딥러닝 기반 비디오 특징을 이용한 장면 검출 기법

고민수*, 송혁*
*한국전자기술연구원
{kmsqwet, hsong}@keti.re.kr

Scene Detection Method using Deep Learning-based Video Features

Min Soo Ko*, Hyok Song*
*Korea Electronics Technology Institute

요 약

스마트 기기, 인터넷 환경의 변화로 스트리밍 방식의 플랫폼 서비스가 증가하고 있다. 콘텐츠 제작자 또는 사용자들의 비디오 콘텐츠가 매순간 업로드 되고 있으며 축적되는 데이터양이 급격히 증가하고 있다. 다양한 비디오 콘텐츠의 급격한 증가로 인해 이를 효과적으로 분석할 수 있는 기술의 필요성이 높아지고 있다. 연속적인 행동, 상황 등으로 이루어진 비디오를 분석하기 위해서는 먼저 의미 있는 단위로 분리할 수 있어야 한다. 장면은 같은 시간과 공간에서 촬영된 영상들을 연결한 비디오의 단위를 의미한다. 비디오를 자동으로 장면 단위로 분리할 수 있다면 효과적인 비디오 분석에 도움이 될 수 있다.

본 논문에서는 딥러닝 기반의 비디오 특징을 이용한 장면을 검출하는 기법을 제안한다. 먼저 입력 비디오를 샷 경계 검출 네트워크를 이용하여 샷 단위로 분리한다. 분리된 샷의 중심 클립영상을 입력으로 비디오 특징 검출 네트워크 통해 샷 단위의 비디오 특징을 생성한다. 생성된 특징들 간의 유사도를 계산하고 시퀀셜 그룹핑을 통해 유사한 샷들을 연결하여 최종 장면이 분리되는 지점들을 검출한다. 실험을 통해 제안하는 기법이 0.5680의 성능을 보였으며, 기존 기법보다 우수함을 확인하였다.

Ⅰ. 서 론

인터넷 환경의 발전으로 OTT와 같은 스트리밍 기반의 디지털 플랫폼들이 차세대 미디어 콘텐츠 시장의 주역이 되고 있다. 이러한 변화로 인하여 다양한 편집 형태의 비디오 서비스들이 생겨나고 있으며, 매일 수많은 비디오 데이터가 축적되고 있다. 서비스의 차별화, 고도화를 위해 비디오 데이터를 분석하는 기술을 필요로 하고 있으나, 비디오는 이미지와 다르게 수많은 프레임으로 구성되어 있기 때문에 분석에 더 높은 수준의 기술이 필요하다. 따라서 이러한 비디오의 분석을 위해서는 이를 의미 있는 단위로 분리할 수 있는 기술이 먼저 적용되어야 한다.

비디오는 단위에 따라 프레임(frame), 샷(shot), 장면(scene), 시퀀스(sequence)로 나눌 수 있다. 프레임은 카메라로 촬영된 한 장의 이미지를 의미한다. 샷은 촬영을 시작하고 멈추는 동안의 영상을 의미한다. 장면은 같은 시간과 장소에서 일어나는 하나의 이야기를 의미하는 단위로 같은 시공간에서 촬영된 샷들을 연결하면 하나의 장면이 된다. 마지막으로 시퀀스는 장면들이 모여 만들어지는 에피소드 단위를 의미한다.

비디오를 단위로 분리하는 기술로는 샷 경계 검출(shot boundary detection) 기법과 장면 검출(scene detection) 기법이 있다. 최근 딥러닝 기술의 발전으로 기존 단일 이미지를 분석하는 기법을 확장해 연속적인 프레임의 비디오를 분석할 수 있는 기법들이 활발히 연구되고 있다. 이러한 발전으로 비디오를 분할하는 기술들 또한 딥러닝 기반의 특징 정보를 이용한 기법들로 연구되고 있으며, 기존의 사람에 의해 설계된 수제 특징(hand-crafted features)을 이용하는 기법들과 비교해서 큰 성능의 개선이 있는 것으로 발표되고 있다.

본 논문에서는 딥러닝 기반의 비디오 특징을 이용하여 장면을 검출하는 기법을 제안한다. 하나의 의미 있는 단위인 장면을 검출하는 기술은 비디오 요약, 비디오 편집, 비디오 검색 등의 많은 분야에 적용 가능할 것으로 기대된다.

Ⅱ. 제안하는 기법

그림 1은 제안하는 딥러닝 기반 비디오 특징을 이용한 장면 검출 기법의 흐름도를 나타낸다. 먼저 입력 비디오를 샷 경계 검출 네트워크를 통해 샷 단위로 분리한다. 분리된 샷 중심의 클립영상을 비디오 특징 추출 네트워크의 입력으로 하여 샷의 비디오 특징을 추출한다. 샷 간의 비슷한 정도를 추정하기 위해 샷 단위로 추출된 비디오 특징 간에 유사도를 계산하여 유사도 행렬(similarity matrix)을 생성한다. 마지막으로 유사한 샷들을 연결하는 시퀀셜 그룹핑(sequential grouping)을 통해 비디오를 장면 단위로 분리한다.

A. 샷 경계 검출 네트워크

샷 경계 검출 네트워크 학습을 위해 ClipShots 데이터 셋을 사용한다[1]. 영상이 갑자기 변화하는 하드 컷(hard cut)과 점진적인 변화(gradual transition)를 모두 검출 할 수 있으려면 다수의 프레임이 딥러닝 네트워크 에 한 번에 입력될 수 있어야 한다. 샷 검출은 아주 작은 이미지의 변화보다는 전체 이미지의 변화를 분석해야하기 때문에 입력의 해상도를 크게 줄이고 대신 많은 프레임을 입력으로 사용하도록 한다. 본 논문에서는 48×27 해상도에 100 프레임을 사용한다.

B. 비디오 특징 추출 네트워크

비디오 추출 네트워크로 SlowFast-ResNet101 구조를 이용한다. CNN을 통한 비디오 특징이 샷 클립영상의 특성을 잘 반영할 수 있도록 액션 검출 기법을 위한 학습 데이터인 AVA Action 데이터 셋을 이용하여 사전 학습을 진행한다[2]. 액션 검출을 학습한 네트워크에서 최종 액션 판단을 위한 레이어 부분을 제거하고 앞 단의 CNN 네트워크 부분만 이용한다. CNN을 통해 생성된 특징맵에 MAC(maximum activation of convolution)과 정규화를 통해 1차원의 비디오 특징을 생성한다.

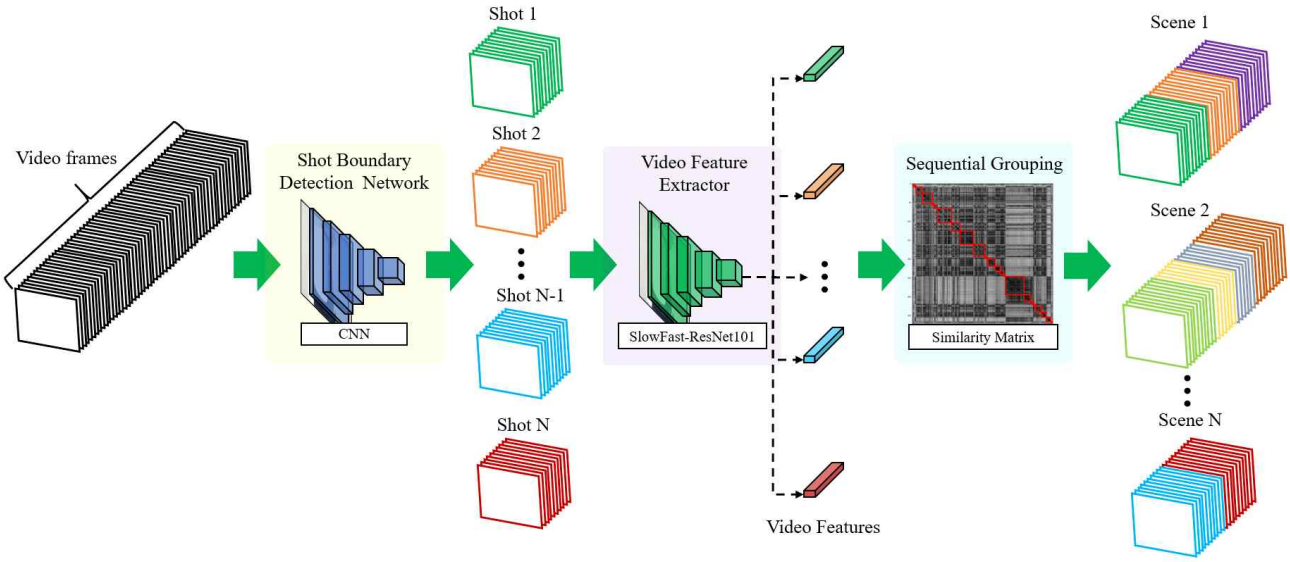

그림 1. 제안하는 딥러닝 기반 비디오 특징을 이용한 장면 검출 기법의 흐름도

C. 장면 지점 검출을 위한 시퀀셜 그룹핑

추출된 샷의 비디오 특징간의 모든 유사도를 계산하여 유사도 행렬을 생성한다. 유사도 행렬의 대각 방향으로 블록의 크기를 바꿔가며 유사도의 합을 이용하는 비용함수를 통해 비용을 계산하고, 비용이 최소가 되도록 최적화 한다. 최소가 되는 블록들을 유사한 샷들의 그룹으로 판단하여 최종 장면의 분리 지점들로 검출한다.

Ⅲ. 실험 결과

제안하는 기법의 성능을 평가하기 위해 OVSD(open video scene detection) 데이터 셋을 이용한다[3]. 이 데이터는 21개의 오픈소스 비디오와 사람이 직접 판단하여 작성한 장면 지점에 대한 정답 데이터를 제공한다. 성능평가 지표는 모델을 통해 추정되는 장면 검출 지점과 정답을 비교하여 각 비디오별로 F_1-score를 계산하고 21개 비디오에 대한 평균값을 계산한다.

표 1은 제안하는 기법과 기존 기법의 성능을 보여준다. OSG는 샷 키프레임의 HSV 히스토그램을 특징으로 이용한 결과이며, DL_Img_Feat은 ImageNet으로 사전 학습된 EfficientNet-B4 네트워크를 이용하여 생성된 이미지 특징을 이용한 기법이다[3, 4]. 결과와 같이 제안하는 기법의 성능이 우수함을 알 수 있다. 그림 2는 제안하는 기법을 통해 검출된 장면의 시작과 끝 프레임을 보여준다.

표 1. 제안하는 기법의 성능 비교

Method	F_1-score
Ours	**0.5680**
DL_Img_Feat[4]	0.5214
OSG[3]	0.46

Ⅳ. 결론

본 논문에서는 딥러닝 네트워크로 추출된 비디오의 특징 정보를 이용하여 장면을 검출하는 기법을 제안하였다. 장면을 검출하기 위하여 더 세부 단위인 샷으로 먼저 나누고, 비디오 특징 추출 네트워크를 통해 샷의 특징 정보를 생성하였다. 샷의 비디오 특징간의 유사도를 계산하고 이를 군집화 할 수 있는 시퀀셜 그룹핑을 통해 최종 장면 단위로 분리하였다. 제안하는 기법을 통해 비디오를 의미 있는 단위로 분리할 수 있었으며, 향후 이를 응용한 동영상 분석 시스템을 개발할 예정이다.

그림 2. 제안하는 기법의 장면 검출 결과
(좌: 시작 프레임, 우: 끝 프레임)

ACKNOWLEDGMENT

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)

참 고 문 헌

[1] Shitao T., Litong F., Zhanghui K., Yimin C., and Wei Z. “Fast Video Shot Transition Localization with Deep Structured Models,” Asian Conference on Computer Vision, pp. 577-592, Dec. 2018.

[2] Chunhui G., Chen S., David R., Carl V., Caroline P., Yeqing L., Sudheendra V., George T., Susanna R., Rahul S., Cordelia S., and Jitendra M. “AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions,” Computer Vision and Pattern Recognition, pp. 6047-6056, June 2018.

[3] Daniel R., Dror P., and Gal A. “Robust and Efficient Video Scene Detection Using Optimal Sequential Grouping,” International Symposium on Multimedia, pp. 275-280, Dec. 2016.

[4] Min Soo K., Hyok S., Jisang Y. “Scene Extraction Technology on Deep Learning for Media Production,” Broadcast and Media Engineers Summer Conference, pp. 153-154, June 2022.