표적인 텍스트 기반 메신저 서비스에서는 입력된 텍스트로부터 특정 키워드를 식별 및 이모티콘과 연계하여 사용자가 원하는 이모티콘을 선택하여 대화할 수 있는 서비스를 도입하였다. 또한 모바일 메신저에서 텍스트를 사용하지 않고 이미지만으로 소통하는 오픈 대화방인 '고독방'이 따로 생겨날 정도로 텍스트가 아닌 데이터에 의존한 의사소통 방법은 흥미를 유도하고 친밀도를 향상시키는 역할을 하고 있다[2]. 그러므로 이러한 모바일 의사소통 환경에서 사용자들은 자신이 원하는 문맥 정보를 정확하고 효율적으로 전달해 줄 수 있는 이미지를 빠르고 정확하게 찾아서 공유하는 것에 보다 집중하게 되었다. 이런 상황에서 숏폼(Short-form) 콘텐츠 트렌드가 등장하였다. 숏폼 콘텐츠란 최대 10분 이내의 영상, 최소 15초 단위의 콘텐츠를 뜻한다[3]. 기존의 상영시간이 긴 동영상을 단순히 편집하여 줄이는 것이 아닌, 완결된 기획과 스토리텔링을 바탕으로 원하는 정보를 짧은 동영상을 통해 보여주는 콘텐츠를 뜻한다. 이를 위해 숏폼 동영상에는 상황 정보를 흥미 있고 보다 효율적으로 표현하기 위해 다양한 이미지와 자막들을 추가하여 표현한다. 동영상에서 자막은 영상의 정보에 부가하여 의미 전달을 위해 사용된다[4]. 따라서 모바일 플랫폼 사용자들을 위해서 숏폼 동영상에 존재하는 이미지, 자막 데이터들을 빠르고 정확하게 검색 및 재사용하기 위한 방법이 필요하다. 동영상에 표현된 자막을 추출하는 방법은 기존에 많은 연구가 되었지만 숏폼 콘텐츠 동영상에 빈번하게 등장하는 특수 자막(이미지-한글 자모 결합, 비정형 문자 형태 등)을 식별에는 다음과 같은 문제점이 존재한다. 첫 번째로 기존의 텍스트 추출 방식은 이미지 속 자막 이외의 텍스트도 자막으로 추출하는 경우가 발생한다. 따라서 이미지에 포함된 모든 텍스트를 추출할 경우 이미지에 대한 텍스트 검색 시 이미지 상황을 대표적으로 표현해주는 자막 이외의 텍스트가 검색 결과로 출현할 수 있어 이미지 검색에 대한 정확도 성능을 저하시킬 수 있다. 그러므로 이미지에 포함된 다수의 텍스트 중에서 자막 텍스트 추출에 집중할 필요가 있다. 두 번째로 자막의 표현 형태가 다양화 되고 있다. 기존의 동영상 콘텐츠에서는 설명 혹은 대사가 영상 하단에 위치하면서 이미지의 상황 정보를 정확하게 전달하는데 중점을 두었다면 이제는 영상에 재미를 증가시키기 위해 다양한 모양의 자막 디자인이 적용되고 있다. 그림 1과 같이 문법이나 글자 형태를 파괴, 해체하는 것은 물론 사람의 얼굴을 활용하여 한글 자모 역할로 사용하는 등의 자막이 등장한다. 따라서 본 논문에서는 영상처리와 인공 신경망을 사용하여 정형 글자 형태로 표현된 자막이 아닌 특수 자막을 인식할 수 있는 모폴로지 연산을 적용을 하여 특수 자막을 인식 성능을 향상시키는 시스템을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 기존 연구들을 살펴보고 기존 연구의 한계점을 보인다. 3장에서는 특수 자막 추출을 위한 모폴로지 연산과 인공신경망을 포함한 제안 시스템의 아키텍처를 소개한다. 4장에서는 제안 시스템에 대한 실험 방법 및 실험 결과를 보이고, 5장에서는 향후 계획과 결론을 맺는다.

그림 1. 다양한 특수 자막 유형 예시
Fig. 1. Examples of special subtitle types included in short-form contents

Ⅱ. 관련 동향

2.1 SNS 기반 커뮤니케이션 트렌드

1980년대 초에서 2000년대 초에 출생한 밀레니얼 세대(M)와 1990년대 중반 이후부터 2000년대 초반

에 출생한 세대(Z)를 통칭하는 MZ세대들은 장문의 텍스트 보다 적절한 타이밍에 보유한 이모티콘, 이미지 파일을 보내는 것에 더 흥미를 보인다[5]. 이렇듯 최근 SNS 상에서 의사소통 양상은 텍스트 위주로 소통하는 초기 모습과 다른 방식으로 빠르게 변화하고 있다. 특히 코로나 19의 판데믹 상황에서 비대면을 통한 온라인 환경에서 기존 텍스트 형태로 전달하기 어려운 미묘한 감정과 상황을 표현함에 있어서 이모티콘과 이미지 파일들은 전통적인 방법의 대안으로 채택되어 널리 사용되고 있다.

그림 2는 메신저 서비스를 통해 생성된 '고독방'을 보이고 있다. 참여 사용자들 간에 텍스트 형태로 대화하는 것이 아닌 이미지 파일만을 사용하여 대화를 진행한다.

그림 2. 이미지 기반 의사소통 채널 '고독방' 예시
Fig. 2. Example of an image-based communication channel 'Solo Room'

2.2 숏폼 콘텐츠

한국의 스마트폰 보급률은 약 95%로 전세계에서 가장 높은 수치를 보이며, 스트리밍 기술의 발달로 인해 사람들은 끊어짐 없이 수시로 영상을 시청할 수 있게 되었다. 그 결과 최근 대표적인 동영상 공유 플랫폼인 틱톡, 인스타그램, 유튜브 등을 통해 숏폼 콘텐츠의 소비는 급증하며 콘텐츠 소비문화가 정착하였다[6]. 이와 더불어 여가활동이 다양해지고 소비할 콘텐츠 양이 방대해지면서 콘텐츠 소비에 있어서 시간적인 효율을 추구하는 경향이 나타나면서 효율적으로 콘텐츠를 소비하는 것을 선호하게 되었다. 또한 동영상 콘텐츠의 재생 시간이 단축되고 있으며 유튜브의 경우 5~10분의 짧은 영상을 게시하는 채널의 인지도가 높아지고 있으며 상위 채널의 평균 재생 시간은 12분인 것으로 나타났다.

2.3 자막 영역 인식

기존 연구들로 이미지에 포함된 자막 영역을 식별 및 추출하는 알고리즘들이 제안되었다. 먼저 통계학적인 방법을 이용하여 색상극성을 결정하고 이에 따라 잡음을 제거하여 문자영역을 인식하는 방법은 지역 필터링(Region filling)과 색생 군집화(Color clustering) 기법을 사용하여 배경을 제거하고 문자영역을 추출한다[7]. 하지만 이러한 방안은 색상을 이용하여 문자영역의 여부를 결정하기 때문에 다양한 색상의 자막이 사용되면 정확도가 떨어질 수 있다. [8]은 배경과 자막 사이에 천이 영역이 있다는 것을 인지하여 자막영역을 추출하는 방법이다. 이 방법은 자막영역 인식률이 뛰어나지만, 실제 예능 같은 TV 프로그램에서의 자막영역을 추출할 때 자막이 아닌 다른 플로팅 데이터까지 추출할 수 있다. 또한 예능과 같은 영상에 자막이 정형화된 텍스트 모양으로 존재하는 것이 아니라 이모티콘 또는 다양한 심볼이 추가되는 특수 자막 유형에 대한 추출이 어렵다. 그러므로 최근 유튜브 및 예능 분야 동영상 콘텐츠에서 빈번하게 출현되는 영상에서의 자막 및 특수 자막 추출 방법이 필요하다. [9]은 인공 신경망을 이용하여 자막 픽셀의 수와 에지의 중앙 픽셀의 밝기들의 표준 편차 등의 특징들로 학습을 하여 자막을 인식한다. [10]은 자막 영역을 인식하는 방법을 제안하여 자막 영역의 모양을 학습하여 예측한다. 또한 자막 영역을 효율적으로 인식할 수 있도록 일부 개선된 모델이 OCR API로 제안되었다[11]. [12]는 자막 텍스트의 모양 식별을 위해 텍스트의 형상을 기반으로 에지를 추출하고, 후보 자막 영역을 도출 및 레이블링을 통해 비자막 영역을 식별한다.

비록 자막의 모양을 고려하여 자막 영역을 식별하지만 특수 자막에 해당되는 데이터를 반복적으로 적용한 기계 학습 모델이 존재하지 않기 때문에 다양한 형태의 특수 자막 및 비자막 식별이 어렵다. 따라서 기존에 제안된 방법들은 이미지에 포함된

모든 텍스트 추출을 목표로 하여 자막 영역에 포함되지 않은, 즉 자막 영역 이외의 배경 및 사물에 포함된 문자 이미지, 텍스트 등을 자막으로 인식될 수 있다. 따라서 이미지에서 자막영역 식별하고 자막영역으로부터 일반 및 특수 자막 추출을 위한 방법이 필요하다.

III. 제안 시스템

이 장에는 제안 시스템의 전체적인 프로세스를 기술한다. 제안하는 숏폼 이미지 자막 추출 시스템은 3단계로 구성된다. 첫 번째 단계로는 영상 데이터를 수집하는 단계로 숏폼 콘텐츠 동영상으로부터 이미지를 추출한다. 두 번째 단계에서는 수집된 개별 이미지에 대해 모폴로지 연산을 수행하여 자막영역이라 추정되는 부분만을 잘라내어 자막 데이터를 생성한다. 세 번째 단계에서는 인공 신경망을 통해 자막/비자막 및 특수 자막 데이터를 분류하고 자막으로 분류된 자막 데이터를 Kakao OCR API[13]를 활용하여서 텍스트로 변환한다.

그림 3은 제안 시스템의 흐름도를 나타낸다.

3.1 프레임 수집

숏폼 콘텐츠 동영상을 프레임 별로 분리한다. 정확한 모폴로지 연산을 위해 선별된 프레임들을 1280x720로 고정한다. 동영상에 포함된 모든 프레임을 대상으로 자막 식별을 수행할 경우 중복 장면이 포함되며 연산 속도가 또한 느려질 수 있으므로 2초에 한 번씩 프레임을 추출한다.

3.2 자막영역 인식

기존에 이미지에서 문자 영역을 인식하는 다양한 알고리즘들이 제안되었다[14][15]. 그렇지만 기존방법들은 다양한 크기와 이모티콘을 사용하는 특수 자막을 인식을 못하는 경우가 발생한다. 따라서 제안 시스템은 특수 자막을 인식하기 위한 모폴로지 연산을 제안한다. 이를 위해 자막영역 인식은 다음과 같다. 먼저 1) RGB 이미지를 회색조(Gray scale) 변환 단계, 2) 기울기(Gradient) 연산을 통한 글자의 외곽선만 추출 단계, 3) 외곽선 중 특정 기준 값(threshold) 이상인 부분만 남기고 나머지 부분은 제거하는 이진 분류 단계, 4) 자막 부분을 하나의 영역으로 합치(Closing)는 단계, 마지막으로 5) 이미지에서의 윤곽(Contour)을 찾아 텍스트 영역 단위로 잘라내는 단계로 구성된다.

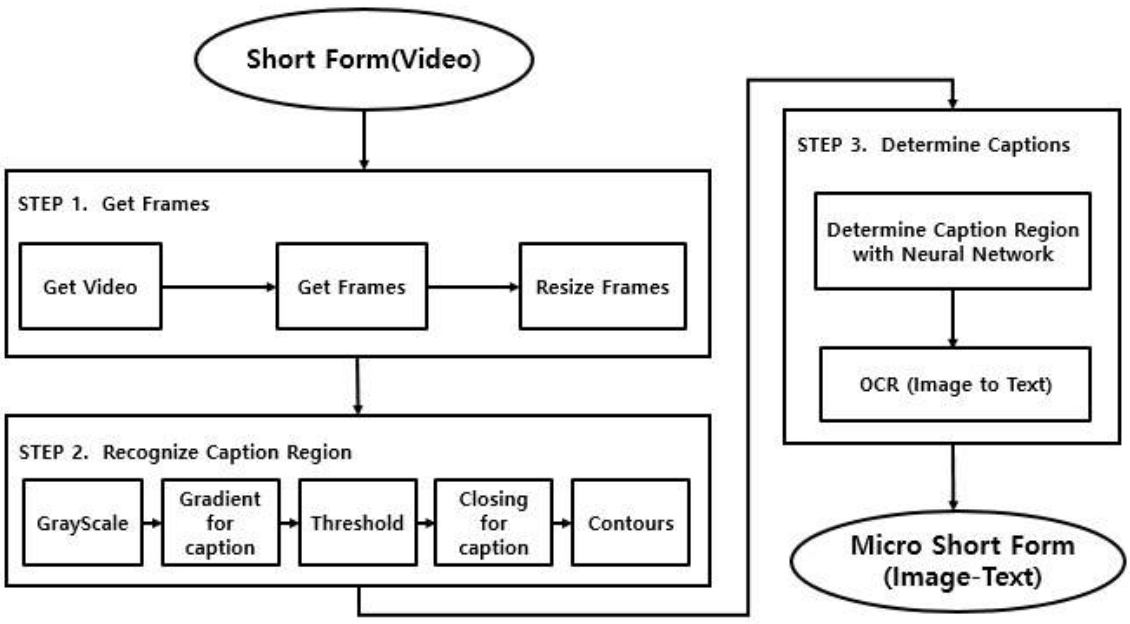

그림 3. 자막/비자막/특수 자막 추출 시스템 흐름도
Fig. 3. Flowchart for extracting subtitles

3.3 자막 판별

자막영역 인식 과정을 통해 자막영역이 인식된 데이터들에 대한 자막 여부 판별을 수행한다. 이를 위해 CNN 모델을 사용하여 개체 탐지, 이미지 분류, 랜드마크 식별에 널리 사용되는 MobileNet[16]을 이용한다. MobileNet으로 자막영역이 인식된 데이터들을 분류하고, 해당 영역이 자막이라 판단이 되면 OCR(Optical Character Recognition)을 적용하여 이미지에서 텍스트를 최종 출력한다. 이때 Kakao OCR을 이용하며 이미지와 텍스트 쌍은 키-값 형태로 저장한다. 키(Key)는 숏폼 이미지를 나타내며 값(Value)은 해당 이미지에 존재하는 자막에 표현한다.

IV. 실험 및 평가

4.1 실험 환경

제안 시스템의 실험을 위한 상세 환경은 표 1과 같다.

표 1. 실험 환경
Table 1. Experimental environment

Feature	Specification
OS	MacOS big sur
CPU	2.3GHz quadcore Intel core i5
Memory	16.00GB
GPU	Intel iris plus graphics 655 1536MB
Python	3.7
Opencv	4.5.1.48
OCR	Kakao OCR

4.2 실험 데이터

실험을 위한 자막 및 비자막 데이터 구축을 위해 숏폼 콘텐츠 동영상, 즉 7분~9분의 영상 50개(총 21,624 프레임)를 수집하였다. 이때 육안으로 자막 식별이 가능한 해상도인 1080이상의 화질의 영상을 수집하였다. 제안 시스템의 성능 평가를 위해 크게 두 가지 실험 데이터 집합으로 구분한다.

첫 번째 실험 데이터 집합은 기존 방법과 제안 시스템의 비교 평가를 위한 것으로 자막/비자막 데이터 각각 4,000개로 구성한다.

(a) 특수 자막과 비 자막
(a) Special caption and non-caption

(b) 일반 자막과 비 자막
(b) Caption and non-caption
그림 4. 자막 영영과 비 자막영역
Fig. 4. Caption region and Non-caption region

자막 데이터는 이미지에 존재하는 자막을 의미하며 하나의 이미지에는 여러 개의 자막을 포함할 수 있다. 비자막 데이터는 이미지에 포함된 임의의 텍스트를 의미한다. 또한 자막 데이터는 일반 자막과 특수 자막으로 구성한다. 특수 자막은 눈에 띄는 색상이나 크기, 다양한 그림 등을 통해 표현하여 정보 전달보다는 재미와 흥미 유도를 목적으로 사용하는 것을 뜻한다. 따라서 일반 자막과 특수 자막의 학습 데이터 집합과 테스트 데이터 집합은 각각 1,400개, 600개로 구성하고 비자막 데이터의 학습 및 테스트 데이터 집합은 각각 2,800개와 1,000개로 구성한다.

그림 4는 일반 자막, 특수 자막, 비자막을 낸다. 특수 자막은 그림 4의 (a) 녹색 영역에 해당하며 일반 자막은 그림 (b)의 녹색 영역과 같다. 비자막은 (a)와 (b)의 적색 영역에 나타난 텍스트를 나타낸다. 표 2는 실험용 자막(일반, 특수) 및 비자막 데이터를 학습데이터와 테스트 데이터로 구축한 건수를 보인다.

표 2. 실험용 자막/비자막 데이터 건수
Table 2. Number of experimental data set

Caption data \ Model		Train-set	test-set	all
Caption	Normal	1,400	600	2,000
	Special	1,400	600	2,000
Non-caption		2,800	1,000	4,000

두 번째 실험 데이터 집합은 모폴로지 연산의 성능을 실험 평가하기 위해 숏폼 콘텐츠로부터 추출한 이미지 1,500장으로 구성한다.

표 3. 영상 속 자막 데이터 집합
Table 3. Caption data set in Image

Category \ Caption	Caption		Non-caption
	Normal	Special	
Products	413	187	1,224
Entertainment	45	825	7,66
Game	0	650	4,534
Results	448	1,762	2,524

표 3은 위와 같은 방법으로 구축한 자막 데이터 건수를 나타낸다. 동영상에서 자막 또는 비자막의 개수가 편향될 수 있기 때문에 비교적 일반 자막이 많이 포함된 "제품 리뷰 영상", 특수 자막이 많이 포함된 "예능 영상", 비자막이 포함된 "게임 방송 영상"을 대상으로 데이터를 구축한다.

4.3 실험 평가

이 장에서는 제안 시스템의 성능을 평가하기 위해 두 가지의 방법으로 실험을 진행한다. 첫 번째로는 자막 유형에 따른 성능을 비교 평가 하는 것으로 이를 위해 4.2에서 정의된 첫 번째 실험 데이터 집합을 이용한다. 두 번째로 기존 방법과 제안 시스템을 비교하기 위해 4.2에서 정의된 두 번째 자막 데이터 집합을 이용한다.

4.3.1 다종 학습 데이터에 대한 자막 인식성능

첫 번째 실험은 일반 자막으로만 학습한 모델, 특수 자막으로만 학습한 모델, 일반 자막과 특수 자막 모두 포함하여 모델을 학습하고 학습 데이터에 따른 자만 인식 성능을 비교한다.

표 4. 각 데이터 집합에 대해 학습 시 정확도
Table 4. Accuracy at learning for each dataset.

Captions \ Model		Normal captions	Special captions	Total
Normal	Number	582/600	570/600	588/600
	Accuracy	0.97	0.95	0.98
Special	Number	252/600	590/600	587/600
	Accuracy	0.42	0.983	0.978
Normal + special	Number	415/600	578/600	583/600
	Accuracy	0.69	0.963	0.971

표 4는 일반 자막, 특수 자막, 일반자막 특수자막 혼용된 데이터에 대한 각각의 학습 정확도를 나타낸다. 학습 결과 일반 자막으로만 학습하였을 경우 일반 자막은 97%의 정확도로 식별하지만 특수 자막은 42%로 정확도가 낮아짐을 알 수 있다. 또한 특수 자막만을 대상으로 학습하였을 때 일반 자막에 대한 식별 정확도는 95%로 일반 자막 데이터만으로 학습한 결과와 성능이 비슷하지만, 특수 자막 데이터에 대한 정확도는 98%로 높다. 마지막으로 일반 자막과 특수 자막을 함께 모델에 학습한 경우는 97%의 정확도를 보이며 일반 자막, 특수 자막 각각을 개별로 학습한 모델보다 학습 성능이 향상됨을 알 수 있다.

그림 5는 신경망 모델의 자막 데이터 건수를 증가시키며 측정된 자막 인식 정확도를 나타낸다. 자막 데이터 개수가 초기 500개, 800개, 1,000개일 때의 정확도가 각각 87%, 94%, 97%로 점차 높은 성능 정확도를 보인다.

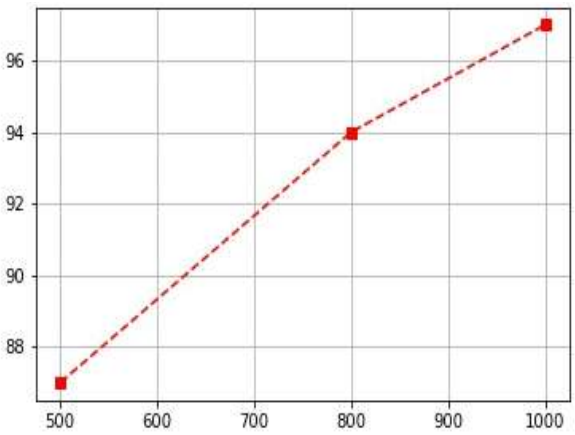

그림 5. 자막 데이터 건수에 대한 정확도
Fig. 5. Accuracy for the number of caption data sets

4.3.2 모폴로지와 신경망 정확도

두 번째 실험은 모폴로지와 신경망 학습이 포함된 제안 모델과 Kakao OCR과의 자막 인식성능을 표 5에서 비교한다. Kakao OCR은 이미지로부터 문자 영역과 해당 문자영역의 문자를 추출하는 기능을 제공하고 있다. 이때 제안 시스템에 대한 성능 측정은 모폴로지 연산만 수행하였을 경우와 모폴로지와 신경망을 함께 적용한 경우 2가지로 구분하여 정확도를 측정한다. Kakao OCR의 경우 문자 영역에서의 문자 추출의 정확도(97%)는 높게 나타났지만 재현율(0.26%) 은 낮다. 또한 제안 모델에서 모폴로지 연산만을 사용한 재현율(28%)은 모폴로지 연산과 인공신경만을 함께 적용한 재현율(86%) 보다 58%p 낮음을 알 수 있다. 그러므로 자막의 모양과 함께 일반 및 특수 자막의 데이터를 통한 인공신경망 학습이 재현율에 영향을 주는 것을 알 수 있다.

표 5. 자막 및 특수 자막 인식 성능 비교 결과
Table 5. Performance comparison results for subtitle and special subtitle recognition

Method		Captions	Normal	Special	Results
Kakao OCR[10]		Precision	0.97	0.98	0.97
		Recall	0.07	0.22	0.26
		Accuracy	0.07	0.22	0.26
Ours	Morphology	Precision	0.98	0.98	0.98
		Recall	0.07	0.22	0.28
		Accuracy	0.07	0.22	0.28
	Morphology +neural network	Precision	0.96	0.98	0.97
		Recall	0.52	0.84	0.86
		Accuracy	0.95	0.95	0.95

표 6. 비자막 인식에 대한 성능 비교 결과
Table 6. Performance results of the proposed method to non-caption recognition

Method	Caption	Non-captions
Kakao OCR[10]	Precision	0.512
	Recall	0.413
	Accuracy	0.505
Ours	Precision	0.965
	Recall	0.949
	Accuracy	0.958

표 6은 비자막 인식에 대한 비교 평가 결과를 나타낸다. 모폴로지와 인공신경망을 함께 적용한 제안 모델의 정확도(97%)와 재현율(95%)이 Kakao OCR 보다 정확도에서는 46%p, 재현율에서는 54%p 높음을 알 수 있다. 따라서 제안 방법이 비자막 식별에서 그 성능이 우수함을 알 수 있다.

표 7. 비자막 인식에 대한 성능 비교 결과(혼동행렬)
Table 7. Performance results of the proposed method to non-caption recognition in Confusion matrix

Method	Caption	Non-captions	
		Negative	Positive
Kakao OCR[13]	False	1,472	989
	True	1,474	1,041
Ours	False	2,583	89
	True	131	2,477

표 7은 비자막 인식성능 결과를 혼동행렬로 보인다. 제안 시스템이 카카오 OCR 보다 맞는 것을 올바르게 예측한 결과(TP, True Positive)가 높게 나타났으며 맞는 것을 틀렸다고 잘못 예측한 결과(FN, False Negative), 틀린 것을 정답으로 잘못 예측한 결과(FP, False Positive)가 제안 모델이 낮게 나타내는 것으로 제안 시스템의 성능이 우수한 것으로 판단할 수 있다.

Ⅴ. 결론 및 향후 과제

스마트 폰을 통한 동영상 콘텐츠 소비가 급증하고 수많은 동영상 콘텐츠가 생성되고 있다. 따라서 보다 효율적인 콘텐츠 소비를 위한 활동이 MZ세대를 통해 일종의 트랜드로 자리잡고 있다. 그러므로 시청 시간이 짧지만 핵심 내용과 흥미를 이끌어 낼 수 있는 숏폼 콘텐츠가 더욱 중요한 역할을 하게 되었다. 특히 비대면 온라인을 통해 의사소통을 할 수 밖에 없는 코로나 19 판데믹 상황에서 온라인 의사소통 채널에서의 의사소통 방법은 더욱 중요하게 되었다.

따라서 기존의 텍스트 형태만을 공유하던 소통에서 특정 상황에 맞는 이미지 파일을 공유함으로써 텍스트로 표현하기 어려운 감정 및 메시지를 담아서 의사소통할 수 있게 되었다. 이를 위해서 숏폼

콘텐츠 들은 기존의 정형화된 텍스트 자막에서 벗어나 다양한 형태로 표현되는 특수 자막을 사용하여 콘텐츠에 대한 흥미와 집중을 유도하게 되었다. 그러므로 특수 자막을 빠르게 검색하고 재사용하는 것 또한 의사소통에서의 한 가지 새로운 트랜드로 빠르게 자리매김 하고 있다.

본 논문에서는 숏폼 동영상 콘텐츠로부터 일반 및 특수 자막을 식별하고 식별된 이미지의 텍스트를 획득하는 시스템을 제안하였다. 제안 시스템을 통해 특수한 형태의 자막에 포함된 텍스트를 추출할 수 있게 됨으로써 텍스트로만 상황 묘사가 아닌 이미지 기반의 상황 묘사를 지원하게 되었다. 기존의 제안 방법과 비교하여 높은 정확도를 보였다. 제안 시스템을 기반으로 콘텐츠 큐레이션, 사용자 맞춤형 동영상 추천 등에 활용할 수 있을 것으로 기대된다. 따라서 중복 자막 제거와 함께 자막 추출 시간 최적화를 향후 연구로 한다.

References

[1] A. Kim and J. H. Han, "A Study of the way of choosing a emoticon on Mobile Instant Messenger", Proceedings of Korea HCI Conferences, pp. 435-439, Mar. 2020.

[2] S. Y. Bae and S. H. Kwon, "Communication Style in ‘Solitude Room’ in Social Media: Communication with an Image in the Space of Anonymity", Journal of Research in Curriculum Instruction, Vol. 24, No. 3, pp. 269-308, Mar. 2020.

[3] H. Kim, S. Oh, and S. Cho, "A study on short-form trend", Excellence Marketing for Customer, Vol. 54, No. 7, pp. 60-69, Jul, 2020.

[4] B. Chun, "A Study on Analyzing Caption Characteristic for Recovering Original Images of Caption Region in TV Scene", Jounal of JIIBC, Vol. 10, No. 4, pp. 177-182, Aug. 2010.

[5] H, Lim, "90's are coming", Whalebooks, pp. 285-298, 2018.

[6] J. Lee, "A Study on Types of Short-form Video Contents", Humanities Contents, Vol. 58, pp. 121-139, Sep. 2020.

[7] J. Kim, S. Kim, and Y. Moon, "Study on video character extraction and recognition", Proceedings of the IEEK Conference, Vol. 24, No. 1, pp. 141-144, Jun. 2001.

[8] W. Kim and C. Kim, "A new approach for overlay text detection from complex video scene", Journal of Broadcast Engineering, Vol. 13, No. 4, pp. 544-553, Jul. 2008. https://doi.org/10.5909/JBE.2008.13.4.544

[9] I. Jang, B. Ko, K. Kim, and H. Byun, "Automatic Text Extraction from News Video using Morphology and Text Shape", Journal of KISS : Computing Practices, Vol. 8 No. 4, pp. 479-488, Aug. 2002.

[10] J. Jeong, T. Yun, D. Kim, and J. Lee, "Connected Component-based Regardless of Caption Size Caption Extraction with Neural Network", Proceedings of Autumn Conference, Vol. 17, No. 7, pp. 172-175, Nov. 2007.

[11] X. Zhou, C. Yao. H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, "EAST: an efficient and accurate scene text detector", Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 5551-5560, Jul. 2017.

[12] H. Yassin and L. J. Karam, "Morphological text extraction from images", IEEE Transactions on Image Processing, and Image Processing Vol. 9. No. 11, pp. 1978-1983, Nov. 2000.

[13] Kakao OCR, https://vision-api.kakao.com/#ocr, [accessed: May. 4, 2021]

[14] M. Liao, Z. Wan, C. Yao, K Chen, and X. Bai, "Real-time Scene Text Detection with Differentiable Binarization", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 7, pp. 11474-11481, Apr. 2020. https://doi.org/10.1609/aaai.v34i07.6812.

[15] B. Shi, X. Bai and C. Yao, "An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.

39, No. 11, pp. 2298-2304, Nov. 2017. https://doi.org/10.1109/TPAMI.2016.2646371.

[16] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications", 2017. arXiv preprint arXiv:1704.04861.

저자소개

조 준 영 (Junyoung Jo)

2019년 ~ 현재 : 군산대학교 소프트웨어융합공학과 학부생
관심분야 : 데이터 마이닝, 자연어처리, 강화학습, 딥러닝

김 장 원 (Jangwon Gim)

2005년 : 상명대학교 컴퓨터소프트웨어공학과(학사)
2008년 : 고려대학교 컴퓨터학과(석사)
2012년 : 고려대학교 컴퓨터·전파·통신공학과(박사)
2013년 : 한국과학기술정보연구원 선임연구원
2017년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 텍스트 마이닝, 빅데이터 분석, 메타데이터, 공공데이터, 지식 그래프

온 병 원 (Byung-Won On)

2007년 : The pennsylvania state U. 컴퓨터공학과(박사)
2008년 : 캐나다 U. of British Columbia 박사 후 연구원
2010년 : U. of Illinois at Urbana-Champaign ADSC 연구소 선임연구원
2011년 : 차세대융합기술연구원 공공데이터 연구센터 센터장
2014년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터 마이닝, 정보검색, 빅데이터, 인공지능

정 동 원 (Dongwon Jeong)

1997년 : 군산대학교 컴퓨터과학과(학사)
1999년 : 충북대학교 전산학과(석사)
2004년 : 고려대학교 컴퓨터학과(박사)
2005년 ~ 현재 : 군산대학교 소프트웨어융합공학과 교수
관심분야 : 데이터베이스, 시맨틱 서비스, 빅데이터, 사물인터넷, 엣지컴퓨팅