ETRI

2016
Electronics and
Telecommunications
Trends

방송·전파·위성 & 스마트 미디어
기술 특집

스마트 방송서비스를 위한 방송콘텐츠 분석 기술동향

Trends on Broadcasting Content Analysis Techniques for Smart Broadcasting Service

손정우 (J.W. Son) 스마트미디어플랫폼연구실 선임연구원
김선중 (S.J. Kim) 스마트미디어플랫폼연구실 책임연구원

* 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음[B0126-16-1002, 개방형 미디어 생태계 구축을 위한 시맨틱 클러스터 기반 시청상황 적응형 스마트방송 기술 개발].

스마트TV, 스마트폰, 태블릿 컴퓨터 등 다양한 스마트 기기의 이용이 급격하게 확산됨에 따라 방송콘텐츠의 소비패턴 또한 변화하고 있다. 시청자는 더 이상 방송콘텐츠를 시청하기 위해 TV 앞에 앉아 기다리지 않으며, 관심 있는 콘텐츠를 추천받거나 선택하는 방법을 적극적으로 이용하고자 한다. 이러한 소비패턴의 변화는 새로운 방송서비스에 대한 요구사항의 형태로 나타나고 있다. 스마트 방송서비스는 이와 같이 변화된 시청자 소비패턴에 대응하기 위한 새로운 방식의 콘텐츠 전달 서비스로, 스마트 방송서비스 실현을 위해서는 다양한 기술의 개발 및 적용이 요구된다. 본고에서는 스마트 방송서비스를 제공하는 데 필요한 기술 중 방송콘텐츠 분석 기술에 대한 연구동향을 살펴보고, 더불어 한국전자통신연구원에서 개발하고 있는 방송콘텐츠 분석 기술에 대해 소개하고자 한다.

Ⅰ. 서론

방송서비스는 수십 년간 가장 쉽고 간편한 콘텐츠 소비경로를 시청자에게 제공해왔다. 방송사는 정해진 시간에 맞춰 콘텐츠를 송출하고, 시청자가 TV가 설치된 장소에서 전달되는 콘텐츠만을 소비하는 것은 수년 전까지만 해도 매우 일상적이며 당연한 콘텐츠 소비 시나리오였다. 하지만, 최근 이러한 소비패턴이 크게 흔들리고 있다. 스마트폰, 태블릿PC 등의 보급과 고속 인터넷 환경은 시청자들이 특정 장소나 시간에 얽매이지 않고 원하는 콘텐츠를 소비할 수 있도록 했으며, 콘텐츠를 소비할 때에도 다양한 경로를 통해 관련된 정보에 접근할 수 있도록 했다[1]. 이에 따라 시청자의 콘텐츠 소비패턴 변화에 대응하기 위한 새로운 방송서비스의 필요성이 최근 몇 년간 대두되어 왔다.

스마트 방송서비스는 이와 같은 변화에 대응하기 위한 새로운 형태의 방송콘텐츠 전달 서비스이다. 스마트 카(smart car), 스마트 홈(smart home) 등의 분야와 마찬가지로 '스마트' 방송서비스 또한 물리적/논리적 연결성에 기초한 방송서비스를 포함한다. 예컨대, 데스크톱, 스마트폰, 태블릿 PC 등 시청자가 보유한 다수 기기 간의 연결성을 고려한 N스크린 서비스나 촬영 시 사용되는 다수 카메라 간의 연결성을 기반으로 한 다시점 방송서비스 등이 물리적 연결성을 토대로 한 대표적인 스마트 방송서비스다. 반면, 콘텐츠 추천 및 검색 기반의 Video on Demand(VoD) 서비스는 시청자와 콘텐츠, 콘텐츠와 콘텐츠 간의 논리적 연결성에 기초한 스마트 방송서비스로 볼 수 있다.

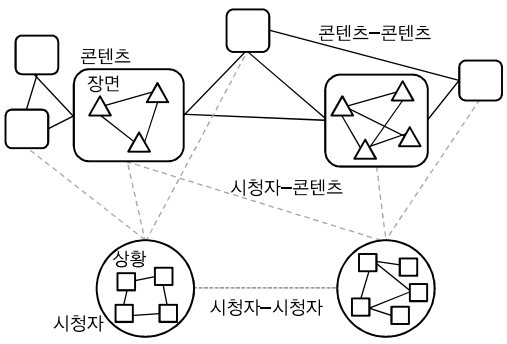
다양한 형태의 스마트 방송서비스 중, 논리적 연결성에 기초한 서비스를 실현하기 위한 필수적인 요구사항은 콘텐츠와 콘텐츠 간, 그리고 콘텐츠와 시청자 간의 연결 관계를 밝히는 기술이다. 가장 널리 쓰여지는 Collaborative Filtering(CF) 기반 추천 기술의 경우, 시청자와 콘텐츠 간의 연결 관계를 시청 이력을 통해 구성하며, 콘텐츠와 콘텐츠 간의 연결 관계는 공유하는 시청자를 통해 구축한다[2]. 최근에는 간단한 이력 정보에서 벗어나, 콘텐츠와 시청자에 대한 다양한 데이터를 기반으로 의미정보를 추출하여 고차원의 연관성을 밝히고 이를 통해 보다 정확하고 많은 연결 관계를 구축하려는 노력이 있다[3]. 이러한 연구가 이루어지는 것은 보다 많은 의미적 연결관계가 다양한 시청자 요구사항에 부합하는 새로운 서비스의 단초가 되기 때문이다. 이들 연구의 핵심은 더 정확하고 가치 있는 의미를 데이터로부터 도출할 수 있는 분석기술에 있다.

본고에서는 스마트 방송서비스 실현을 위한 분석기술, 그중에서도 방송콘텐츠에 대한 분석기술과 동향을 살펴보고 한국전자통신연구원에서 연구 중인 방송콘텐츠 분석기술을 소개하고자 한다.

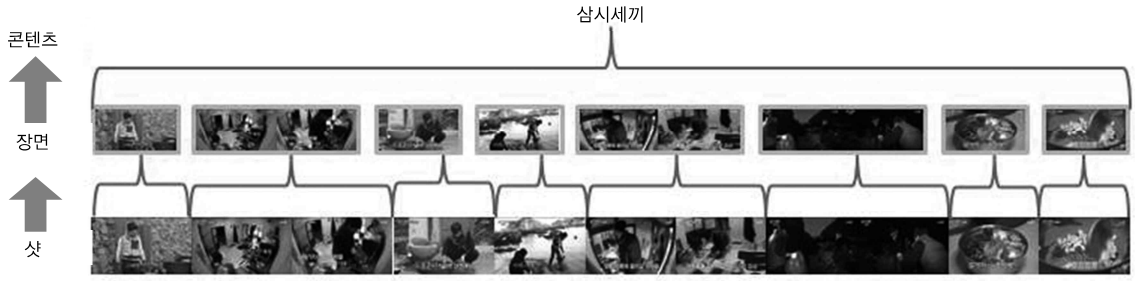
Ⅱ. 스마트 방송서비스 개요

대부분의 스마트 방송서비스는 앞서 간단히 언급한 바와 같이 콘텐츠와 시청자를 분석하여 추출된 고차원 의미정보를 이용하여 구축된 연결 관계를 기반으로 사용자에게 콘텐츠를 제공하는 서비스를 의미한다. (그림 1)은 스마트 방송서비스에서 활용할 수 있는 관계정보를 보여준다.

(그림 1)에서 보여 주듯이 기본적으로 구축되어야 하는 관계는 콘텐츠-콘텐츠 관계, 콘텐츠-시청자 관계,



(그림 1) 스마트 방송서비스에서 활용되는 관계정보



(그림 2) 방송콘텐츠의 구성

그리고 시청자–시청자 관계를 둘 수 있으며, 복합적인 스토리 혹은 의미로 구성되는 콘텐츠를 고려할 경우, 단일 콘텐츠 또한 장면–장면으로 구성된 관계 구조로 해석될 수 있다[4]. 콘텐츠와 마찬가지로 시청자 또한 시간 및 장소 등의 정보를 토대로 상황–상황으로 구성된 관계의 복합체로 볼 수 있다[5].

스마트 방송서비스는 (그림 1)과 같은 관계정보를 바탕으로 시청자에게 적합한 콘텐츠 혹은 관련 정보를 전달한다. 예컨대, 현재 시청 중인 콘텐츠의 의미정보를 시각화하여 부가 정보나 맞춤광고 등을 제공하는 서비스나, 주변 시청자의 콘텐츠 관계정보를 기반으로 다양한 형태의 장면 추천이나 검색 혹은 장면 간 큐레이션 서비스를 생각해볼 수 있다. 앞서 언급했듯이, 이와 같은 관계정보를 구축하는 데 있어 필수적인 기술 중 하나는 콘텐츠에 대한 분석기술이다.

Ⅲ. 방송콘텐츠 분석기술 동향

본 절에서는 방송콘텐츠 분할 기술 중, 샷 및 장면분할, 장면 내 객체인식, 콘텐츠 어노테이션(annotation) 기술에 대해 알아보고 최근 동향을 소개한다.

1. 샷 및 장면분할 기술

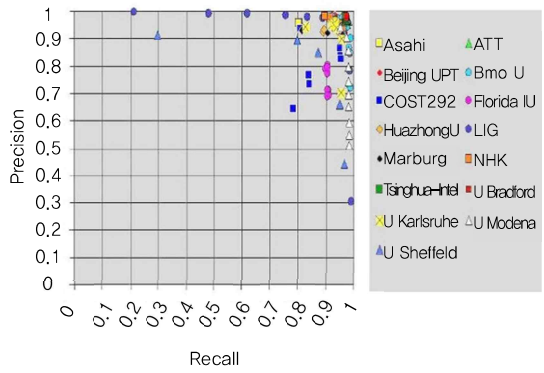
방송콘텐츠의 최소 단위는 한 장의 이미지를 의미하는 프레임이며, 시적으로 연속인 프레임들의 집합을 샷으로, 연속된 샷의 집합을 장면으로 정의할 수 있다[(그림 2) 참조]. 실제적인 방송콘텐츠의 촬영과정을 통해 알아보면, 샷은 단일 카메라에서 촬영된 연속적인 프레임의 집합이다. 반면, 장면은 동일 장소, 동일 시간에 촬영된 샷의 집합으로 볼 수 있다.

〈표 1〉 방송콘텐츠 내 샷의 종류와 설명

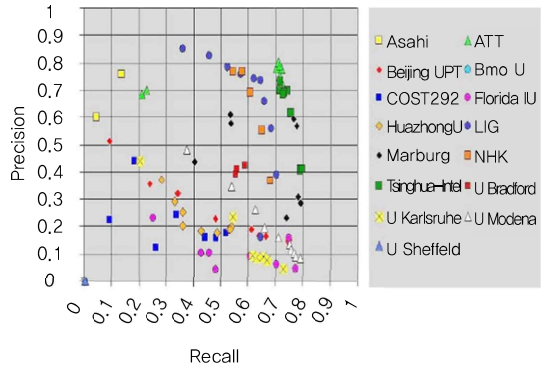
샷의 종류	설명
하드컷	– 샷 사이의 변화가 프레임과 프레임 사이에서 온전히 일어나는 전환 기법
Fade	– 샷의 경계 부분에서 밝기가 어두워지거나 밝아지는 전환 기법
Dissolve	– 두 샷의 시작과 마지막 프레임이 겹쳐지며 샷이 전환되는 기법

샷 및 장면분할 기술은 수동편집 후 생성된 하나의 방송콘텐츠를 그 이전의 상태인 샷 혹은 장면으로 분할하는 방법이다. 일반적인 흐름은, 방송콘텐츠를 입력으로 샷을 분할하고, 분할된 샷을 결합하여 장면을 생성한다.

먼저 샷 분할 기술에 대해 살펴보면, 영상 콘텐츠에 대한 샷 단위 분할은 TRECVID[6]에서 2001년부터 2007년까시 수행된 shot boundary annotation task를 통해서 다양한 방법들이 제안된 바 있다[7]–[9]. TRECVID에서는 샷의 종류를 하드컷(hard cut), 페이드(fade in/out), 디졸브(dissolve)로 구분하여 각각에 대한 분할 성능을 발표하였다. 샷의 종류에 따른 설명은 〈표 1〉에서 확인할 수 있다. (그림 3)은 최종적으로 2007년에 발표된 샷 분할 성능을 보여준다. 그림에서 알 수 있듯이, 하드컷에 대해서는 정확도(precision)와 재현율(recall) 모두 90% 이상을 보였으나, 그 외에는 최대 정확도 약 85%, 최대 재현율 약 80% 정도를 보였다. 이후 샷 추출과 관련된 연구는 거의 없는 편이다.



(a) 하드 컷에 대한 분활성능



(b) 페이즈/디졸브에 대한 분할성능

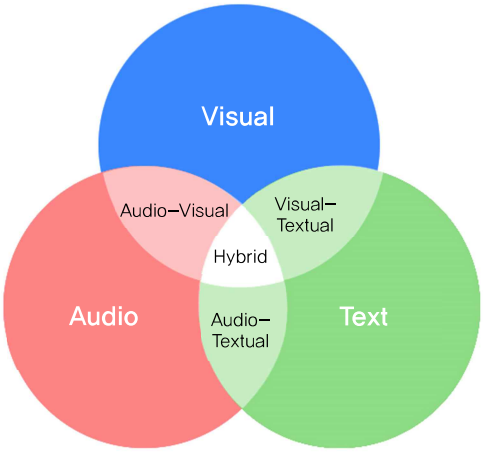
(그림 3) TRECVID 2007에 발표된 샷 분할성능[7]

이미 제안된 대부분의 샷 분할 기술들은 서로 다른 샷 간에 나타나는 색상, 모션 등의 물리적인 차이를 고려하여 인접 프레임 간 벡터 비교를 통해 높은 성능을 얻었다. 따라서 급격한 특징의 차이를 보일 수 없는 페이드나 디졸브의 경우, 여전히 만족할 만한 성능을 얻을 수 없었다. 하지만, 방송콘텐츠에서 페이드와 디졸브 샷의 수가 하드컷에 비해 상대적으로 매우 낮음을 고려하면, 실제 서비스에 적용 가능한 수준의 성능으로 볼 수 있다.

샷 분할 기술에서부터 서비스의 구현까지 과정을 고려해볼 때, 정확도보다는 샷 분할속도 향상을 위한 고속화 기술이나 샷을 대표하는 키프레임에 대한 효과적인 추출기술 등을 추가적인 연구대상으로 볼 수 있다. 특히, 순차적이며 독립적으로 샷의 분할이 이루어지는 기존 기술의 특성상 접근이 용이한 고속화 기술[10]보다 서비스로 표출될 수 있는 샷의 키프레임 추출을 위한 기술이 비교적 최근까지도 연구되고 있는 상황이다[11].

명확한 샷의 정의와 달리, 영상의 장면은 영상의 종류에 따라 다양하게 정의된다[12]-[13]. 방송콘텐츠의 경우, 콘텐츠 제작의 기본이 되는 대본을 바탕으로 정의하면 드라마의 경우 장면은 장소와 시간이 연속적인 샷의 집합으로 정의할 수 있다. 다만 이와 같은 정의에서는 교차 편집된 동일한 스토리의 장면을 하나의 장면으로 포함하기는 어렵다. 서비스에 따라서는 장면을 동일한 스토리를 표현하며 순차적으로 보여지는 샷의 집합으로서 다수의 장소와 시간대를 포함하도록 정의하는 등, 장면 정의는 다양할 수 있다. 일반적으로 장면의 정의가 달라지더라도, 물리적으로 시간상 연속된 샷의 집합으로 정의하는 공통점이 있다. 장면의 정의에서도 알 수 있듯이, 장면은 색상, 모션, 음성 등 특정 특징에 의해 표현되기 힘들며, '스토리'를 고려할 경우, 의미 단위에서의 차이까지 분석해야 장면을 분할 할 수 있다. 따라서 샷 분할에 비해 장면분할에서는 더 다양한 특징을 바탕으로 복합적인 분석기술을 적용함으로써 가능하다.

(그림 4)는 장면분할에서 기본이 되는 샷을 어떤 특징으로 표현하는 가에 따른 장면분할 기술의 분류를 보여준다. (그림 4)에서 알 수 있듯이, 장면분할 기술은 영상, 음성, 텍스트를 이용하거나 이들 셋을 조합하여 이용하고 있다. 예컨대, Closed Circuit Television(CCTV) 영상에서의 장면분할은 영상특징에 크게 의존한다. 반면, 방송콘텐츠의 경우 영상과 음성특징을 사용하는 것이 일반적이며, 방송콘텐츠의 제작 환경을 고려하여 자막, 대본 등의 텍스트 데이터를 이용하기도 한다. 몇몇 경우에는 웹 데이터나 영상에 태깅된 텍스트를 이용하기도 한다. 따라서 대부분의 방송콘텐츠 장면분할 기술은 하이브리드(hybrid) 장면분할 기술로 간주할 수 있다. 초기 콘텐츠에 대한 장면분할 기술은 특징 벡터 간 유사



(그림 4) 사용하는 특징에 따른 장면분할 기술의 분류[14]

도를 바탕으로 규칙을 이용하여 특정 샷에서 장면의 경계를 검출하였다[15]–[16]. 규칙을 정의하여 사용할 수 있었던 것은 콘텐츠를 촬영하는 기법이 존재하기 때문이며, film grammar라 불리는 몇몇 기법을 바탕으로 규칙이 정의되었다. 이후에는 Hidden Markov Model(HMM)과 같은 시퀀스 데이터 처리 모델을 이용하여 장면을 분할하였다. 이러한 방법에서는 하나의 장면을 구성하는 샷의 종류를 밝히고 이들 간의 전환 확률을 학습함으로써 장면을 분할하였다.

최근에 개발되고 있는 장면분할 기술은 클러스터링 방법에 기반을 두고 있다. 시퀀스 모델을 사용할 경우 이전 혹은 이후 샷 간의 관계를 반영하는데 한계가 있다. 많은 시퀀스 정보는 높은 계산량을 요구할뿐만 아니라, '주변' 샷을 어디까지 정의할 것인가가 큰 문제였다. 클러스터링 방법을 활용한 장면분할 기술에서는 전체 샷 간의 유사도를 바탕으로 샷의 군집을 밝힘으로써 콘텐츠 전반에 걸친 정보를 활용할 수 있어, 상대적으로 높은 성능을 보였다. 최근에는 스펙트럴 클러스터링(spectral clustering)과 같은 그래프 기반의 분할기법들이 활용되고 있다[17]. 그래프 기반의 분할기법은 장면분할문제를 샷 간의 유사도를 기반으로 구성된 샷 그래프를 특정 수의 서브그래프(장면)로 분할하는 것으로 간주하였다. 이러한 문제 정의의 장점은 일반적인 클러스터링 기술에서 가정하는 convex 군집을 가정할 필요가 없다는 것이다. 장면을 convex 군집으로 가정할 경우, 장면 내 샷들은 모두 서로 간 높은 유사도를 가져야 한다. 콘텐츠에서의 스토리 흐름을 고려할 경우, 위의 가정은 특정 장면에서만 유효하다. 그래프 기반의 클러스터링을 이용하여 장면을 분할할 경우, non-convex 군집까지 분할이 가능하기 때문에 다양한 형태의 장면을 분할할 수 있다.

규칙 기반 방법에서 시퀀스 모델, 클러스터링으로 이어져 온 장면분할 기술의 흐름은 최근 다시 한 번 전환기를 맞고 있다. 딥러닝(deep learning) 기술의 일종인 Long Short Term Memory(LSTM) 모델[18]은 데이터에서 이전 시퀀스의 함축된 의미정보를 선택적으로 활용하여 다양한 문제에서 높은 성능을 보이고 있다. Bi-directional LSTM(BLSTM)[19]은 시퀀스 데이터에서 특정 시점을 기준으로 이전과 이후 의미정보를 모두 활용할 수 있어 콘텐츠의 장면분할에 활용할 수 있다. 이러한 모델을 활용한 장면분할 기술이 이후 제안될 것으로 예상된다.

다만 이들 모델은 지도학습 기반의 학습 모델이기 때문에 대용량의 학습 데이터 구축이 선행되어야 한다. 장르별, 콘텐츠별 장면의 경계를 결정하는 특성은 매우 상이하다. 따라서 이 경우, 모델의 학습 기술보다는 모델 학습을 위해 적절한 레이블 데이터를 구축하는 것이 어려운 문제가 된다. 기존의 장면분할문제와 마찬가지로 비지도 학습환경에서의 모델 구축을 고려할 경우, LSTM이나 BLSTM 등 딥러닝 기반 시퀀스 모델은 복합적인 샷의 의미 벡터를 추출하는 데 사용하고 이후 클러스터링과 같은 결합방법을 적용해볼 수 있다. 이 경우에는 레이블 데이터 구축이 필요 없어지나, 딥러닝이 가지는 오류 전이(error propagation)를 통한 모델 학습의 장점은 반감한다.

한국전자통신연구원에서는 방송콘텐츠에 대한 스펙트럴 클러스터링 기반의 장면분할 기술을 개발하였다. 한국전자통신연구원의 장면분할 기술은 콘텐츠에서 분할된 샷에 대해 10여종 이상의 영상, 오디오, 텍스트 특징을 추출하고 이들 간의 상관관계를 기반으로 특징 간 정보를 전달하는 협업 학습 기법을 적용한 것이 특징이다. 개발한 기술에서는 특징 간 정보전달을 통해 샷을 표현함에 있어 정보손실을 최소화하여 성능을 높였다. 개발 기술의 검증은 총 10편의 드라마 콘텐츠를 대상으로 했으며, 평균 81.3%의 Adjusted Rand Index(ARI)를 보였으며, 이는 기존 스펙트럴 클러스터링 기반 기술[20] 대비 최소 1.57 % 최대 12.04% 향상된 결과이다. 장면분할 기술을 바탕으로 한국전자통신연구원은 방송콘텐츠를 구성하는 샷-장면-콘텐츠의 트리 구조를 벗어나, 장면 간 수평적인 복합 연결 관계를 고려한 비선형의 장면 집합과 하나의 스토리 단위를 이루는 시맨틱 클러스터를 구성하는 기술을 개발하고 있다.

2. 장면 내 객체인식

객체인식은 콘텐츠 내에 등장하는 다양한 인물, 제품, 장소 등의 객체를 인식하는 기술이다. 객체인식은 크게 카테고리를 인식하는 객체검출과 인물의 이름, 제품명 등 상세한 정보를 인식하는 객체식별로 나눌 수 있다. 객체인식은 일반적으로 영상처리 기술을 기반으로 이루어지며, 본고에서 소개하는 다양한 기술 중 최근 가장 크게 발전하고 있는 분야이다. 객체인식의 최근 경향에서 가장 주목해야 할 점은 앞서 언급한 바 있는 딥러닝이다.

영상처리 분야에서는 객체인식을 위한 다양한 데이터를 공유하고 있다. 가장 큰 데이터 중 하나로 ImageNet을 들 수 있다. ImageNet 데이터를 활용한 Imagenet Large Scale Visual Recognition Challenge(ILSVRC)[21]에서는 다양한 문제에 대한 세계각국의 연구를 매년 비교하여 발표하고 있다. 이를 위해 ImageNet은 200여 카테고리에 대한 태깅 데이터를 제공하고 있다. 최근 ILSVRC에서는 이미지 내의 객체인식을 위하여 딥러닝을 이용한 다양한 방법들이 제안되고 있으며, 특히 Convolution Neural Network(CNN)를 기반으로 한 기술들은 기존의 SVMs와 같은 기술들의 성능을 크게 뛰어 넘고 있다[22].

〈표 2〉 ILSVRC 2015의 문제 및 설명

문제	설명
Object Detection	– 이미지가 의미하는 객체의 카테고리 검출 (200 카테고리)
Object Localization	– 이미지 내에서 객체 및 객체의 위치를 검출 (150,000 이미지, 1,000 카테고리)
Object Detection from Video	– 200 카테고리 객체에 대한 비디오 데이터에서의 객체검출
Scenec Cassification	– 이미지의 장면 분류(100만장 이상의 이미지, 400 이상의 장면 카테고리)

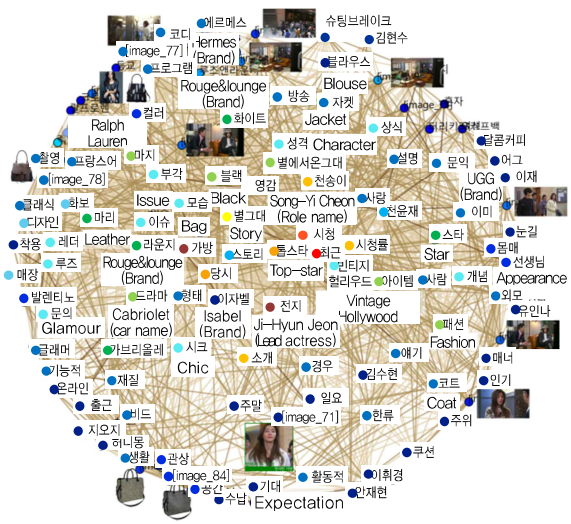
ILSVRC에서 고려하는 문제는 〈표 2〉에서 알 수 있듯이, 총 4종으로, 객체검출, 지역화, 장면 분류 등이 포함되어 있다. 객체검출의 경우 2015년을 기점으로 비디오 데이터에서의 검출문제가 추가되어 콘텐츠 내 객체검출과 직접적으로 관련된 다양한 기술들이 발표되고 있지만, 객체식별은 수행하지 않았다.

최근 영상처리 분야는 딥러닝을 기반으로 많은 문제들을 해결해 나가고 있다. 이와 같은 상황에도 불구하고 콘텐츠의 의미 분석을 위한 장면 내 객체인식 기술 개발에는 여전히 많은 어려움이 있다. 객체검출의 경우 많은 카테고리에서 이미 높은 성능을 내고 있지만, 실제 부가서비스를 제공하고자 하거나, 콘텐츠 간 연결관계를 구축하기 위해서는 좀 더 자세한 인식 정보를 제공할 수 있는 객체식별 기술이 매우 중요하다.

방송콘텐츠에 대한 객체식별이 어려운 이유는 대상 객체가 매우 다양한데 일차적인 원인이 있다. 일반적인 드라마 콘텐츠의 경우 3~5명의 주연급 캐릭터와 20여

명의 조연급 캐릭터가 존재한다. 이 경우, 23~5명의 인물에 대한 식별이 가능해야 하며, 이를 위한 식별기 학습 및 학습 데이터 구축이 선행되어야 한다. 출연 인물에 대한 학습 데이터를 구축하는 것은 이전에 촬영된 콘텐츠를 태깅하여 구축하거나 웹에서 수집을 통해 이루어질 수 있다. 하지만, 방송콘텐츠의 특성상 동일인물이라 하더라도, 콘텐츠에 따라 완전히 다르게 표출되기 때문에 만족스러운 크기의 데이터를 구축하기 힘들다. 인물 객체의 경우, 한정된 크기이기에 시도해 볼만 한 여지가 있으나, 휴대폰, 자동차 등의 제품 객체의 경우 그 수가 많고 데이터가 한정되어 있어 식별모델을 학습하는 데 한계가 있다.

위와 같은 이유로 객체검출과 달리 현재까지 객체식별 문제를 해결하기 위한 기술 개발은 초기 개발단계로 볼 수 있다. 이를 해결하기 위해 외부 지식을 활용하는 기술들이 일부 제안된 바 있다[23]-[24]. Tapaswi et al.[23]의 경우 시트콤에 등장하는 인물들과 관련된 자주 입는 의상의 색상, 출연장소의 배경정보, 인물 간 관계정보 등을 바탕으로 방송콘텐츠에서의 인물을 식별하는 기술을 제안한 바있다.



(그림 5) 드라마 '별에서 온 그대'에 대해 자동구축된 지식망

한국전자통신연구원에서는 웹에서 수집한 드라마 관련 블로그 포스팅을 바탕으로 구축된 콘텐츠 지식망[(그림 5) 참조]과 출연진 정보, Product placement(PPL) 정보를 바탕으로 객체를 식별하는 기술을 개발하였다[24].

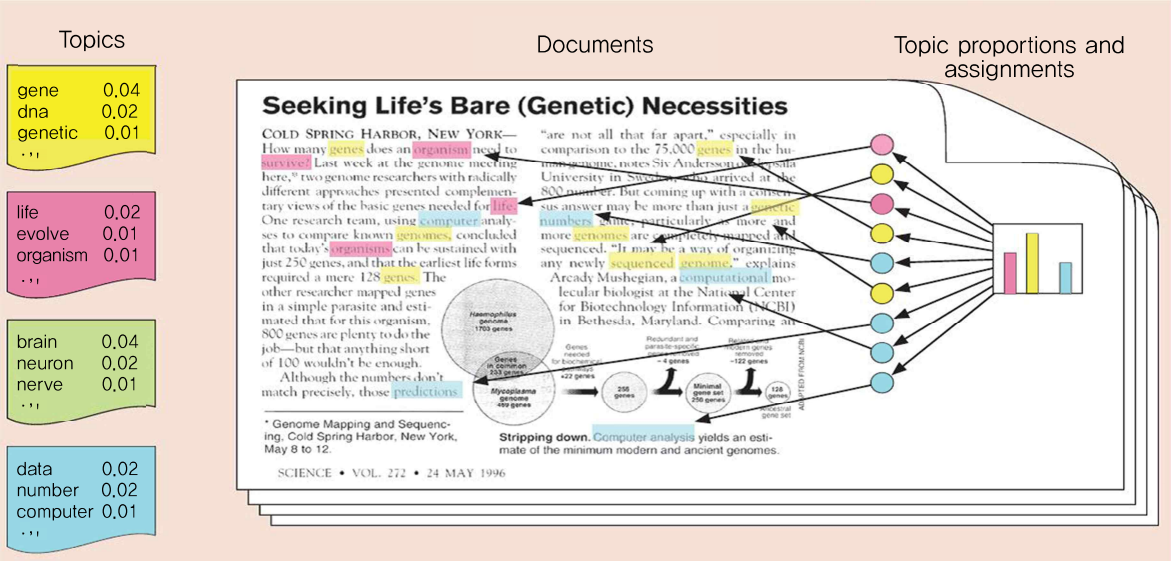
3. 콘텐츠 어노테이션

콘텐츠 어노테이션이란, 방송콘텐츠와 관련된 다양한 정보를 콘텐츠에 태깅(tagging)하는 기술을 의미한다. 넓은 의미에서는, 객체인식 결과를 콘텐츠에 태깅하는 기술이나, 장면의 경계정보를 태깅하는 기술까지도 콘텐츠 어노테이션이라 할 수 있다. 본 장에서는 객체나 분할 정보를 제외한 스토리, 의미, 시청자 감정 등의 정보를 분석하고 그 결과를 콘텐츠 전체 혹은 일부분에 태깅하는 기술에 대해서 다루고자 한다.

콘텐츠 어노테이션에서의 정보 추출은 방송콘텐츠 외부 데이터를 이용하여 이루어진다. 일례로, Masuda et al.[25]은 온라인 비디오 콘텐츠에 태깅된 사용자의 태그나 짧은 글 등을 수집하고 이 중 주요 단어들을 추출하여 콘텐츠에 어노테이션 한 후, 이를 통해 비디오 콘텐츠에 대한 검색 서비스를 제안한 바 있다.

콘텐츠 어노테이션은 크게 온톨로지(ontology)나 사전과 같은 사전에 준비된 지식 체계를 이용하는 경우[26]와 데이터에서 자동학습된 의미정보를 이용하는 경우로 나눌 수 있다. 온톨로지나 사전을 이용하는 경우에는 인물 간 관계, 극의 전개 과정, 갈등의 종류 등을 사전에 정의하고 주어진 콘텐츠 전체 혹은 일부가 기 정의된 온톨로지 등의 어느 부분에 해당하는 지를 분석한다. 이 경우, 사전에 정의된 지식을 활용하기 때문에 태깅되는 정보를 쉽게 제어할 수 있는 점과 태깅하는 정보의 수가 명확하고 한정되기 때문에 비교적 높은 정확률(precision)을 가지는 태거를 구축할 수 있는 장점이 있다.

Francois et al.[27]은 콘텐츠 어노테이션을 위한 객체와 이벤트 온톨로지를 구축하고 이를 기반으로 CCTV



(그림 6) 토픽 모델의 개념도[29]

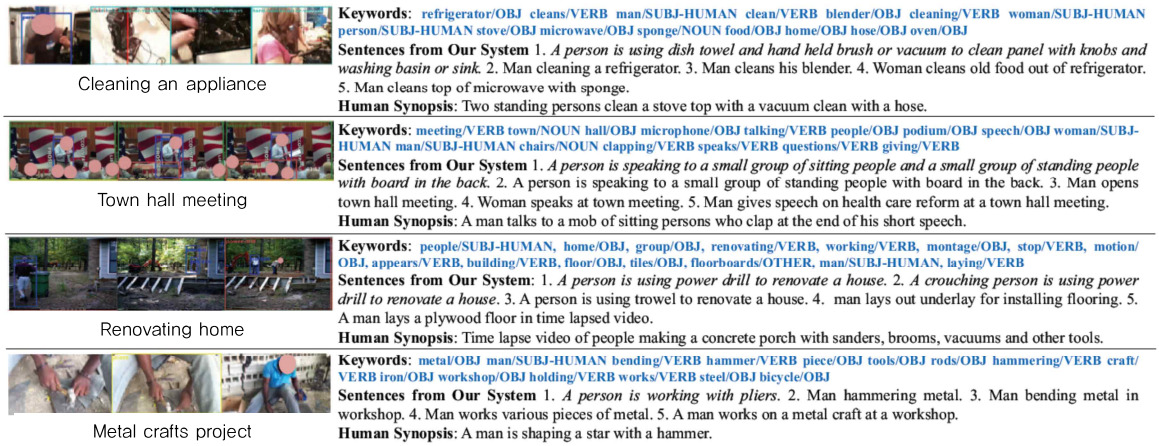
콘텐츠를 태깅한 바 있다. 기존 연구에서도 알 수 있듯이 사전에 정의된 지식 체계를 기반으로 한 어노테이션의 경우, 특정 장르나 환경을 염두에 두고 이루어 진다. 이는 사전에 구축할 수 있는 지식 체계가 한정되기 때문에 어노테이션 또한 다양한 장르나 환경에 적용할 수 없는 단점 때문이다.

지식 체계를 활용한 어노테이션의 단점을 보완하기 위한 시도로 데이터에서 자동학습된 의미정보를 이용한 콘텐츠 어노테이션 기술들이 제안된 바 있다. 대표적으로 토픽모델에 기반한 방법들을 들 수 있다[28]. 토픽모델이란, 문서를 이루는 토픽 집합을 문서로부터 자동학습하는 기술이다[(그림 6) 참조]. (그림 6)에서 보여주듯이 토픽모델은 모든 데이터가 토픽으로부터 생성된 단어로 이루어져 있음을 가정하고 문서로부터 문서를 구성하는 토픽을 추정한다. 학습 과정은 먼저, 각 문서를 구성하는 토픽의 밀도 함수와 토픽 내 단어의 밀도 함수 형태를 가정한 다음, 문서로부터 가정된 밀도 함수의 파라미터를 추정하는 순서로 이루어진다.

Das et al.[30]은 토픽모델을 이용하여 비디오 콘텐츠에 대한 키워드를 생성하고 이를 통해 장면의 의미나 더 나아가 장면에 대한 문장 단위의 표현을 자동 생성하는 기술을 제안한 바 있다[(그림 7) 참조]. 지식 체계를 이용한 이전 기술들과 달리, 토픽모델은 주어진 데이터를 기반으로 자동학습되기에 사전에 지식 체계를 구축하기 위한 비용이 들지 않으며, 적용 대상에 대한 제한이 적은 편이다.

다시 말해 토픽모델의 장점은 임의의 토픽에 대해 자동학습되기 때문에 사전의 지식체계 구축이 필요치 않으며, 적용 대상에 따라 다양한 형태의 새로운 토픽을 자동으로 학습한다는 데 있다. 다만, 토픽이 단어와 단어의 확률값으로 이루어진 집합으로 표현되기 때문에 가독성이 낮은 점과 이론적으로는 잘 정의되어 있으나 이론적 기반이 약한 초기 사용자의 경우 파라미터 설정 등에 어려움을 겪을 수 있는 단점이 있다.

한국전자통신연구원에서는 장면 단위로 분할된 방송콘텐츠에 대해, 자막 및 대본을 기반으로 토픽을 생성하는 Guided Hierarchical Dirichlet Process(GHDP) 모델을 개발하였다. GHDP에서는 초기 방송콘텐츠에 대한 토픽 학습에서 가장 큰 문제가 되는 적은 양의 짧은 데이터(대사, 지문 등)로부터의 토픽학습을 위해 Point-wise

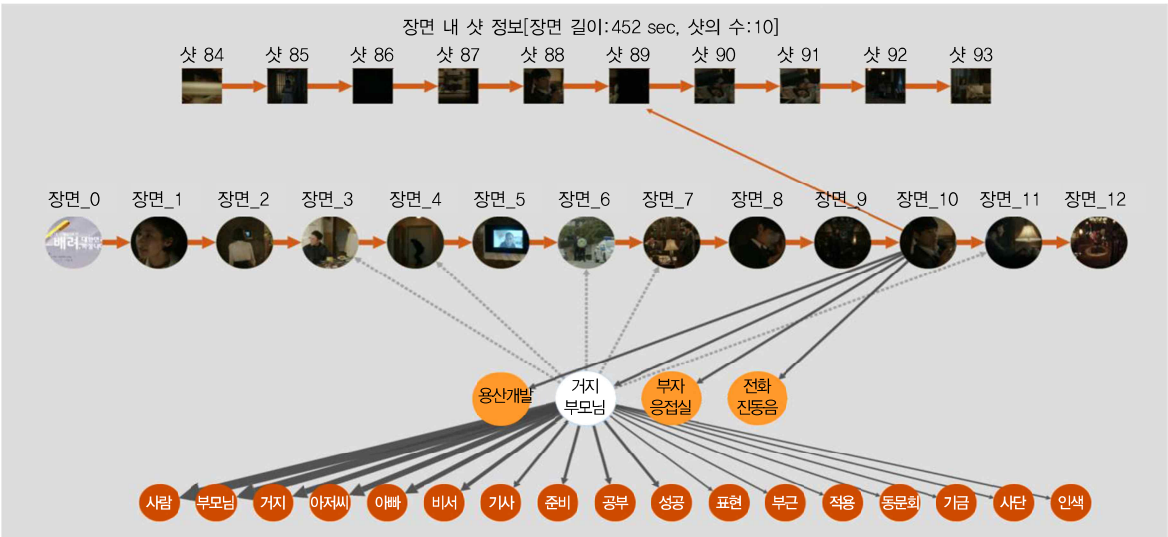


(그림 7) 토픽기반의 비디오 어노테이션 예[30]

Mutual Information(PMI) 기반의 시드(seed)단어 추출 기술을 적용하였다. 시드단어는 생성하고자 하는 토픽의 방향성을 초기에 모델에 제공함으로써 적은 양의 데이터에서도 토픽이 학습될 수 있도록 한다. 한편의 방송콘텐츠에 속한 다수의 장면은 GHDP를 통해 학습된 토픽을 공유함으로써 장면 간 연결성을 가지게 된다. (그림 8)은 GHDP를 기반으로 생성된 장면 간 연결성을 시각화하는 인터페이스를 보여준다. 그림에서 알 수 있듯이, 방송콘텐츠를 이루는 장면은 시간순서에 관계 없이 유사한 의미를 가질 경우, 토픽을 통해 연결된다.

Ⅳ. 결론

본고에서는 스마트 방송서비스를 위한 방송콘텐츠 분석기술의 동향을 소개하였다. 전통적인 방송시장의 규모가 축소되고 있는 상황에서도 방송콘텐츠가 가지는 가치는 타 콘텐츠와 비교했을 때 매우 높게 평가되고 있다. 높은 부가가치를 가지는 방송콘텐츠를 활용한 서비스 발굴은 기존 방송시장의 부흥뿐만 아니라 다양한 시장의 생성이라는 측면에서 매우 중요하다. 방송콘텐츠를 이용한 서비스의 일환으로 스마트 방송서비스는 고부가가치를 가지는 방송콘텐츠에 대한 다각화된 비즈니



(그림 8) GHDP 기반 토픽 생성 예시

스 모델을 제공하기 위한 시도이며, 기존과 달리 콘텐츠에 대한 다양한 지식정보를 기반으로 서비스를 제공한다.

스마트 방송서비스의 성공을 위해서는 콘텐츠 분석을 통해 콘텐츠를 아우르는 정확하고 풍부한 지식정보를 획득하는 기술이 필요하다. 본고에서는 다양한 방송콘텐츠 분석기술 중, 샷 및 장면분할, 객체인식, 콘텐츠 어노테이션 기술의 개발동향을 소개하였다. 최근 딥러닝을 기반으로 촉발된 인공지능 기술의 도약은 방송콘텐츠 분석기술에도 큰 영향을 미칠 것으로 예상된다. 따라서 앞으로 더 다양하고 정확한 의미정보를 추출할 수 있을 것으로 기대되기에, 새로운 기술을 바탕으로 한 지속적인 방송콘텐츠 분석기술 개발이 요구된다.

용어해설

딥러닝 기계 학습 기술의 일종으로 입력-은닉-출력의 층을 가지는 선형 모델의 계층적 구조로 정의되는 모델의 학습 기술

토픽 모델 생성 모델을 이용하는 자연어 처리 기술의 일종으로 문서 집합을 구성하는 토픽을 학습하고, 입력된 문서에 대한 토픽과 그 확률을 추정하는 모델. 토픽은 단어와 해당 단어의 확률값 집합으로 정의됨.

시맨틱 클러스터 콘텐츠를 구성하는 장면의 비선형 집합으로 시간적으로 연속적이진 않으나, 의미상 연속되는 장면의 집합을 의미

약어 정리

ARI	Adjusted rand index
BLSTM	Bi-directional LSTM
CCTV	Closed Circuit Television
CF	Collaborative Filtering
CNN	Convolution neural network
GHDP	Guided hierarchical dirichlet process
HMM	Hidden markov model
ILSVRC	Imagenet Large Scale Visual Recognition Challenge
LSTM	Long short term memory
PMI	Point-wise mutual information
PPL	Product placement
VoD	Video on demand

참고문헌

[1] 정보통신정책연구원, "2014년 방송매체 이용행태 조사 보고서," 2014. 12.

[2] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, 1998, pp. 43-52.

[3] D. park, H. Kim, I. Choi, and J. Kim, "A Literature Review and Classification of Recommender Systems Research," *Expert Systems with Applications*, vol. 38, no. 11, 2012, pp. 10059-10072.

[4] M. Fabro and L. Böszörmenyi, "State-of-the-art and Future Challenges in Video Scene Detection: A Survey," *Multimedia Systems*, vol. 19, no. 5, 2013, pp. 427-454.

[5] B. Clarkson, A. Pentland, and K. Mase, "Recognizing User Context via Wearable Sensors," *Proc. IEEE Inter. Symp. Wearable Computers*, Oct. 2000, p. 69.

[6] NIST, "TREC Video Retrieval Evaluation: TRECVID," http://trecvid.nist.gov/

[7] A. Smeaton, P. Over, and A. Doherty, "Video Shot Boundary Detection: Seven Years of TRECVid Activity," *Computer Vision Image Understanding*, vol. 114, no. 4, 2010, pp. 411-418.

[8] J. Yuan et al., "A Formal Study on Shot Boundary Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no.2, Feb. 2007, pp. 168-186.

[9] P. Over et al., "TRECVID 2007-Overview," July 31st, 2014, pp. 1-27, http://www-nlpir.nist.gov/projects/tvpubs/tv7. papers/tv 7overview.pdf

[10] E. Apostolidis and V. Mezaris, "Fast Shot Segmentation Combining Global and Local Visual Descriptors," *Proc. IEEE Inter. Conf. Acoustic, Speech and Signal Processing*, 2014, pp. 6583-6587.

[11] R. Hannane et al., "An Efficient Method for Video Shot Boundary Detection and Keyframe Extraction using SIFT-point Distribution Histogram," *Inter. J. Multimedia Information Retrieval*, Mar. 16th, 2016, pp. 1-16.

[12] J. Monaco, "How to Read a Film: The World of Movies, Media, Multimedia: Language, History, Theory," Oxford University Press, 2000.

[13] E. Katz, F. Klein, and R.D. Nolen, "The Film Encyclopedia," Harperperennial, 1998.

[14] M. Fabro and L. Boszormenyu, "State-of-the-Art and Future Challenges in Video Scene Detection: a Survey,"

Multidedia Systems, vol. 19, no. 5, 2013, pp. 427-454.

[15] J. Huang, Z. Liu, and W. Yao, "Integration of Audio and Visual Information for Content-based Video Segmentation," *Proc. Inter. Conf. Image Processing*, vol. 3, 1998, pp. 526-529.

[16] J.R. Kender and B.-L. Yeo, "Video Scene Segmentation via Continuous Video Coherence," *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, June 1998, pp. 367-373.

[17] V.T. Chasanis, A.C. Likas, and N.P. Galatsanos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment," *IEEE Transactions on Multimedia*, vol. 11, no. 1, 2009, pp. 89-100.

[18] F. Gers, J. Schmidhuber, and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," *Neural Computation*, vol. 12, no. 10, 2000, pp. 2451-2471.

[19] A. Graves, J. Schmidhuber, "Framewise Phoneme Classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, 2005, pp. 602-610.

[20] A. Kumar and H. Daume III, "A Co-training Approach for Multi-view Spectral Clustering," *Proc. Inter. Conf. Machine Learning*, 2011.

[21] ImageNet Large Scale Visual Recognition Challenge, http://www.image-net.org/challenges/LSVRC/

[22] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems 25, 2012, pp. 1097-1105.

[23] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "Knock! Knock! Who is it? Probabilistic Person Identification in TV-Series," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2012, pp. 2658-2665.

[24] J.W. Son, A. Lee, and S.J. Kim, "Knowledge Construction for the Broadcasting Content by Using Audience Oriented Data," *Proc. IEEE/WIC/ACM Inter. Conf. Web Intelligence and Intelligent Agent Technology*, 2015, pp. 89-92.

[25] T. Masuda et al., "Video Scene Retrieval Using Online Video Annotation," *New Fromtiers in Aritificial Intelligence*, vol. 4914, 2008, pp. 54-62.

[26] S.A. Bhat et al., "Overview of Existing Content Based Video Retrieval Systems," *Inter. J. Advanced Engineering and Global Technology*, vol. 2, no. 2, 2014, pp. 476-483.

[27] A. Francois et al., "VERL: An Ontology Framework for Representing and Annotating Video Events," *IEEE MultiMedia*, vol. 12, no. 4, Oct.-Dec. 2005, pp. 76-86.

[28] V. Ramanathan, P. Liang, and L. Fei, "Video Event Understanding Using Natural Language Descriptions," *Proc. IEEE Inter. Conf. Computer Vision*, Dec. 2013, pp. 905-912.

[29] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, 2003, pp. 993-1022.

[30] P. Das et al., "A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 2634-2641.