특집논문 (Special Paper)
방송공학회논문지 제27권 제4호, 2022년 7월 (JBE Vol.27, No.4, July 2022)
https://doi.org/10.5909/JBE.2022.27.4.519
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

미디어 편집을 위한 인물 식별 및 검색 기법

박 용 석[a)], 김 현 식[a)‡]

Character Recognition and Search for Media Editing

Yong-Suk Park[a)] and Hyun-Sik Kim[a)‡]

요 약

동영상 콘텐츠 편집 시 등장인물을 구분하고 식별하는 작업은 많은 시간과 노력이 요구되는 작업이다. 노동 집약적 특성이 있는 미디어 편집 작업 시 인공지능 기술을 활용하면 미디어 제작 시간을 획기적으로 줄일 수 있어 창작과정의 효율성 향상에 도움을 줄 수 있다. 본 논문에서는 동영상 편집을 위한 인물 식별 및 검색 작업을 자동화하기 위해 다수의 인공지능 기술을 혼합하여 활용하는 기법을 제안한다. 객체 검출, 얼굴 검출, 자세 예측 기법을 사용하여 인물 객체에 대한 특징 정보를 수집하고, 수집된 정보를 바탕으로 얼굴 인식, 색 공간 분석 기법 등을 활용하여 인물 객체 식별 정보를 생성한다. 인물 특징 및 식별 정보는 편집 대상 영상의 각 프레임에 대해서 수집되며 영상 편집을 위한 프레임 단위 검색을 위한 메타데이터로 사용된다.

Abstract

Identifying and searching for characters appearing in scenes during multimedia video editing is an arduous and time-consuming process. Applying artificial intelligence to labor-intensive media editing tasks can greatly reduce media production time, improving the creative process efficiency. In this paper, a method is proposed which combines existing artificial intelligence based techniques to automate character recognition and search tasks for video editing. Object detection, face detection, and pose estimation are used for character localization and face recognition and color space analysis are used to extract unique representation information.

Keyword : Video editing, Character recognition, Object detection, Face recognition, Feature extraction

Ⅰ. 서 론

소셜 미디어(social media) 기반의 동영상 공유 플랫폼이 오락과 휴식 그리고 정보와 지식 습득의 중심으로 자리 잡으면서 다양한 분야에서 미디어 콘텐츠의 활용이 증가하고 있다. 동영상 공유 플랫폼에 콘텐츠 제공을 통해 수익을 창출할 기회가 발생하면서 많은 사람이 동영상 콘텐츠 제작

a) 한국전자기술연구원 콘텐츠응용연구센터(Contents Convergence Research Center, Korea Electronics Technology Institute)
‡ Corresponding Author : 김현식(Hyun-Sik Kim)
E-mail: hskim@keti.re.kr
Tel: +82-2-6388-6613
ORCID:https://orcid.org/0000-0002-8552-6751
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00804, Media production technology using learning based directing methods).
· Manuscript May 17, 2022; Revised July 6, 2022; Accepted July 13, 2022.

에 관심을 두고 참여하고 있다[1]. 이와 같은 미디어 콘텐츠 제작은 기획, 촬영, 편집 등 전 과정에 걸쳐 많은 시간과 노력이 요구되는 노동 집약적 특성이 있다. 동영상 콘텐츠 편집 단계에서 등장인물을 구분하고 식별하는 작업은 많은 시간과 노력이 요구되는 작업 중 하나이다. 인위적 연출이 없는 일상 환경에서 촬영된 영상의 경우, 배경 장면에 불특정 인물들이 포함될 수 있다. 이 같은 경우 사생활 보호 및 초상권 침해 방지를 위해 배경 인물에 대하여 비식별화(de-identification) 작업이 진행될 수 있으며, 배경 인물을 구분할 필요성이 발생한다. 등장하는 주역 인물에 대한 구분 및 식별은 주역 인물이 등장하는 장면의 편집 및 인물별 출연 분량을 확인하는 데 필요하다. 등장인물을 구분하고 식별하기 위해 영상 프레임(frame) 단위로 인물 추적을 진행해야 하며, 이는 반복적이고, 고도의 집중력이 요구되고, 시간이 많이 소요되고, 오류나 누락이 빈번히 발생하는 작업이다.

인공지능(Artificial Intelligence, AI)을 활용한 지능적 제작, 편집 기술은 미디어 제작 시간을 획기적으로 줄여줄 수 있어 창작과정의 효율성을 향상할 수 있다[2]. 본 논문에서는 동영상 편집을 위한 인물 식별 및 검색 작업을 자동화하기 위해 다수의 AI 기술을 혼합하여 활용하는 기법을 제안한다. 인물 객체에 대한 특징 정보 수집을 위해 객체 검출(object detection), 얼굴 검출(face detection), 자세 예측(pose estimation) 기법을 사용하고, 인물 객체 식별 정보 수집을 위해 얼굴 인식(face recognition) 및 색 공간 분석(color space analysis) 기법을 적용한다. 인물 특징 및 식별 정보는 편집 대상 영상의 각 프레임에 대해서 수집되며 영상 편집을 위한 프레임 단위 검색(query)용 메타데이터(metadata)로 사용된다. 이를 활용하여 편집 작업자는 특정 조건의 인물 정보를 검색하고 해당 조건을 충족하는 영상 프레임을 찾아 원하는 편집작업을 진행할 수 있게 된다.

Ⅱ. 관련 연구

AI 기반 영상 편집은 추출하고자 하는 정보에 따라 다양한 AI 기술이 적용될 수 있다. 많이 연구되고 활용되고 있는 기술에는 객체 검출 및 추적, 얼굴 검출 및 인식, 장면 검출(scene detection), 감성 분석(sentiment analysis), 영상 캡션(video captioning) 등이 있다. 객체 검출 및 추적 기술은 관심 객체를 식별하고 위치를 파악하여 해당 객체의 프레임 간 움직임을 추적할 수 있다. 얼굴 검출 및 인식 기술은 영상에서 사람의 얼굴을 찾아내고 누구의 얼굴인지 식별할 수 있다. 장면 검출 기술은 프레임 구성 요소들의 시각(visual) 및 의미(semantics) 유사성을 분석하여 관련성이 있는 프레임들을 추출하는 기술이다. 감성 분석은 주어진 콘텐츠가 내포하는 감성(기쁨. 슬픔 등)을 유추해내는 기술이다. 영상 캡션 기술은 주어진 장면에 대해 구성요소 및 상황 등을 분석하여 자연어로 설명을 제공하는 기술이다.

객체 검출 기술은 딥러닝(deep learning)의 발전과 함께 합성곱 신경망(Convolutional Neural Network, CNN)이 적용되면서 성능이 크게 향상되었다[3]. 특히 ImageNet, Open Images Dataset 등과 같은 인공지능 학습을 위한 대규모 이미지 데이터 세트(dataset)가 다수 구축되면서, CNN 기반 객체 검출 기술 발전을 진전시켰다. 딥러닝 기반 객체 검출 기법은 region proposal과 classification 단계를 순차적으로 진행하는 2단계(2-stage) 구조를 가진 알고리즘에서 이를 동시에 진행하는 1단계(1-stage) 구조의 알고리즘이 개발되면서 속도와 정확도가 향상되었다. 대표적인 객체 검출 알고리즘으로 YOLO(You Only Look Once)가 있으며, YOLO는 이미지를 격자 그리드(grid)로 나누어 한 번에 객체의 위치와 클래스(class)를 판단하고 최종 객체를 구분하기 때문에 실시간 운용과 연산 경량화가 가능하다.

얼굴 검출 기술은 이미지에서 사람의 얼굴 영역(위치와 크기)을 감지하며 얼굴 인식, 표정 인식 등 다수 기법을 진행하기 위한 기반 기술로 활용된다. 기존에는 Viola-Jones 알고리즘이 많이 사용되었으나 최근에는 다양한 얼굴 크기, 비율, 각도, 표정, 조도, 가림 현상(occlusion) 등에 효과적으로 대응할 수 있는 CNN 기반 알고리즘이 많이 사용되고 있다. 얼굴 검출 모델은 범용 객체 검출 모델을 얼굴 검출에 특화되도록 수정하여 학습시킨 것이 많다[4]. 대표적인 얼굴 검출 알고리즘으로 MTCNN(Multi Task Cascaded Convolutional Network)이 있으며, MTCNN은 눈, 코, 입의 좌표를 알아내는 얼굴 기울기 조정(face alignment) 과정과 얼굴 위치를 나타내는 테두리의 미세 조정(bounding box regression) 과정을 동시에 학습하여 높은 검출 정확도를 달

성한다.

얼굴 인식 기술은 사람의 얼굴이 가지는 고유 특징을 분석하여 동일인 여부를 판단할 수 있게 해주는 기술이다. 딥러닝 기반의 얼굴 인식은 사람 얼굴 이미지를 사용하여 CNN 모델을 학습하고, 일반적으로 분류(classification) 계층 전 단계 계층의 activation을 입력 데이터를 encoding 하는 기술자(descriptor)로 사용한다[5]. Descriptor는 벡터 형태로 표현(vector representation)되며 얼굴 이미지 간 구분 및 식별을 위해 유크리드 거리(Euclidean distance) 또는 코사인 유사도(cosine similarity) 등의 거리 척도를 사용한다. 페이스북(Facebook, 현재 Meta)이 개발한 DeepFace 얼굴 인식 모델의 경우, 8개의 CNN 계층으로 구성되며, 152x152 픽셀 크기의 얼굴 이미지를 입력으로 사용하며, 4096 차원(dimension) 크기의 벡터를 descriptor로 생성한다. 현재 대다수 딥러닝 기반 얼굴 인식 모델은 정면 얼굴(frontal face)을 기반으로 학습되기 때문에, 얼굴 회전으로 인하여 각도가 변경되어 측면 얼굴(side-view face) 등이 입력되면 인식률이 현저히 저하된다. 이런 단점을 개선하기 위해 얼굴 정면과 측면 모습을 함께 인식할 수 있는 알고리즘 연구도 진행되고 있다[6].

자세 예측 기술은 신체의 특징점(keypoint)을 구분하여 사람이 취하고 있는 자세를 예측한다. 자세 예측은 행동 인식(action or activity recognition)을 위한 기반 기술로 활용된다. 자세 예측은 일반적으로 머리, 어깨, 팔꿈치, 손목, 골반, 무릎, 발목 등 관절 특징점을 구분하는 단계와 추출된 관절 특징점을 연관 지어 유효한 자세 설정(configuration), 즉 뼈대구조(skeleton)를 구축하는 단계로 구분된다[7].

영상 객체 추적은 다수의 연이은 프레임에서 특정 관심 객체의 위치를 파악하는 작업을 진행하는 것이다. 객체 추적은 관심 객체 특징점의 이전 위치 및 움직임의 방향을 기반으로 다음 위치를 예측하고, 예측한 위치에 관심 객체가 있는지 확인하는 과정으로 구성된다. 영상에서 객체 추적을 위해 상관 필터(correlation filter)를 이용한 객체 추적 알고리즘이 많이 사용되었다[8]. 하지만 이러한 방식은 추적 대상의 크기(scale) 변화에 대처하기 어렵고, 전체 또는 부분 가림 현상이 발생 시 추적이 안 되는 현상이 발생한다. 이를 보완하기 위해 관심 객체 검출 단계에서 CNN 기반 모델을 적용하여 사용하는 추적 방법이 등장하였다[9]. 편집되는 영상은 다양한 배경을 가지는 다수의 장면으로 구성될 수 있으며, 이런 경우 장면의 선형성이 보장되지 않기 때문에 관심 객체를 추적하기 위해 기존의 추적 알고리즘을 적용하기 어려울 수 있다. 영상 편집 시 편집 배경 장면이 하나로 고정되어 있지 않은 경우가 많고, 다른 여러 장면이 빠르게 교차 전환되는 상황도 발생하기 때문에, 프레임 단위로 관심 객체를 직접 검출하는 것이 더 효율적일 수 있다.

Ⅲ. 제안 기법

동영상 편집을 위한 인물 식별 및 추적 작업을 자동화하기 위해 다수의 AI 기술을 혼합하여 활용하는 기법을 제안

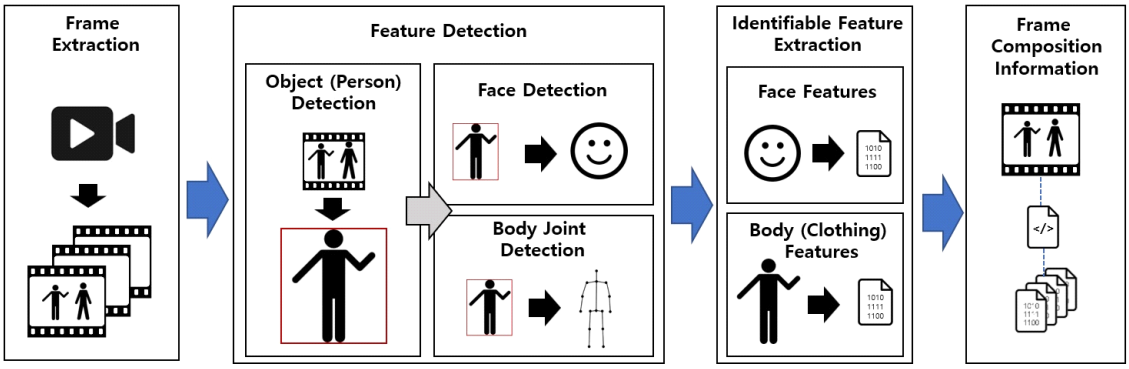

그림 1. 제안 기법 개념도
Fig. 1. Conceptual diagram of the proposed method

한다. 제안된 기법의 전체적인 개념적 구성은 그림 1과 같다. 편집 대상 영상을 구성하는 프레임들을 모두 추출한 다음, 각 프레임에서 객체 검출 알고리즘을 이용하여 인물(사람)의 영역(bounding box)를 검출한다. 검출된 인물 영역에서 얼굴 검출과 신체 주요 관절 특징점 검출을 실행하고 관련 좌표를 수집한다. 수집된 얼굴 및 특징점 정보를 이용하여 식별 가능한 얼굴 및 의상의 특징점 정보를 생성한다. 각 프레임 단위로 생성된 구성 인물 정보 및 각 인물에 해당하는 특징 및 식별 정보는 메타데이터 형태로 저장한다. 작업자는 생성된 메타데이터 정보를 검색하여 편집작업을 진행할 수 있다.

1. 특징 정보 수집

편집 영상에 등장하는 인물 객체에 대한 특징 정보 수집을 위해 그림 2와 같이 객체 검출, 얼굴 검출 그리고 자세 예측 기법을 사용한다. 편집 대상 영상을 프레임 단위로 분할 한 다음 각 프레임에 대해 사람 객체 검출을 실행하고 사람이 존재하는 영역에 대한 bounding box 정보를 수집한다. 각 사람 bounding box 이미지 영역에 대해서 얼굴 검출 알고리즘을 실행하여 얼굴 영역에 해당하는 bounding box를 구하고, 자세 예측 알고리즘을 실행하여 신체의 특징점 정보를 수집한다.

사람 객체 검출을 위해 YOLOR(You Only Learn One Representation) 알고리즘을 사용한다[10]. YOLOR은 explicit knowledge와 implicit knowledge를 함께 활용하여 다양한 이미지 처리 작업을 가능하게 하는 알고리즘으로 객체 검출 작업의 경우 처리 속도와 정확도 측면에서 Scaled-YOLOv4 알고리즘보다 좋은 성능을 나타낸다. 사람 객체 영역에서의 얼굴 검출은 RetinaFace 알고리즘을 사용한다[11]. RetinaFace는 추가학습과 자가지도 다중작업 학습(extra-supervised and self-supervised multi-task learning) 기법을 이용하여 밀집된 군중에서 작은 크기의 얼굴도 검출할 수 있도록 특화되어있어 세밀한 편집 작업 시 용이하다. 검출된 사람들의 자세 예측을 위해 BlazePose 알고리즘을 사용한다[12]. BlazePose는 실시간으로 동작할 수 있도록 경량화 설계되었으며, 검출 과정에서 총 33개의 신체 및 관절 특징점을 생성한다. BlazePose는 모든 관절의 heatmap을 예측한 다음 regression을 이용하여 각 관절의 좌표를 예측하여 경량화된 자세 검출이 가능하게 하였다.

2. 식별 정보 생성

인물 객체 식별 정보 생성을 위해 그림 3과 같이 얼굴 인식 및 색 공간 분석 기법을 적용한다. 일반적으로 인물 식별을 위해 얼굴 인식 정보를 단독으로 사용해도 충분하지만, 동영상에서는 등장인물의 움직임에 따라 얼굴의 방향과 크기 변화로 인해 얼굴 인식이 불가능할 수 있다. 이를 보완하기 위해 인물이 착용하고 있는 의상의 색상 정보를 함께 사용한다. 이전 단계를 통해 수집된 정보 중 얼굴 영역

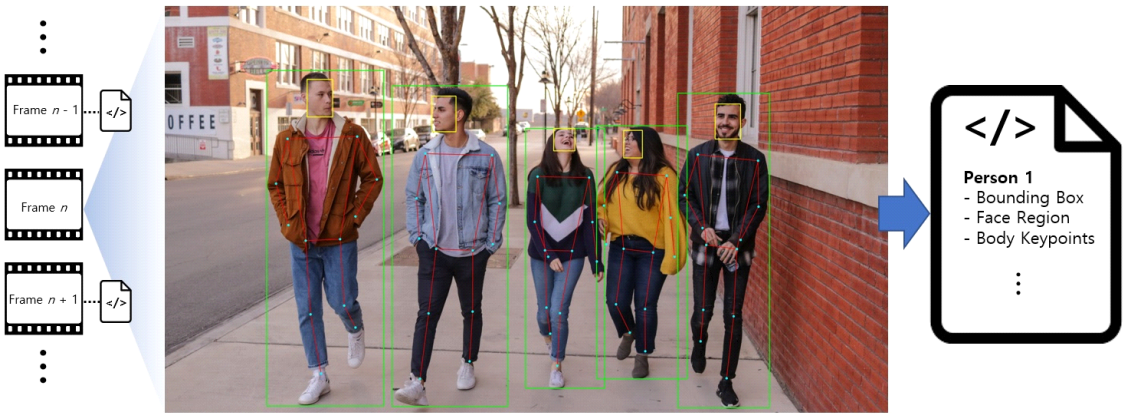

그림 2. 인물 객체 특징 정보 추출
Fig. 2. Character feature information extraction

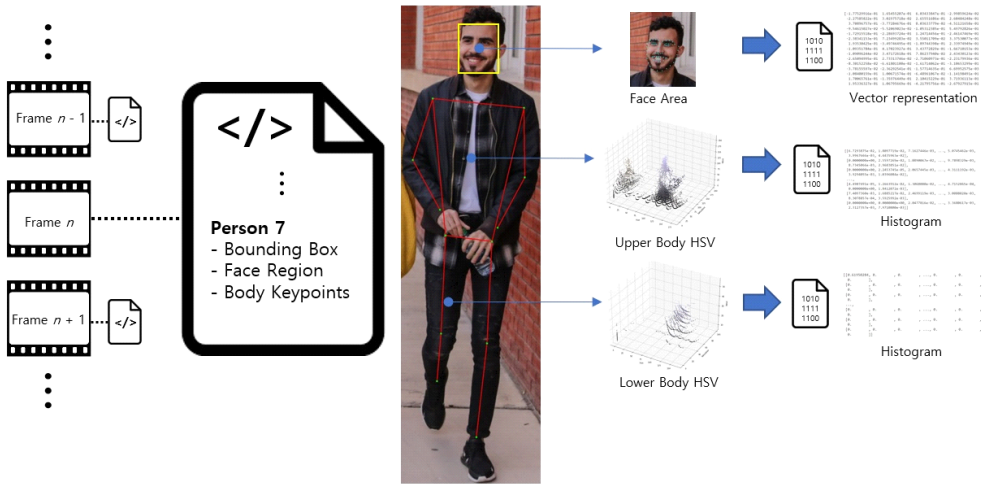

그림 3. 인물 객체 식별 정보 생성
Fig. 3. Character identity information generation

bounding box는 얼굴 인식을 위해 사용되며, 신체 특징점은 상체, 하체 영역을 구분하고 해당 영역의 색상 특징을 추출하기 위해 사용된다. 얼굴 인식을 통해 얼굴의 특징을 나타내는 vector representation을 생성하고, 색상 분석을 통해 의상의 색상 히스토그램(histogram) 정보를 생성한다. 식별 정보 생성은 각 프레임에서 식별된 모든 사람 객체에 대해서 실행한다.

얼굴 인식을 위해 ArcFace 알고리즘을 사용한다[13]. ResNet32 모델을 기반으로 하는 ArcFace는 가산 각도 여백(additive angular margin loss)을 사용하여 얼굴 인식을 위한 식별 가능한 특징(discriminative features)을 보강한다. 얼굴 인식의 결과로 512 차원 크기의 벡터를 생성한다. 색상 정보 추출을 위해서는 색상(hue), 채도(saturation), 명도(value) 성분으로 변환하여 진행한다. HSV 색 공간은 인간이 색을 인지하는 방법과 유사하여, 조명 조건이 변해도 색상에는 영향을 주지 않아 색 추출이 용이하다. 생성되는 히스토그램은 색상과 채도 정보를 저장하며 180x256 크기의 배열로 구성된다.

3. 비교 및 검색

각 프레임에 대해서 인물 객체 정보 메타데이터를 추출 및 생성한 다음, 작업자는 생성된 정보를 이용하여 그림 4와 같이 관심 인물 검색 작업을 진행할 수 있다. 검색하고자

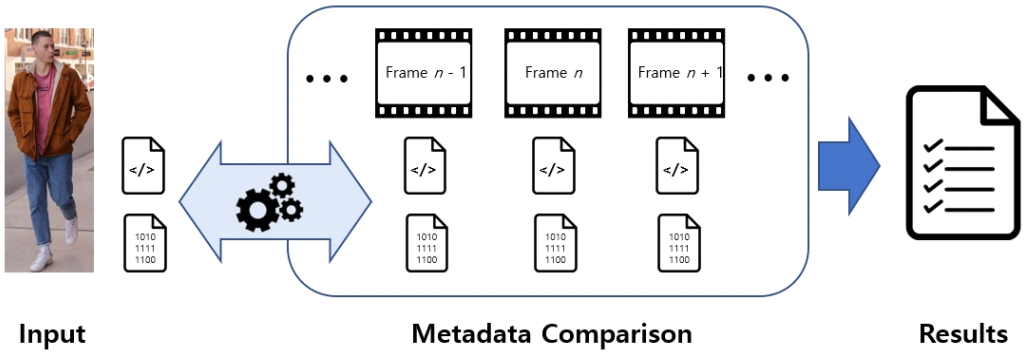

그림 4. 관심 인물 비교 및 검색 과정
Fig. 4. Character comparison and search process

하는 인물의 이미지를 입력으로 제공하면 해당 인물에 대해서 특징 정보를 추출하고 식별 정보를 생성한다. 생성된 식별 정보는 기생성된 프레임 메타데이터 정보와 비교되어 유사성 검사를 하고, 어떤 프레임에 인물이 등장하는지 결과가 출력된다.

얼굴 이미지 정보 비교 시 거리 척도(distance metric)를 사용하며 결괏값의 차이가 작을수록 동일 인물일 가능성이 크다. 유크리드 거리 척도를 사용하였으며 이는 (1)과 같다. 이때 p_i와 q_i는 각각 비교 대상 벡터의 값을 나타낸다.

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \tag{1}$$

색상 정보도 픽셀값의 분포가 서로 비슷하면 유사한 이미지일 확률이 높고, 분포가 다를 경우 서로 다른 이미지일 확률이 높다는 사실을 기반으로 유사도를 측정한다. 색상 히스토그램의 유사성 비교를 위해서는 상관관계(correlation metric)를 사용하며 이는 식 (2)와 같다.

$$d(I,M)=\frac{\sum_{j=1}^{n}(I_j-\bar{I})(M_j-\overline{M})}{\sqrt{\sum_{j=1}^{n}(I_j-\bar{I})^2\sum_{j=1}^{n}(M_j-\overline{M})^2}} \tag{2}$$

식 (2)에서 I와 M은 비교 대상 히스토그램이며 $\bar{I}$와 $\overline{M}$은 각각 I와 M의 평균(mean)값을 나타낸다. 상관관계의 경우 비교 결괏값이 클수록 유사성이 높다.

Ⅳ. 실험 결과 및 분석

영상 편집을 위한 관심 인물 식별 및 검색 시스템은 Intel Core i9 CPU@3.70GHz, NVIDIA GeForce RTX 3090

표 1. 인물 비교검색 테스트 결과(얼굴 특징만 사용)
Table 1. Character comparison and search test results (face features only)

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
True Positives	104	113	180	173	135
False Negatives	26	56	0	3	64
False Positives	0	0	0	0	0
True Negatives	285	398	274	404	801
Accuracy	93.73%	90.12%	100%	99.48%	93.6%
Recall	80%	66.86%	100%	98.30%	67.84%
Precision	100%	100%	100%	100%	100%
F1-score	88.89%	80.14%	100%	99.14%	80.84%

표 2. 인물 비교검색 테스트 결과(얼굴 및 의상 특징 사용)
Table 2. Character comparison and search test results (face and garment features)

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
True Positives	106	114	180	176	145
False Negatives	24	55	0	0	54
False Positives	0	0	0	0	0
True Negatives	285	398	274	404	801
Accuracy	94.22%	90.3%	100%	100%	94.6%
Recall	81.54%	67.46%	100%	100%	72.86%
Precision	100%	100%	100%	100%	100%
F1-score	89.83%	80.57%	100%	100%	84.3%

GPU, 128GB 메모리가 장착된 윈도우 운영체제 환경에서 구현 및 테스트하였다. 제안된 인물 식별 및 검색 기법을 테스트하기 위해 Full HD(1920x1080 픽셀) 해상도의 토크쇼 인터뷰 형식의 가편집 영상과 유튜브 영상을 사용하였다. 영상 30개를 선별하여 등장하는 인물들에 대하여 프레임 단위로 수작업으로 경계 박스를 설정하고 인물들을 라벨링하여 ground truth 데이터를 생성하였다.

관심 인물의 식별 및 검색 정확도를 확인하기 위해 영상에 등장하는 특정 인물의 이미지를 시스템에 입력으로 제공하고 해당 인물의 비교검색 결과 리포트를 반환받고, 리포트 결과를 ground truth 데이터와 비교하였다. 표 1과 2는 일부 테스트 영상에 대한 인물 비교검색 테스트 결과를 보여준다. 표 1은 얼굴 특징만 비교했을 때의 결과를 보여주며, 표 2는 얼굴과 의상 색상 특징을 모두 비교했을 때의 결과이다.

표에서 양성(True Positive, TP)은 실제로 동일 인물을 동일 인물로 맞게 판단한 경우를 의미하고, 음성(True Negative, TN)은 실제로 다른 인물을 다른 인물로 맞게 판단한 경우를 나타내며, 거짓 양성(False Positive, FP)은 실제로는 다른 인물인데 결과는 동일 인물로 틀리게 판단한 경우이며, 거짓 음성(False Negative, FN)은 실제로는 동일 인물인데 결과는 동일 인물이 아닌 것으로 틀리게 판단한 경우이다. 정확도(accuracy)는 올바르게 예측된 데이터 수(TP + TN)를 전체 데이터 수로 나눈 값이며, 재현율(recall)은 실제로 True인 데이터를 모델이 True로 인식한 데이터의 수(TP/TP+FN))를 나타낸다. 정밀도(precision)는 모델이 True로 예측한 데이터 중 실제로 True인 데이터의 수(TP/(TP+FP))이며, F1 score는 정밀도와 재현율의 조화평균을 나타낸다.

표 1과 2의 결과를 비교해 보면, 얼굴 특징 외 의상 색상 정보를 추가로 사용할 때 거짓 음성 결과가 감소하는 것을 볼 수 있다. 결과를 살펴보면 대체로 비교검색 결과가 정확한데, 거짓 음성 결과가 다수 발생하는 현상을 볼 수 있다. 이미지에서 인물들이 가까이 겹쳐있거나, 얼굴은 식별이 안 되는데 착용한 의상의 색 조합이 유사한 경우 거짓 음성 결과가 발생하였다. 이 같은 경우의 수에 대해 대안 마련이 필요하다. 그림 5는 거짓 음성이 발생한 이미지 예시를 보여준다.

그림 5. 거짓 음성 결과의 이미지 예시
Fig. 5. Sample images resulting in false negatives

V. 결 론

본 논문에서는 동영상 편집 시 편집 시간을 단축하고 작업 효율성 향상을 위해 AI 기반 기술을 활용하여 관심 인물 식별 및 검색을 자동으로 할 수 있는 기법을 제시하였다. 편집 대상 영상의 각 프레임에 대해서 객체 검출, 얼굴 검출, 자세 예측 기법을 이용하여 인물 객체의 특징 정보 메타데이터를 생성하고, 얼굴 인식 및 색 공간 분석을 통해 인물 식별 정보를 생성하였다. 편집자가 찾고자 하는 인물의 이미지를 입력으로 제공하면, 기생성된 인물 데이터의 비교검색을 통해 해당 인물이 등장하는 모든 프레임에 대한 정보를 제공한다. 실험 결과, 대다수의 경우 검색 대상이 있는 프레임을 성공적으로 찾을 수 있었다. 단, 얼굴이 식별되지 않고 다른 등장인물과 유사한 색상의 의상을 착용한 경우, 거짓 음성 결과가 발생하는 것을 확인할 수 있었으며, 향후 이에 대한 보완을 진행하여 검색 결과의 성능을 향상할 계획이다.

참 고 문 헌 (References)

[1] Q. Tang, B. Gu, and A. Whinston, “Content Contribution in Social Media: The Case of YouTube,” Proceeding of 2012 45th Hawaii International Conference on System Sciences, Maui, HI, USA, pp.

4476-4485, 2012.
doi: https://doi.org/10.1109/HICSS.2012.181

[2] T. Soe, “AI video editing tools. What editors want and how far is AI from delivering?” arXiv:2109.07809 [cs.HC], pp. 1-7, 2021. doi: https://doi.org/10.48550/arXiv.2109.07809

[3] L. Jiao et al., “New Generation Deep Learning for Video Object Detection: A Survey,” IEEE Transactions on Neural Networks and Learning Systems (Early Access), pp.1-21, Feb. 2021.
doi: https://doi.org/10.1109/TNNLS.2021.3053249

[4] Y. Feng, S. Yu, H. Peng, Y. -R. Li, and J. Zhang, “Detect Faces Efficiently: A Survey and Evaluations,” IEEE Transactions on Biometrics, Behavior, and Identity Science, Vol.4, No.1, pp.1-18, Jan. 2022.
doi: https://doi.org/10.1109/TBIOM.2021.3120412

[5] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, “Deep Face Recognition: A Survey,” Proceeding of 22018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, pp. 471-478, 2018.
doi: https://doi.org/10.1109/SIBGRAPI.2018.00067

[6] G. -S. Hsu and C. -H. Tang, “Dual-View Normalization for Face Recognition,” IEEE Access, Vol.8, pp.147765-147775, July 2020.
doi: https://doi.org/10.1109/ACCESS.2020.3014877

[7] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, “The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation,” IEEE Access, Vol.8, pp.133330-133348, July 2020.
doi: https://doi.org/10.1109/ACCESS.2020.3010248

[8] S. Du and S. Wang, “An Overview of Correlation-Filter-Based Object Tracking,” IEEE Transactions on Computational Social Systems, Vol.9, No.1, pp.18-31, Feb. 2022.
doi: https://doi.org/10.1109/TCSS.2021.3093298

[9] A. Gautam and S. Singh, “Trends in Video Object Tracking in Surveillance: A Survey,” Proceeding of 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 729-733, 2019.
doi: https://doi.org/10.1109/I-SMAC47947.2019.9032529

[10] C. -Y. Wang, I-H. Yeh, and H. -Y. Liao, “You Only Learn One Representation: Unified Network for Multiple Tasks,” arXiv:2105.04206 [cs.CV], pp. 1-11, 2021.
doi: https://doi.org/10.48550/arXiv.2105.04206

[11] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild,” Proceeding of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 5202-5211, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00525

[12] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, “BlazePose: On-device Real-time Body Pose tracking,” arXiv:2006.10204 [cs.CV], pp. 1-4, 2020.
doi: https://doi.org/10.48550/arXiv.2006.10204

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” Proceeding of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 4685-4694, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00482

저 자 소 개

박 용 석

- 1997년 5월 : 미국 Carnegie Mellon University 전기/컴퓨터공학과 학사
- 1998년 5월 : 미국 Carnegie Mellon University 전기/컴퓨터공학과 석사
- 2018년 2월 : 연세대학교 전기전자공학과 박사
- 1998년 6월 ~ 2000년 6월: 에스원 기술연구소 주임연구원
- 2000년 6월 ~ 2003년 11월: 아이앤씨테크놀로지 주임연구원
- 2003년 12월 ~ 현재 : 한국전자기술연구원 콘텐츠융합연구센터 책임연구원
- ORCID : https://orcid.org/0000-0002-6694-5125
- 주관심분야 : 영상처리, 컴퓨터비전, 인공지능

김 현 식

- 2002년 2월 : 인하대학교 전기공학과 학사
- 2004년 2월 : 인하대학교 정보통신공학과 석사
- 2017년 2월 : 연세대학교 전기전자공학과 박사
- 2004년 3월 ~ 현재 : 한국전자기술연구원 콘텐츠융합연구센터 책임연구원
- ORCID : https://orcid.org/0000-0002-8552-6751
- 주관심분야 : 블록체인, 실감미디어, 인공지능