비디오 시각적 관계 이해 기술 동향

Trends in Video Visual Relationship Understanding

권용진 (Y.J. Kwon, scocso@etri.re.kr) 시각지능연구실 선임연구원
김대희 (D.H. Kim, dhkim19@etri.re.kr) 시각지능연구실 선임연구원
김종희 (J.H. Kim, jhkim27@etri.re.kr) 시각지능연구실 선임연구원
오성찬 (S.C. Oh, sungchan.oh@etri.re.kr) 시각지능연구실 선임연구원
함제석 (J.S. Ham, jsham@etri.re.kr) 시각지능연구실 연구원
문진영 (J.Y. Moon, jymoon@etri.re.kr) 시각지능연구실 책임연구원

ABSTRACT

Visual relationship understanding in computer vision allows to recognize meaningful relationships between objects in a scene. This technology enables the extraction of representative information within visual content. We discuss the technology of visual relationship understanding, specifically focusing on videos. We first introduce visual relationship understanding concepts in videos and then explore the latest existing techniques. Next, we present benchmark datasets commonly used in video visual relationship understanding. Finally, we discuss future research directions in video visual relationship understanding.

KEYWORDS 관계 트리플렛, 시각적 관계 이해, 시공간 정보 탐색

Ⅰ. 서론

시각지능 기술은 영상을 기계가 사람처럼 인식하고 이해하는 것이 궁극적인 목표이다. 최근 인공지능 기술의 급격한 발전으로 이러한 목표에 한층 더 가까워진 것으로 생각되지만, 여전히 영상을 사람처럼 매끄럽게 이해하는 것은 어려운 일이다. 특히, 단순히 영상에서 나타나는 사람(Person)이나 사물(Object)을 찾는 단계를 넘어, 영상에서 발생하는 사건이나 상황을 파악하여 이를 상세히 설명하거나, 상황을 면밀히 분석하여 의사결정 과정에 반영하거나, 더 나아가 향후 발생할 일이 무엇인지 사전에 예측하는 등 영상으로부터 고수준의 정보를 얻어 실세계 문제에 활용하는 기술은 극히 한정된 분야를 제외하고는 아직까지 발전이 요원한 상태이다.

영상을 높은 수준에서 포괄적으로 이해하기 위해서는 영상에서 의미 있는 정보가 무엇인지 알 수 있어야 한다. 이를 위해서는 영상에 등장하는 사

* DOI: https://doi.org/10.22648/ETRI.2023.J.380602
* 이 논문은 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No. 2020-0-00004, 장기 시각 메모리 네트워크 기반의 예지형 시각지능 핵심기술 개발].

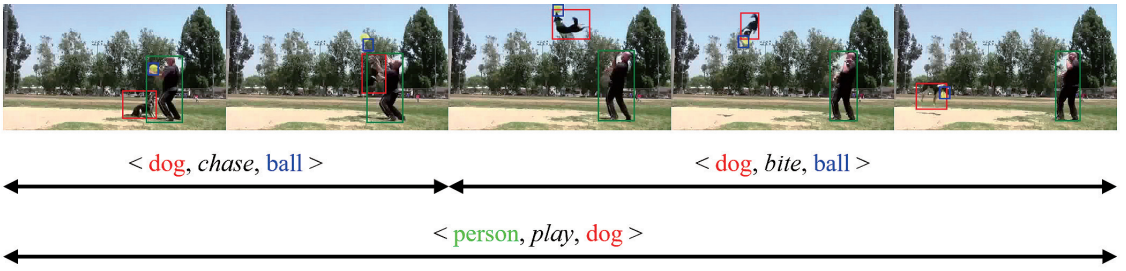
람/사물 인식을 넘어, 인식된 사람/사물 간의 관계나 상호작용을 인식하거나 행동/이벤트의 주체와 대상을 파악하는 등 의미 있는 시각적 관계 정보를 인식할 수 있어야 한다[1-3]. 본고에서는 이처럼 영상에서 등장하는 사람/사물 간 의미 있는 관계 정보를 인식하는 기술을 시각적 관계 이해(Visual Relationship Understanding) 기술로 지칭한다. 시각적 관계 이해 기술을 이용하면 영상에서 단순한 인식 정보를 넘어, 영상에 포함된 의미 있는 정보들을 파악할 수 있기 때문에 시각적 관계 이해 기술은 행동 인식(Action Recognition)[4], 영상 설명 생성(Image/Video Captioning)[5-7], 시각적 질의응답(Visual Question Answering)[8-10], 의미 기반 영상 검색(Semantic Image/Video Retrieval)[1,11,12] 등 고수준의 복잡한 시각지능 응용에서 중요한 기반 기술로 고려되고 있다.

이처럼 시각적 관계를 인식하기 위해서는 먼저 영상에서 등장하는 객체를 찾고, 두 객체씩 하나의 쌍으로 조합하여 두 객체가 어떤 관계를 갖는지 인식해야 한다. 문제는 두 객체의 조합이 굉장히 다양하게 나올 수 있으며, 동일한 객체 종류의 쌍이라 할지라도 영상에서 나타난 형상(Appearance), 위치(Location), 주변 상황(Context) 등에 따라 관계가 서로 다르게 나타날 수 있다는 점이다. 또한, 두 객체 간 관계가 복수 개가 나타날 수 있다는 것과 당연한 (Trivial) 관계 정보를 넘어 희소하면서도 중요한 의미를 담는 관계 정보도 찾아야 하는 것도 중요한 문제이다. 특히, 지금까지는 단일 이미지를 대상으로 한 시각적 관계 이해 연구[13]가 집중적으로 수행되었지만, 현업에서 필요로 하는 시각지능 응용은 비디오를 대상으로 하는 경우가 많이 있기 때문에 비디오 시각적 관계 이해 기술의 중요성은 더욱 크다고 볼 수 있다.

본고에서는 비디오 시각적 관계 이해를 위한 최신 기술 동향에 대해 알아본다. 먼저 Ⅱ장에서는 비디오 시각적 관계 이해 기술의 정의와 특징을 살펴본다. Ⅲ장에서는 비디오 시각적 관계 이해 기술의 일반적인 접근 방법이 무엇인지 살펴보고, 각 접근 방법에 따른 최신 기술 동향에 대해 알아본다. Ⅳ장에서는 비디오 시각적 관계 이해 연구를 수행함에 있어 모델 학습 및 성능 검증에 필요한 벤치마크 데이터셋을 살펴본다. 마지막으로 Ⅴ장에서는 비디오 시각적 관계 이해 기술 동향을 정리하고 향후 전망을 소개한다.

Ⅱ. 비디오 시각적 관계 이해 기술 개요

비디오 시각적 관계 이해 기술은 비디오 안에 나타나는 시각적 관계 정보를 찾는 것으로, 여기서 시



출처 Reproduced from [14].

그림 1 비디오 시각적 관계 이해 예시

각적 관계는 그래프의 형식으로 표현한다. 그래프의 노드(Node)는 비디오에 등장하는 객체(Instance)와 두 객체의 시각적 관계(Visual Relationship)를 나타내고, 그래프의 에지(Edge)는 해당 시각적 관계가 어떤 객체와 연관되었는지 나타낸다. 일반적으로 각각의 시각적 관계는 관계 트리플렛(Relation Triplet) 〈subject, predicate, object〉의 형식으로 표시하기도 한다. 그림 1은 비디오 시각적 관계 이해 기술의 예시 그림이다.

비디오 시각적 관계 이해 기술은 언뜻 보기에 이미지에서의 시각적 관계 이해 기술[13]을 확장함으로써 해결할 수 있을 것처럼 보인다. 하지만 이와 같은 접근 방법은 비디오에서만 나타나는 시각적 관계의 특징을 간과할 수 있어 해결 방안이 될 수 없다[15]. 첫 번째로, 이미지와 달리 비디오는 행동(Action)을 포함하거나 동적인 의미가 있는 시각적 관계를 담을 수 있다[16]. 예를 들어, 〈dog, run past, person〉이나 〈dog, chase, frisbee〉 등과 같은 시각적 관계는 비디오에서만 얻을 수 있다. 이러한 시각적 관계를 찾기 위해서는 비디오의 시공간 특징을 활용할 수 있어야 한다. 두 번째로, 비디오에서의 시각적 관계 정보는 시간이 지남에 따라 변할 수 있다[17]. 특히, 같은 비디오 안에서 동일한 두 객체가 계속해서 등장함에도 불구하고 두 객체 사이의 시각적 관계는 시간이 지남에 따라 변할 수 있다. 예를 들어, 그림 1에서 볼 수 있듯이 두 객체 “dog”와 “ball” 사이의 관계가 시간이 지남에 따라 “chase”에서 “bite”로 변하는 것을 확인할 수 있다.

Ⅲ. 비디오 시각적 관계 이해 기술 동향

비디오 시각적 관계 이해 기술은 크게 두 종류로 생각할 수 있다. 첫 번째는 비디오 단위(Video-level)의 시각적 관계 이해 기술로, 주어진 입력 비디오에 대해 비디오 전체에서 등장하는 객체와 시간에 따른 객체의 이동 경로를 구한 후 객체 간 관계를 찾는 방법이다. 이 방법은 먼저 비디오를 여러 개의 작은 세그먼트(Segment)로 분할한 후 세그먼트별로 객체 탐지(Object Detection)와 추적(Object Tracking) 알고리듬을 이용하여 세그먼트 내 객체의 이동 경로(Object Tracklet Proposal)를 구한다. 그다음, 세그먼트별로 두 개의 객체 쌍을 구성하여 두 객체가 어떠한 시각적 관계를 갖는지를 파악한다. 이때, 세그먼트 내 모든 객체 쌍을 고려할 수도 있고, 관련성이 높은 객체 쌍을 따로 분류해내어 고려할 수도 있다. 마지막으로 세그먼트별 시각적 관계 정보를 기반으로 동일한 관계를 포함하는 인접한 두 세그먼트를 연결(Association)하여 전체 비디오의 시각적 관계 이해 결과를 도출해낸다.

두 번째는 프레임 단위(Frame-level)의 시각적 관계 이해 기술로, 프레임별로 객체 간 관계를 찾은 후 시간에 따른 프레임별 그래프를 연결하여 전체 비디오의 시각적 관계를 찾는 방법이다. 이 방법은 먼저 비디오의 프레임별로 객체를 탐지한 후 공간 맥락 정보(Spatial Context)를 추출한다. 이렇게 구한 공간 맥락 정보를 이용하여 프레임 간 시간 맥락 정보(Temporal Context)를 계산한다. 이때, 일정 개수의 프레임 간 시간 맥락 정보를 계산할 수도 있고, 현재 프레임과 과거 프레임들 간의 시간 맥락 정보를 계산할 수도 있다. 이렇게 구한 시공간 맥락 정보(Spatio-temporal Context)는 연관된 객체 간 시각적 관계를 분류하는 데 사용된다. 프레임 단위의 시각적 관계 이해 기술은 공간상의 객체 간 연관 관계를 찾는 것뿐만 아니라 서로 다른 프레임 간 연관 관계를 찾을 수 있어야 하므로 많은 연구에서 트랜스포머(Transformer) 모델[18]을 활용하는 경우가 많다.

1. 비디오 단위의 시각적 관계 이해 기술

VidVRD[17]는 처음으로 비디오 단위의 시각적 관계 이해 기술을 제안한 방법이다. 비디오를 30프레임 단위의 작은 세그먼트로 나눈 후 세그먼트 내 객체의 이동 경로(Tracklet Proposal)를 계산한다. 세그먼트별로 두 개의 객체 쌍을 구성한 후 두 객체의 특징 벡터를 합쳐 하나의 관계 특징 벡터(Relation Feature Vector)를 생성한다. VidVRD는 이렇게 구한 관계 특징 벡터를 이용하여 하나의 관계 트리플렛 〈subject, predicate, object〉의 subject, predicate, object를 각각 분류해내는 3개의 분류기를 동시에 학습시킨다. 마지막으로 세그먼트별로 구한 관계 트리플렛들을 대상으로, 인접한 세그먼트 내 동일한 관계 트리플렛 중 위치가 충분히 겹치는 트리플렛은 장시간에 걸친 하나의 시각적 관계 트리플렛으로 합치는 작업을 greedy algorithm으로 수행한다.

VidVRD 방법은 비디오가 갖는 많은 시공간 정보를 효과적으로 활용했다고 보기 어렵다. 이러한 점에서 비디오의 시공간 정보를 활용하기 위한 방법들이 고안되었다. GSTEG[19]는 각 객체를 노드로, 노드 간 통계적 관계를 에지로 표현하는 완전히 연결된 시공간 그래프상의 Conditional Random Field(CRF)로 구성하여 객체 간 시공간적 관계를 포착하고자 하였다. VRD-GCN[20]은 각 객체를 노드로 하는 완전히 연결된 시공간 그래프를 구성하고, Graph Convolutaioal Network(GCN)[21]을 활용하여 시각적 관계 정보를 찾고자 하였다. 여기에 더해 VRD-GCN은 인접한 세그먼트 간 시각적 유사성과 공간상의 위치 정보를 더 잘 반영하기 위한 Siamese Network[22] 기반 시각적 관계 정보 병합을 제안하였다.

VidVRD 방법의 또 다른 문제점은 이미 설정된 길이의 세그먼트 단위로 관계 트리플렛을 계산한다는 점이다. 이로 인해 장시간에 걸쳐 나타나는 관계 정보를 얻기 힘들고, 세그먼트별로 독립적으로 계산을 수행하므로 인접한 세그먼트 안에 동일한 시각적 관계 정보를 포함함에도 불구하고 중복해서 계산을 수행한다는 상황이 발생하게 된다. 이러한 문제를 해결하기 위해 PPN-STGCN[15]은 슬라이딩 윈도우(Sliding Window) 방법을 이용하여 다양한 크기의 객체의 부분 경로(Tracklet Proposal)를 생성한다. 이를 통해 장시간에 걸친 시각적 관계 트리플렛을 찾을 수 있고, 연속된 프레임에 보이는 동일한 관계를 한 번에 찾게 되어 계산의 효율성을 높일 수 있다.

몇몇 연구에서는 비디오 시각적 관계 이해의 관계 트리플렛 〈subject, predicate, object〉의 각 요소 subject, predicate, object는 의미적으로 서로 긴밀하게 연결되어 있는 점을 주목하였다. 이를 통해 시각적 관계 이해 기술에서 각 요소를 인식하고자 할 때 다른 요소의 정보를 같이 활용할 수 있다면 더 향상된 결과를 얻을 수 있을 것으로 생각할 수 있다. 3DRN[23]은 관계 트리플렛의 subject, object의 특징 정보뿐만 아니라 각 객체의 클래스 정보, 상대적 위치 정보를 반영할 수 있는 추가적인 branch를 구성하여 predicate 인식에 활용하였다. VidVRD-II[24]는 여러 가지의 이유로 subject, object를 제대로 인식하지 못함에 따라 predicate 인식에도 영향을 주는 경우를 고려하였다. 앞서 언급했듯이 관계 트리플렛의 각 요소는 다른 요소와 긴밀하게 연결되어 있기 때문에, 어느 한 요소가 시각적으로 불분명할지라도 다른 두 요소를 통해 어느 정도 해당 요소가 무엇을 의미하는지 한정할 수 있다는 점을 생각할 수 있다. VidVRD-II는 시각적 관계 인식 과정을 다단계로 구성하고 이전 단계의 각 요소의 인식 결과를 다음 단계의 다른 요소 추론의 입력으로 넣어 각 요소의 인식 결과를 점진적으로 개선하는 방법을 제안

하였다.

한편, 기존의 비디오 단위의 시각적 관계 이해 기술들은 세그먼트별 객체 검출과 추적, 세그먼트별 두 객체 간 관계 분석, 인접 세그먼트의 관계 정보 연결 과정을 거치는 동일한 파이프라인을 적용하여 구현되었다. 근래에는 이러한 틀을 벗어나 다른 형태의 접근 방법을 제안한 연구들이 발표되었다. Social Fabric[25]은 비디오를 작은 단위의 세그먼트로 나누는 대신, 전체 비디오에서 두 객체의 쌍이 관계가 있을 구간을 먼저 선별한다. 먼저 비디오 전체 구간에서 객체 검출과 추적을 수행하여 모든 객체의 이동 경로를 구한 후 두 객체가 같이 등장하는 각 프레임에 대해 두 객체가 서로 상호작용하는 관계 형성 여부(Interactivityness)를 계산한다. 그다음 관계 형성 가능성이 높은 구간에 대해 객체 간 어떤 시각적 관계가 있는지를 파악하여 관계 트리플렛을 얻을 수 있다. BIG[26]은 기존의 방법과 반대로, 전체 비디오에서 등장하는 관계 트리플렛을 구한 후 각각의 관계 트리플렛이 어느 구간에서 등장하는지를 파악한다. 먼저 모든 객체의 이동 경로로부터 구한 특징 정보를 이용하여 트랜스포머 모델[18]을 이용하여 시각적 관계를 갖는 두 객체(subject, object)를 찾은 후 해당 관계(predicate)를 분류한다. 그다음, 관계 트리플렛을 일종의 언어 질의(language query)로 생각하는 video grounding problem[27]으로 구상하여 각 관계 트리플렛의 시간 구간을 찾게 된다.

2. 프레임 단위의 시각적 관계 이해 기술

TRACE[28]는 시공간 맥락 정보의 계산과 객체 추적 연산을 분리하여 문제를 단순화하였다. 먼저 공간상의 거리를 기준으로 가능한 시각적 관계 후보들을 계층적인 트리(Tree) 구조로 관리한 후 각 관계 후보들을 대상으로 어텐션 메커니즘[18]을 적용하여 시각적 관계 인식에 필요한 시공간 맥락 정보를 계산한다. 마지막으로 프레임 단위의 시각적 관계 이해 결과를 비디오 단위의 결과로 만들기 위해 객체 추적 결과를 활용하여 동일한 시각적 관계들을 합치는 간단한 작업을 수행한다.

STTran[29]은 기본적인 트랜스포머 모델을 기반으로 시공간 맥락 정보를 반영하는 시각적 관계 이해 기술이다. STTran은 비디오의 프레임별로 공간 맥락 정보를 계산하는 spatial encoder와 이미 정의된 개수의 여러 프레임 간의 시간 맥락 정보를 계산하는 temporal decoder로 구성되어 있다. 먼저 프레임별로 두 객체의 정보를 결합하여 하나의 관계 특징 벡터(Relation Representation)를 구한 후 spatial encoder의 입력으로 넣는다. 그 다음 공간 맥락 정보를 반영하는 프레임별 특징 벡터를 일정 개수만큼 모아 temporal decoder의 입력으로 넣는다. 이렇게 시간 맥락 정보까지 반영하여 얻은 최종 특징 벡터는 시각적 관계를 결정하는 분류기를 통해 어떤 관계인지를 판별하게 된다. 이때, 두 객체 사이에 여러 개의 시각적 관계가 나타날 수 있다는 점을 고려하여 일정 임계치(Threshold)를 넘는 신뢰도(Confidence)를 갖는 모든 시각적 관계를 출력으로 내놓는다.

STTran 방법의 단점은 프레임 간 시간 맥락 정보를 계산할 때 너무 짧은 구간만을 고려하기 때문에 장기간에 걸친 시간 맥락 정보가 누락될 수 있다는 점이다. 이를 해결하기 위해 장기간의 시간 맥락 정보를 계산하기 위한 몇몇 연구가 진행되었다. Li 등[30]은 과거 프레임만으로 아직 관측하지 않은 현재 프레임의 시각적 관계를 예측하는 사전 학습(Pre-training) 단계, 현재 프레임을 관측한 후 현재 프레임의 시각적 관계 결과를 조정하는 미세 조정(Fine-tuning) 단계로 구분하여 시각적 관계 이해를 수행하고자 하였다. 사전 학습 단계에서 과거 프레임들 간의 장단기 시간 맥락 정보를 계산하기

위해 동일한 객체 쌍의 프레임별 관계 특징 벡터를 입력으로 하는 short-term encoder와 short-term encoder의 출력과 더 긴 길이의 프레임별 관계 특징 벡터를 입력으로 하는 long-term encoder를 포함하는 progressive temporal encoder를 제안하였다. DSG-DETR[31]은 장시간에 걸친 객체 단위의 일관성과 객체 간 관계의 시간적 특성을 반영하는 것이 시각적 관계 이해에 도움이 된다는 점에서 착안하였다. DSG-DETR의 가장 큰 특징은 객체 추적을 통해 얻은 객체의 이동 경로를 object transformer의 입력으로 하여 안정적인 객체 정보를 얻고 이를 기반으로 객체 간 관계의 시간 맥락 정보를 계산한다는 점이다. 이때 객체 추적은 온라인 방식으로 동작하며, 정확한 객체의 이동 경로를 찾는 것보다는 트랜스포머 모델에게 관련된 특징 정보만을 효율적으로 주목할 수 있게끔 안내해주는 역할을 수행한다.

3. 시각적 관계 정보의 편향성 문제

시각적 관계 이해 기술에서 겪는 주요 고충 중 하나는 시각적 관계 정보의 편향성(Bias)이다. 시각적 관계 이해 데이터셋을 세부적으로 살펴보면 “next to”, “in front of”, “behind” 등 당연하고 중요하지 않은 시각적 관계 정보는 상당히 자주 등장하는 반면, “eating”, “writing on” 등 보다 상세한 의미를 담는 시각적 관계 정보는 드물게 존재한다[32]. 이처럼 균등하지 않은 시각적 관계 정보의 편향성으로 인해 기존의 시각적 관계 이해 기술들은 자주 등장하는 시각적 관계 정보는 잘 맞추지만 자주 등장하지 않는 시각적 관계 정보에서는 성능을 보장할 수 없는 문제가 발생하였다. 특히, 비디오 시각적 관계 이해의 경우 앞서 언급한 것처럼 행동이나 동적인 관계까지 반영하는 시각적 관계가 추가로 더 등장하기 때문에 이러한 편향성은 더욱 심화되어 나타난다[16].

이러한 문제를 해결하기 위한 몇 가지 기술이 제안되었다. IVRD[16]은 인과 추론(Causal Inference) 방법을 도입하여 시각적 관계 정보의 편향성을 완화하고자 하였다. IVRD는 각각의 시각적 관계 정보를 대표하는 관계 프로토타입(Predicate Prototype)을 구축한 후 시각적 관계 추론 시 모든 관계 프로토타입을 공정하게 고려할 수 있도록 강제함으로써 드물게 나타나는 시각적 관계도 잘 찾을 수 있도록 하였다. MVSGG[33]는 모델 학습 시 메타 학습(Meta Learning) 방법을 적용하여 편향성을 배제하고 시각적 관계가 갖는 일반적인 특징을 학습할 수 있는 방법을 제안하였다. MVSGG는 학습 데이터셋을 서로 다른 데이터 분포를 갖는 support set과 query set으로 나누어 모델을 support set으로 학습시킨 후 query set으로 평가한 결과를 모델의 학습에 추가로 반영함으로써 편향성의 내재화를 방지하고 모델을 일반화시킬 수 있도록 하였다. TEMPURA[32]는 로우샷 학습(Low-shot Learning) 방법에서 자주 사용되었던 기억 유도 학습(Memory-guided Training) 방법을 활용하는 방법을 제안하였다. TEMPURA는 시각적 관계 정보의 프로토타입(Prototype)을 계산하여 메모리 뱅크(Memory Bank)에 저장해놓은 후 모델 학습 시에는 메모리 뱅크를 이용하여 관련 특징 정보를 계산한 결과를 일정 비율로 모델 업데이트에 활용하고, 시각적 관계 추론 시에는 메모리 뱅크를 시각적 관계 특징 추출 시 편향성을 줄여주는 역할로 사용한다.

Ⅳ. 벤치마크 데이터셋

이 장에서는 비디오 시각적 관계 이해 기술의 모델 학습과 성능 검증을 위해 사용되는 벤치마크 데이터셋에 대해 알아본다. 표 1은 최신 연구에서 주

표 1 벤치마크 데이터셋의 통계

데이터셋	영상 개수	평균 길이	객체 종류	관계 종류
ImageNet-VidVRD	1,000개	약 10초	35개	132개
VidOR	10,000개	약 35초	80개	50개
Action Genome	9,848개	약 30초	36개	25개

로 사용하는 벤치마크 데이터셋의 간략한 통계 정보를 나타낸다.

1. ImageNet-VidVRD

ImageNet-VidVRD 데이터셋[17]은 ILSVRC2016-VID[13]의 학습 데이터셋과 검증 데이터셋을 이용하여 만든 벤치마크 데이터셋이다. ILSVRC2016-VID 데이터셋의 영상 중 명확하고 충분히 많은 관계, 행동 정보를 포함하는 1,000개의 영상을 선별하고, 이 중 800개는 학습 데이터셋으로, 200개는 테스트 데이터셋으로 구분하였다. ImageNet-VidVRD 데이터셋의 영상의 총길이는 약 3시간이며, 평균 길이는 약 10초 정도이다. ILSVRC2016-VID 데이터셋은 30개 객체 클래스를 다루고 있는데, ImageNet-VidVRD 데이터셋은 여기에 관계, 행동 정보에서 자주 등장하는 5개 객체 클래스를 추가로 레이블링(Labeling)하여 총 35개의 객체 클래스를 다룬다.

ImageNet-VidVRD 데이터셋에서 다루는 predicate의 종류는 다음과 같이 구상하였다. 먼저 "bite", "chase", "ride" 등과 같은 타동사(Transitive Verb)는 바로 predicate로 사용한다. "fast", "tall" 등과 같은 형용사(Adjective)는 "faster", "taller" 등과 같은 비교급(Comparative)으로 변환하여 predicate로 사용한다. "above", "behind", "on the left of" 등 공간 관계를 나타내는 spatial predicate는 임의로 정의하여 사용한다. 이때, 공간 관계의 기준은 카메라 관점(Viewpoint)이다. 자동사(Intransitive Verb)의 경우, "walk behind", "walk next to" 등과 같이 spatial predicate와 결합하거나 "run with" 등과 같이 "with"와 결합하여 predicate로 사용한다. ImageNet-VidVRD 데이터셋에서는 14개의 타동사, 3개의 비교급, 11개의 공간 표현, 11개의 자동사를 사용하여 총 160개의 predicate를 만들었으며, 실제 영상에서 보이는 132개의 predicate를 사용하였다.

2. VidOR

VidOR 데이터셋[34]은 대용량 비디오 데이터셋인 YFCC-100M[35]에서 가져온 비디오를 이용하여 만들어졌다. 이 중에서 사람이 보기에도 어려운 비디오와 객체가 너무 많거나 적은 비디오를 제외한 10,000개의 영상을 선별하였고, 이를 다시 학습/검증/테스트 데이터셋으로 사용하기 위해 각각 7,000개, 835개, 2,165개로 구분하였다. VidOR 데이터셋의 영상 총길이는 약 98.6시간이며, 평균 길이는 약 35초 정도이다. VidOR 데이터셋은 총 80개 객체 클래스를 다루고 있으며, 상위 범주(Category)인 사람(Human), 동물(Animal), 기타(Other) 아래에 각각 3개, 28개, 49개의 하위 객체 클래스를 포함하고 있다. VidOR 데이터셋의 영상에서 등장하는 객체의 비율은 사람 범주가 56.34%, 동물 범주가 35.78%, 기타 범주가 7.98%로 long-tail distribution을 따르고 있음을 알 수 있다.

VidOR 데이터셋에서 다루는 predicate의 종류는 총 50개로, 42개의 단일 행동(Atomic Action)과 8개의 공간 표현을 predicate로 사용한다. 이때, 공간 관계의 기준은 카메라 관점이 아닌 객체의 관점이다. 예를 들어, 트리플렛 〈car, behind, child〉은 카메라의

위치가 어디에 있든 상관없이 child의 등 뒤에 car가 있음을 의미한다. 만약 객체의 방향(Orientation)을 알 수 없다면 레이블링하지 않는다.

3. Action Genome

Action Genome[36]은 실내(Indoor) 활동 장면을 찍은 Charades 데이터셋[37]을 이용하여 만들어졌다. Action Genome 데이터셋은 Charades 데이터셋에 있는 9,848개의 영상을 모두 사용하였으며, Charades 데이터셋의 구성과 같이 7,985개는 학습 데이터셋으로, 1,863개의 테스트 데이터셋으로 구분하였다. 총 영상의 길이는 약 82시간이며, 평균 길이는 약 30초 정도이다. Action Genome 데이터셋은 사람 클래스와 35개의 사물 클래스를 포함하여 총 36개의 객체 클래스를 다루고 있다.

Action Genome 데이터셋은 인지 과학(Cognitive Science)에서 착안하여 사람의 행동(Action)을 시간에 따른 사람과 사물 간 상호작용(Interaction)의 변화로 표현하고자 했다. 따라서 ImageNet-VidVRD와 VidOR 데이터셋과 달리, Action Genome 데이터셋은 사람의 행동과 관련된 사물에 대해서만 트리플렛 레이블링이 되어 있으며 사람의 행동 구간 내에서 프레임들을 샘플링하여 프레임 단위(Frame-level)의 트리플렛 레이블을 제공한다. Action Genome 데이터셋은 총 25개의 predicate를 다루고 있으며, 상위 범주인 attention, spatial, contact 아래에 각각 3개, 6개, 16개의 predicate를 포함한다.

V. 결론

본고에서는 비디오 시각적 관계 이해 기술에 동향과 관련된 벤치마크 데이터셋에 대해 살펴보았다. 먼저 비디오 시각적 관계 이해 기술에 대해 소개하고, 크게 두 갈래인 비디오 단위, 프레임 단위의 시각적 관계 이해 기술로 나누어 각 방법의 특징과 기술 동향에 대해 소개하였다. 더불어 비디오 시각적 관계 이해 기술에서 고려해야 할 편향성 문제에 대해 고찰하고 관련 기술에 대해서도 설명하였다. 그리고 비디오 시각적 관계 이해 기술의 모델 학습과 성능 검증을 위한 벤치마크 데이터셋을 살펴보았다.

비디오 시각적 관계 이해 기술은 비디오를 고수준으로 이해하기 위해 필요한 주요 기반 기술이나, 아직까지는 실제로 기술을 활용할 수 있을 만큼의 진척을 이루었다고 보기 어렵다. 하지만 많은 연구자가 계속해서 연구를 진행하고 있으며, 비디오를 대상으로 한 시각적 관계 이해 기술의 중요성은 더욱 커지고 있기 때문에 향후 비디오 시각적 관계 이해 기술의 발전 가능성이 매우 크다고 판단할 수 있다.

한편, 사람의 시각적 인식 방법에 가깝게 시각적 관계 이해 문제를 보다 정교하게 설계하고 새로운 문제 해결 방법을 고려하는 것도 생각해 볼 수 있다. 예를 들어, PVSG[38]는 기존의 비디오 시각적 관계 이해 기술이 객체 탐지 기술에 의존적이라는 점에 주목하여, 픽셀 단위의 객체/배경 분할(Panoptic Segmentation) 기술 기반의 비디오 시각적 관계 이해 기술을 제안하였다. 이처럼 시각적 관계 이해 기술의 문제를 정교하게 확장하여 영상을 더욱 사람과 비슷하게 이해하기 위한 연구도 앞으로 주목할 필요가 있다.

용어해설

관계 트리플렛 영상에서 등장하는 두 객체 간 시각적 관계를 표현하는 방법. 주어(Subject)와 목적어(Object)가 어떤 관계가 있는지 나타내는 서술어(Predicate)를 기술하여 시각적 관계를 나타내며, 통상 ⟨subject, predicate, object⟩의 형태로 표현함

참고문헌

[1] J. Johnson et al., "Image retrieval using scene graphs," in Proc. IEEE/CVF CVPR, (Boston, MA, USA), June 2015, pp. 3668-3678.

[2] C. Lu et al., "Visual relationship detection with language priors," in Proc. ECCV, Oct. 2016, pp. 852-569.

[3] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," Int. J. Comput. Vis., vol. 123, no. 1, May 2017, pp. 32-73.

[4] J. Ji et al., "Action genome: actions as compositions of spatio-temporal scene graphs," in Proc. IEEE/CVF CVPR, June 2020, pp. 10233-10244.

[5] Y. Zhong et al., "Comprehensive image captioning via scene graph decomposition," in Proc. ECCV, Aug. 2020, pp. 211-229.

[6] X. Yang et al., "Auto-encoding and distilling scene graphs for image captioning," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 5, May 2022, pp. 2313-2327.

[7] X. Lu and Y. Gao, "Guide and interact: SceneGraph based generation and control of video captions," Multimed. Syst., vol. 29, no. 2, Apr. 2023, pp. 797-809.

[8] C. Zhang et al., "An empirical study on leveraging scene graphs for visual question answering," in Proc. BMVC, Sept. 2019.

[9] L. Li et al., "Relation-aware graph attention network for visual question answering," in Proc. IEEE/CVF ICCV, Oct. 2019, pp. 10312-10321.

[10] J. Mao et al., "Dynamic multistep reasoning based on video scene graph for video question answering," in Proc. NAACL, Jul. 2022, pp. 3894-3904.

[11] M. Qi et al., "Online cross-modal scene retrieval by binary representation and semantic graph," in Proc. ACM MM, Oct. 2017, pp. 744-752.

[12] M. Daum et al., "VOCAL: Video organization and interactive compositional analytics," in Proc. CIDR, Jan. 2022.

[13] X. Chang et al., "A Comprehensive survey of scene graphs: generation and application," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, 2023, pp. 1-26.

[14] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis., vol. 115, no. 3, 2015, pp. 211-252.

[15] C. Liu et al., "Beyond short-term snippet: Video relation detection with spatio-temporal global context," in Proc. IEEE/CVF CVPR, June 2020, pp. 10837-10846.

[16] Y. Li et al., "Interventional video relation detection," in Proc. ACM MM, Oct. 2021, pp. 4091-4099.

[17] X. Shang et al., "Video visual relation detection," in Proc. ACM MM, Oct. 2017, pp. 1300-1308.

[18] A. Vaswani et al., "Attention is all you need," in Proc. NIPS, Dec. 2017, pp. 5998-6008.

[19] Y.H.H. Tsai et al., "Video relationship reasoning using gated spatio-temporal energy graph," in Proc. IEEE/CVF CVPR, June 2019, pp. 10416-10425.

[20] X. Qian et al., "Video relation detection with spatio-temporal graph," in Proc. ACM MM, Oct. 2019, pp. 84-93.

[21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. ICLR, Apr. 2017.

[22] L. Bertinetto et al., "Fully-connected siamese networks for object tracking," in Proc. ECCVW, Oct. 2016, pp. 850-865.

[23] Q. Cao et al., "3-D relation network for visual relation recognition in videos," Neurocomputing, vol. 432, 2021, pp. 91-100.

[24] X. Shang et al., "Video visual relation detection via iterative inference," in Proc. ACM MM, Oct. 2021, pp. 3654-3663.

[25] S. Chen et al., "Social fabric: tubelet compositions for video relation detection," in Proc. IEEE/CVF ICCV, Oct. 2021, pp. 13465-13474.

[26] K. Gao et al., "Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs," in Proc. IEEE/CVF CVPR, June 2022, pp. 19475-19484.

[27] C. Lu et al., "DEBUG: A dense bottom-up grounding approach for natural language video localization," in Proc. EMNLP-IJCNLP, Nov. 2019, pp. 5144-5153.

[28] Y. Teng et al., "Target adaptive context aggregation for video scene graph generation," in Proc. IEEE/CVF ICCV, Oct. 2021, pp. 13668-13677.

[29] Y. Cong et al., "Spatial-temporal transformer for dynamic scene graph generation," in Proc. IEEE/CVF ICCV, Oct. 2021, pp. 16352-16363.

[30] Y. Li et al., "Dynamic scene graph generation via anticipatory pre-training," in Proc. IEEE/CVF CVPR, June 2022, pp. 13864-13873.

[31] S. Feng et al., "Exploiting long-term dependencies for generating dynamic scene graphs," in Proc. IEEE/CVF WACV, Jan. 2023, pp. 5119-5128.

[32] S. Nag et al., "Unbiased Scene graph generation in videos," in Proc. IEEE/CVF CVPR, June 2023, pp. 22803-22813.

[33] L. Xu et al., "Meta spatio-temporal debiasing for video scene graph generation," in Proc. ECCV, Oct. 2022, pp. 374-390.

[34] X. Shang et al., “Annotating objects and relations in user-generated videos,” in Proc. ACM ICMR, June 2019, pp. 279-287.

[35] B. Thomee et al., “YFCC100M: The new data in multimedia research,” Commun. ACM, vol. 59, no. 2, 2016, pp. 64-73.

[36] J. Ji et al., “Action genome: actions as compositions of spatio-temporal scene graphs,” in Proc. IEEE/CVF CVPR, June 2020, pp. 10233-10244.

[37] G. A. Sigurdsson et al., “Hollywood in homes: Crowd-sourcing data collection for activity understanding,” in Proc. ECCV, Oct. 2016, pp. 510-526.

[38] J. Yang et al., “Panoptic video scene graph generation,” in Proc. IEEE/CVF CVPR, June 2023, pp. 18675-18685.