영상의 장면 내 장소 및 시간 검출 모델

손정우, 이호재, 곽창욱, 김선중
미디어지능화연구실, 미디어연구본부, 통신미디어연구소
한국전자통신연구원
[jwson, lhjalex, cukwak, kimsj]@etri.re.kr

Location and Time Detection Model for Video Scene

Jeong-Woo Son, Alex Lee, Chang-Uk Kwak, Sun-Joong Kim
Media Intellectualization Research Section, Media Research Division
Telecommunications & Media Research Lab., ETRI

요 약

영화, 드라마, 다큐멘터리 등의 영상 콘텐츠는 다수의 스토리가 복합적으로 구성되어 시각 혹은 청각으로 전달된다. 이에 대한 분석은 작은 스토리 단위인 장면을 기반으로 전체 콘텐츠로 분석 범위를 확장하는 것이 일반적이다. 영상의 장면을 분석하기 위해서는 시청각 정보를 담고 있는 데이터에 대한 높은 수준의 분석을 요구하며, 이 과정에서 장면을 구성하는 객체, 액션 등의 단일 정보를 추출하게 된다. 본 논문에서는 장면에 내재된 정보 중, 장소와 시간 정보를 검출하는 모델을 제안한다. 장소와 시간은 장면을 정의하는 가장 기초적인 기준이며, 영상 콘텐츠 제작을 위한 대부분의 각본은 장소와 시간에 따라 장면을 구분한다. 제안한 모델에서는 영상 콘텐츠에서의 장면 및 시간 정보 추출을 위해 각본을 분석하여 영상 콘텐츠에 적합한 클래스를 정의하고, 영상 콘텐츠의 제작 기법을 고려한 모델 구조를 채택하였다. 실험을 통해 본 논문에서 제안하는 모델이 영상 콘텐츠의 특성을 고려하여 장소 및 시간 정보를 효과적으로 검출하고 있음을 검증하였다.

Ⅰ. 서 론

본 논문에서는 장면 단위 영상에서 장소 및 시간 정보를 검출하기 위한 모델을 제안한다. 영상의 장면을 결정짓는 주요 요소 중, 가장 기본적인 요소는 영상이 담아야하는 장소와 시간이다. 영상의 내용을 초기에 결정하는 각본의 경우, 장소와 시간에 따라 장면을 나누는 것을 원칙으로 두고 있다. 이와 같은 정보의 추출은 장면의 분석을 통한 movie captioning [1], keyword tagging [2] 등 다양한 분석 기술에서 활용 가능하다.

제안한 방법은 영상 콘텐츠의 장면이 정형화된 구성을 가지는데서 착안하여 설계되었다. 영상 콘텐츠의 각 장면을 구성하는 샷은 하나의 구도를 가져야 한다. 영상 콘텐츠의 제작에서는 그림 1과 같이 와이드샷, 니샷,숄더샷,클로즈업샷 등 사전에 정해진 샷의 구도가 존재한다. 영상 분석을 통해, 해당 영상의 장소 및 시간을 검출할 때, 주요 인물이나 객체를 제외하고, 장소와 시간과 관련된 정보가 존재하는 영역은 샷의 구도에 따라 대체로 결정된다. 그림 1에서 보여주듯이, 와이드샷의 경우 전체 화면의 모든 특징이 시간과 장소 정보를 담고 있을 수 있으나, 숄더샷의 경우 인물이 주로 존재하는 영역을 제외하고 영상의 좌상단 혹은 우상단에 정보가 존재할 가능성이 크다. 따라서 제안한 모델은 입력 프레임에 대해 주요 정보를 추출하기 위한 영역을 특징 추출 과정에서 함께 고려함으로써 검출 성능을 높인다. 이를 위해 VGG16 [3]을 기초로한 모델을 block attention [4]을 적용하여 확장하였다.

동일한 구조의 모델을 장소 및 시간 정보를 태깅한 영화 데이터를 이용하여 성능을 검증하였다. 실제 영화 프레임 81,290장, 92,407장에 대해 각

그림 1 영화 프레임의 샷 예

각 시간 5개 클래스, 장소 189개 클래스를 태깅하였으며, 이중 90%는 학습에 이용하고, 나머지 10%로 성능을 측정하였다. 제안한 모델은 시간에 대해 top-1 정확도 기준 90.96%, 장소에 대해 top-1 정확도 17.83%를 달성하였다. 이는 기본 모델인 VGG16 대비 각각 1.2%, 6.2% 향상된 성능으로 샷의 구도를 고려하여 프레임 영역 및 특징을 선택적으로 활용한 모델의 설계가 효과적이었음을 입증한다.

Ⅱ. 장소 및 시간 정보 검출 모델

본 논문에서는 장면의 프레임 이미지로부터 장소와 시간을 추출할 수 있는 검출 모델을 제안한다. 제안한 모델은 영화, 드라마, 다큐멘터리 등 상업 영상의 특성을 분석하여 설계되었다. 기존의 장소 검출 모델들은 일반적인 이미지를 대상으로 하고 있어, 이들 데이터의 장소 클래스를 상업 영상에 적용하기에 맞지 않다. 이에 본 논문에서는 상업 영상의 각복 150여 편을 분석하여 영화에서 자주 나타나는 장소 189개를 정의하였으며, 동일

한 방법으로 시간 클래스를 정의하였다. 장소의 경우, 교무실, 경찰서, 감옥 등 일반인 영상에서 사용되기 힘든 클래스들이 다수 포함되어 있다. 시간은 낮, 밤, 새벽, 오전, 해질녘으로 나눴다. 새벽과 오전은 영상 처리 측면에서 유사성이 높기에 클래스를 분리하기 어려우나, 영화 제작 관점에서 각본상 두 클래스에서 나타나는 이벤트가 상이하여 나눠서 클래스를 정의하였다.

두 정보를 검출하기 위해서 block attention을 적용하여 VGG16을 확장하였다. VGG15을 백본(backbone) 모델로 활용한 것은, place365 등 장소 검출 데이터셋에서 VGG16이 나쁘지 않은 성능을 보인 점을 고려하였다. 뿐만 아니라, 정보 추출의 궁극적인 목적이 이후 적용할 분석 기술의 입력으로써의 역할인데, 과도하게 큰 모델을 활용할 경우 높은 학습 비용과 추론 시간으로 인해 시스템 구축이 어렵기 때문이다.

제안한 모델의 구조를 좀 더 자세하게 설명하면, VGG16은 5개의 블록으로 구성된다. 제안한 모델에서는 각 블록의 마지막 레이어에 block attention을 추가하였다. block attention은 이전 레이어의 출력을 입력으로 attention을 계산하고 이를 다시 이전 레이어의 출력에 적용하여 attention이 반영된 벡터를 출력한다. 제안한 모델에서는 각 블록의 첫 번째 convolution layer의 출력을 기준으로 attention을 계산하고 이를 마지막 convolution layer의 출력과 결합하여 attention이 반영된 벡터를 얻을 수 있도록 수정하였다. 이는 attention을 도출하는 벡터와 적용하는 벡터를 동일하게 두었을 때, 추가적인 정보를 attention 계산시 활용할 수 없는 점을 고려하여 수정되었다.

Block attention에서는 먼저 각 벡터의 channel에 대한 attention을 먼저 계산하고, 이후 공간 영역에 대한 attention을 계산하였다. 공간 영역에 대한 attention을 도출할 때는 attention layer의 입력으로 channel attention이 적용된 벡터를 입력으로 사용하였다. 이를 통해 공간 상에 attention weight를 계산할 때, 이미 attention에 의해 필터링된 채널의 값이 영향을 미치지 않도록 하였다.

제안한 방법의 검증을 위해 100 여 편의 영화로부터 각각 81,290장, 92,407장의 프레임을 샘플링한 후, 5명이 장소 및 시간을 태깅하여 데이터셋을 구축하였다. 구축된 데이터에서 80% 콘텐츠에 해당하는 프레임을 학습에 활용하였으며, 나머지 프레임 중, 10%씩 콘텐츠 별로 나눠 검증(validation)과 테스트 데이터로 사용하였다. 학습은 최대 2,000번의 반복을 통해 파라미터를 추정하였으며 최종 모델은 검증 데이터에서 손실값이 가장 낮은 모델을 사용하였다.

표1과 표2는 실험 결과 얻어진 top-k 정확도를 보여준다. 실험에서는 기본 VGG16모델, channel attention만 적용한 모델, block attention을 적용한 모델의 성능을 비교하였다. 표 1에서 보듯이 시간의 경우 평균적으로 높은 성능을 보였다. 이는 시간 정보의 클래스가 상대적으로 매우 작고, 낮과 밤과 같이 영상에서 정보를 쉽게 구별 가능하기 때문인 것으로 보인다. 표에서는 기본 모델의 높은 성능을 기반으로 attention을 추가함에 따른 성능 향상을 확인할 수 있다. 표2는 장소 검출 성능으로, 상대적으로 시간에 비해 매우 낮은 수치를 보인다. 하지만 189개의 클래스를 고려할 때 모델이 정보를 추출하는데 부족하지 않음을 알 수 있다. 실험에 참가한 모델들은 대체로 30% 정후의 top-3 정확도를 보여주었다. 제안한 모델은 모든 구간에서 가장 높은 정확도를 보였다. 특히 top-1 정확도에서는 유일하게 20%를 넘겨 block attention이 영화 프레임의 특성을 고려하여 정보를 취사선택하고 있음을 보여준다.

표 1 시간 검출 성능

	Top-1	Top-2	Top-3
VGG16	89.74	97.38	99.29
VGG16 + att.	90.09	97.72	99.04
proposed method	**90.96**	**97.84**	**99.35**

표 2 장소 검출 성능

	Top-1	Top-3	Top-6
VGG16	11.69	21.25	26.34
VGG16 + att.	17.83	30.87	37.75
proposed method	**20.65**	**37.72**	**46.44**

Ⅲ. 결 론

본 논문에서는 영화 영상에서 배경 정보에 해당하는 장소와 시간 정보를 추출하는 검출 모델을 제안하였다. 제안한 모델은 VGG16을 기반으로 block attention을 적용하여 설계되었다. 이를 통해 주요 객체 혹은 사람에 의해 가려지는 영역의 정보를 제외하고 배경에 해당하는 특징을 효과적으로 추출함으로써 검출 성능을 높인다. 실험에서는 실제 영화에서 샘플링한 8만여 개의 프레임을 이용하여 모델을 학습하고 성능을 검증하였다. 검증을 통해 제안한 모델인 시간과 장소에 대한 유의미한 정보를 제공할 수 있음을 보였다.

제안한 모델에서 추출한 정보는 장면 분할, video captioning 등에 적용하여 성능을 높일 수 있을 것으로 기대하고 있다. 따라서 후속 연구로는 추출한 정보를 활용한 상영 영상의 분석 기술 개발을 계획하고 있다.

ACKNOWLEDGMENT

본 연구는 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음.[20ZH1200, 초실감 입체공간 미디어·콘텐츠 원천기술 연구]

참 고 문 헌

[1] A. Rohrback, M. Rohrbach, S. Tang, S. J. Oh and B. Schiele, “Generating Descriptions with Grounded and Co-Referenced People,” In Proceedings of the 30th IEEE CVPR, pp. 4979-4989, 2017.

[2] Y. Zhou, X. Sun, D. Liu, Z. Zha and W. Zeng, “Adaptive Pooling in Multi-Instance Learning for Web Video Annotation,” In Proceedings of the 30th IEEE CVPR, pp. 318-327, 2017.

[3] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” In Proceedings of ICLR, 2015.

[4] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” In Proceedings of ECCV, 2018.