비디오 검색 기술 동향

Trends of Video-to-Video Retrieval Technology

조 원 [W. Jo, clockjow@gmail.com] 세종대학교 인공지능학과

In this paper, we analyze the trends of video-to-video retrieval technology. Due to the development of the video streaming industry in recent years, video-to-video retrieval technology has been requested for a variety of applications, including video filtering, recommendation, copyright protection and verification. Mainly, deep learning-based methods are used to retrieve similar videos in complex situations and examine them through various types of datasets. As a result, we introduce how deep learning-based methods are applied to video-to-video retrieval task and what datasets there are.

*본고는 2021 년도 세종대학교 컴퓨터비전 수업의 기말고사 대체로 수행된 연구임

I. 개요

네트워크의 발전 덕택에, 비디오 스트리밍 산업이 기하급수적으로 확장하고 있다. 일례인 Youtube 의 경우, 전세계에서 약 20 억명의 사용자를 보유하고 있으며, 매일 10 억 시간 이상의 비디오가 시청되고 있다[1]. 이러한 비디오 스트리밍 산업의 성장과 함께 다양한 비디오 검색 기술이 요구되고 있다.

비디오 검색 기술로는 입력의 형태에 따라 음성, 문자, 비디오와 같이 세 종류로 나뉜다. 그 중에서 비디오를 입력으로 하여 관련된 비디오를 찾아내는 기술(Video-to-Video Retrieval)이 시각적 요소만 활용하기에 주목 받기 시작하였다. 최초에는 저작권 보호와 같이 시각적으로 복제된 비디오를 찾는 NDVR(Near-Duplicated Video Retrieval)로부터 출발되었으며, 현재는 시각적 복제뿐만 아니라 의미론적으로 유사한지 판단하는 것을 요구하는 FIVR(Fine-grained Incident Video Retrieval)과 AVR(Action Video Retrieval)을 다루고 있다. 이와 같이 비디오를 입력으로 하는 비디오 검색 기술은 다양한 방향으로 연구되고 있으며, 비디오 스트리밍 산업에서의 여러 응용 가능성을 내비친다. 예를 들어, 사용자의 관점에서는 원하는 콘텐츠를 포함한 비디오를 반환하여 연관 비디오 추천에 활용될 수 있다. 또 다른 예시로, 스트리밍 플랫폼 의 관점에서는 사용자가 새롭게 업로드한 비디오에 복제된 장면이 있는지 판단하여 저작권 보호에 활용될 수 있다.

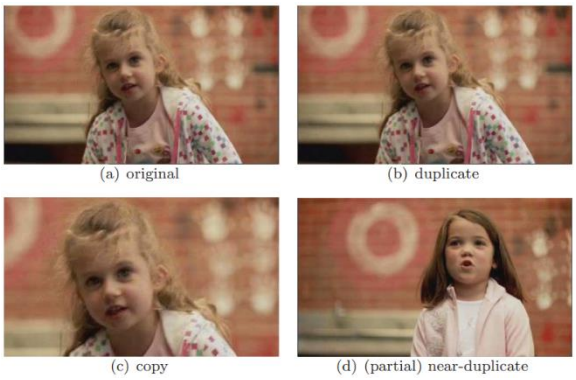
본고에서는 활발히 연구되고 있는 비디오 검색 분야에서도 입력이 비디오인 연구들의 동향을 다룬다. 우선, 해당 연구를 위해 사용되는 데이터 셋을 소개하고, 이후 비디오 검색 문제를 해결하고자 한 방법론에 대해 논의한다.

II. 비디오 검색 데이터 셋

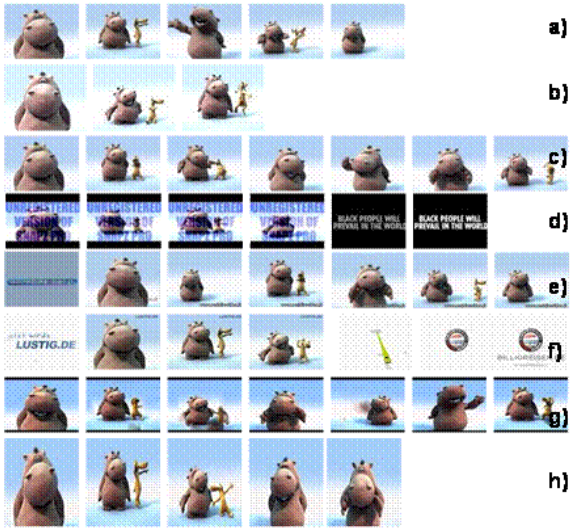
비디오 검색 데이터 셋은 풀고자 하는 비디오 검색 내의 분야에 따라 나뉜다. 이는 시각적 복제를 중점적으로 다루는 NDVR, 의미론적 유사성을 중점적으로 다루는 AVR, 그리고 이 둘을 포괄하고 있는 FIVR 이며, 해당 장에서는 앞선 구분 순서에 따라 설명한다.

2. NDVR

NDVR[2]은 Near-Duplicated Video Retrieval 의 약자로, 시각적으로 복제된 비디오 검색을 주된 목적으로 둔다. 여기서 near-duplicate 란 단순 복제를 포함하는 개념으로써 (그림 1)과 같이 시각적으로 매우 유사한 경우를 의미한다. 비디오를 입력으로 관련 비디오를 찾는 연구들은 최초에 NDVR 을 목적으로 시작되었기에 FIVR 과 AVR 에 비해 보다 다양한 데이터 셋이 존재한다.



(그림 1) 비디오에서 near-duplicate 의 개념,
원본(a) 대비 (b), (c), (d) 모두 near-duplicate 관계
[출처] Shen, H., Liu, J., Huang, Z., Ngo, C.-W., & Wang, W. (2013). Near-Duplicate Video Retrieval: Current Research and Future Trends. Multimedia, IEEE, 45, 1-1. https://doi.org/10.1109/MMUL.2011.39



(그림 2) CC_WEB_VIDEO 의 Near-Duplicate 관계 설명, (a)원본, (b)조도 및 해상도 변화, (c)프레임 레이트 변화, (d)텍스트 겹침 혹은 마지막 위치 콘텐츠 변화, (e, f)시작 및 마지막 위치 콘텐츠 변화, (g)테두리가 추가된 더 긴 버전, (h)해상도 단독 변화

[출처] Wu, X., Hauptmann, A., & Ngo, C.-W. (2007). Practical elimination of near-duplicates from web video search. 218–227. https://doi.org/10.1145/1291233.1291280

가. CC_WEB_VIDEO

CC_WEB_VIDEO[3] 데이터 셋은 2007 년도 카네기 멜론 대학(Carnegie Mellon Univer-sity)에서 NDVR 을 다루기 위해 ACM MM(ACM International Conference on Multimedia)에 공개한 데이터 셋이다. 총 12,790 개의 비디오로 구성되어 있으며, 이 중 서로 다른 24 개의 비디오 쿼리(Query)를 포함하고 있다. 또한, 각 쿼리와의 near-duplicate 관계는 (그림 2)와 같이 크게 7 가지로 나뉜다. Near-duplicate 비디오는 전체의 27% 이며, 이는 <표 1>에서 확인할 수 있다. 해당 비디오들은 각 쿼리의 주제를 Youtube, Google, 그리고 Yahoo 에서 검색한 뒤 수집되었으며, 비디오의 길이는 평균적으로 약 2.5 분 이다.

Queries			Near-Duplicate	
ID	Query	#	#	%
1	The lion sleeps tonight	792	334	42 %
2	Evolution of dance	483	122	25 %
3	Fold shirt	436	183	42 %
4	Cat massage	344	161	47 %
5	Ok go here it goes again	396	89	22 %
6	Urban ninja	771	45	6 %
7	Real life Simpsons	365	154	42 %
8	Free hugs	539	37	7 %
9	Where the hell is Matt	235	23	10 %
10	U2 and green day	297	52	18 %
11	Little superstar	377	59	16 %
12	Napoleon dynamite dance	881	146	17 %
13	I will survive Jesus	416	387	93 %
14	Ronaldinho ping pong	107	72	67 %
15	White and Nerdy	1771	696	39 %
16	Korean karaoke	205	20	10 %
17	Panic at the disco I write sins not tragedies	647	201	31 %
18	Bus uncle (巴士阿叔)	488	80	16 %
19	Sony Bravia	566	202	36 %
20	Changes Tupac	194	72	37 %
21	Afternoon delight	449	54	12 %
22	Numa Gary	422	32	8 %
23	Shakira hips don't lie	1322	234	18 %
24	India driving	287	26	9 %
Total		12790	3481	**27 %**

<표 1> CC_WEB_VIDEO 의 쿼리 및 Near-Duplicate 비디오 정보

[출처] Wu, X., Hauptmann, A., & Ngo, C.-W. (2007). Practical elimination of near-duplicates from web video search. 218–227. https://doi.org/10.1145/1291233.1291280

나. VCDB

VCDB[4] 데이터 셋은 2014 년도 푸단 대학(Fudan University)에서 비디오 복제 탐지(Video Copy Detection)를 목적으로 ECCV(European Conf-erence on Computer Vision)에 공개한 데이터 셋이다. 쿼리 528 개와 백그라운드(Background) 비디오 100,000 개로 구성되어 있으며, 비디오 검색을 위한 연관 관계에는 쿼리 내부적으로만 존재한다. 주로 비디오 수준(video-level)의 연관 관계만을 포함하는 다른 데이터 셋 들과는 달리, VCDB 는 (그림 3)과 같이 쿼리들 사이에 세그먼트 수준(segment-level)의 연관 관계 또한 포함하고 있다. 반면, 아무런 연관 관계가 존재하지 않는

(그림 3) VCDB 에서 세그먼트 수준의 연관 정보 예시
위-원본, 아래-복제

[출처] Jiang, Y.-G., Jiang, Y., & Wang, J. (2014). VCDB: A Large-Scale Database for Partial Copy Detection in Videos. ECCV.

백그라운드 비디오는 검색의 난이도를 높이기 위한 노이즈로만 사용된다. VCDB 데이터 셋은 당시에 제안된 비디오 데이터 셋 중에서는 가장 많은 양의 비디오를 포함하고 있으며, 백그라운드 데이터 셋이 쿼리와의 연관 관계가 없다는 점으로 인하여, 여러 비디오 검색 방법론들이 제안된 방식의 일반화 능력을 보이고자 해당 데이터 셋의 백그라운드 데이터 셋으로 학습하고 있다.

2. AVR

AVR 은 Action Video Retrieval 의 약자로, 유사한 행동이 나타나는 비디오 검색을 주된 목적으로 둔다. 시각적으로 유사한 장면, 인물이 포함된 비디오를 연관된 비디오로 검색하는 NDVR 과는 달리, AVR 은 유사한 행동이 나타나는 비디오만을 연관 관계로 판단한다.

가. ActivityNet

ActivityNet[5] 데이터 셋은 2015 년도 콜롬비아 바랑키야 대학(Universidad del Norte, Colombia) 에서 행동 이해(Activity Understanding)을 위해 CVPR(Conference on Computer Vision and Pattern Recognition)에 공개한 데이터 셋이다. (그림 4)와 같이 분류학적으로 행동을 나눠 총 10,024 개의 학습용 비디오, 4,926 개의 검증용

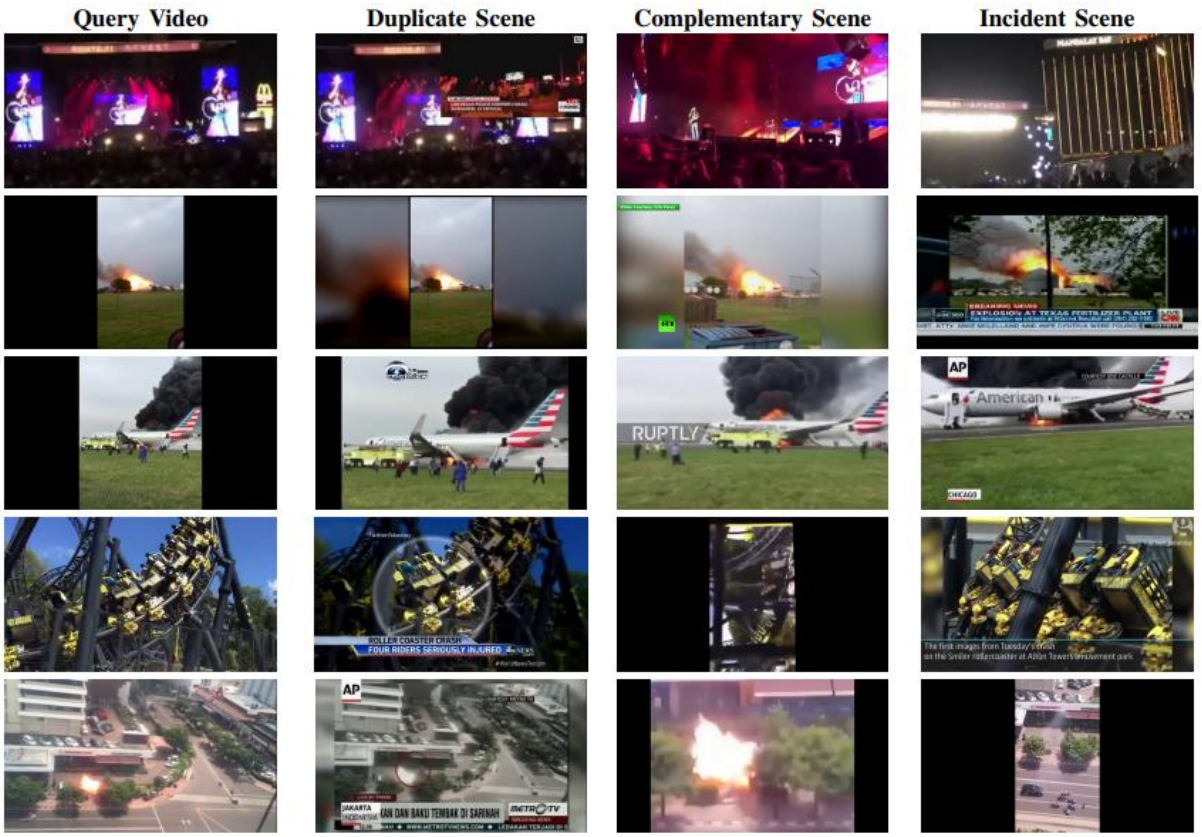
(그림 4) Activitynet 에서 분류학적으로 나눈 행동 시각화

[출처] Fabian Caba Heilbron, B. G., Victor Escorcia, & Niebles, J. C. (2015). ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 961-970.

비디오, 그리고 5,044 개의 평가용 비디오로 구성되어 있으며, 주로 행동 분류(Action Recognition) 혹은 비디오 내 행동 탐지(Temporal Action Detection)을 위해 사용되었다. ActivityNet 을 AVR 에서 활용하기 위해 [6]가 2018 년도 ECCV 에 재구성된 ActivityNet 을 제안하였으며, 현재까지 이처럼 재구성된 데이터 셋으로 AVR 을 다루고 있다.

3. FIVR

FIVR 은 Fine-grained Video Retrieval 의 약자로 유사한 사건이 나타나는 비디오 검색을 목적으로 둔다. 시각적으로 나타나는 장면뿐만 아니라 특정 사건에서 발생하는 인물의 행동 또한 고려해야 하기에 앞선 두 종류를 포함하고 있다. 해당 문제를 해결하기 위해서는 한 프레임 내의 디테일한 장면과 프레임 간의 연속된 정보를 동시에 나타내야 하기에 난이도가 높다.



(그림 5) FIVR-200 에서 정의된 라벨 예시, Duplicate Scene(DS), Complementary Scene(CS), Incident Scene(IS)
[출처] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2018). FIVR: Fine-grained Incident Video Retrieval. CoRR, abs/1809.04094. http://arxiv.org/abs/1809.04094

가. FIVR-200K

FIVR-200K[7] 데이터 셋은 2019 년도 ITI (Infor-mation Technologies Institute) 소속 MKLab 에서 사건 비디오 검색(Incident Video Retrieval)을 목적으로 IEEE Transactions on Multimedia 에 공개한 데이터 셋이다. 해당 데이터 셋에는 쿼리 100 개를 포함한 총 225,960 개의 비디오로 구성되어 있다. 또한, 비디오 간의 연관 관계를 총 5 가지로 정의하여, 쿼리와 나머지 비디오 간의 라벨(label)을 제공한다. 정의된 라벨은 각각 ND, DS, CS, IS, 그리고 DI 이며, ND 에 가까울수록 시각적 유사도를 중요시하고 IS 에 가까울수록 의미론적 유사도를 중요시한다. 또한, 쿼리와 아무런 관계가 없는 비디오는 DI 로 라벨링된다. 이렇게 정해진 라벨을 기준으로 FIVR-200K 내에서 세 가지 검색 태스크 DSVR, CSVR, ISVR 로 나뉜다. DSVR 의 경우 ND, DS 라벨인 비디오를 연관된 비디오로 선정하며, CSVR 의 경우 ND, DS, CS 라벨인 비디오를 연관된 비디오로 선정한다. 그리고 ISVR 의 경우 ND, CS, DS, IS 라벨인 비디오를 연관된 비디오로 선정한다. 이와 같은 세가지 검색 태스크를 통해 제안된 검색 방법론이 어떤 유형의 유사도에 장점을 보이는지 판단할 수 있다.

III. 비디오 검색 기술

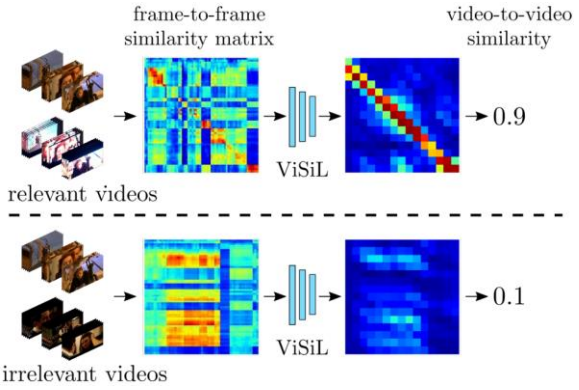
비디오 검색 기술의 핵심은 두 비디오 간의 유사도를 산정하는 것이다. 이러한 유사도를 구하는 방법에 따라 프레임 기술자 기반 방법론과 비디오 기술자 기반 방법론으로 나뉜다.

1. 프레임 기술자 기반 방법론

프레임 기술자 기반 방법론은 두 비디오 간의 유사도 연산 시, 프레임 기술자(frame-level feature)를 활용하는 방식이다. 주로 두 비디오에서 매 프레임 단위로 기술하고 프레임 간의 유사도를 연산한 뒤, 이를 응집(aggregation)하여 비디오 유사도를 산출하는 방식이다. 이 같은 과정으로 인해 계산 복잡도가 높아 검색 속도가 느리지만, 자세하게 판단하기에 일반적으로 검색 성능이 높다. 또한, 각 프레임 내의 정보를 우선시하기에 시각적 유사도를 중요시하는 상황에서 효과적인 방법론이다.

가. ViSiL

ViSiL[8]은 Video Similarity Learning 의 약자로, 2019 년도 ITI (Information Technologies Institute) 소속 MKLab 에서 ICCV (International Conference on Computer Vision)에 제안한 프레임 기술자 기반 방법론이다. 해당 방법론은 두 비디오를 프레임 단위로 기술한 뒤, (그림 6)과 같이 프레임 간의 유사도 매트릭스(matrix)를 구성하고 이를 응집(aggregation) 하여 비디오 유사도를 산출해낸다. 그리고 연관 관계에 놓여 있는 비디오의 경우 계산된 비디오 유사도가 높게, 반대의 경우 낮게 나타나도록 하는 방식의 학습 과정을 거친다.

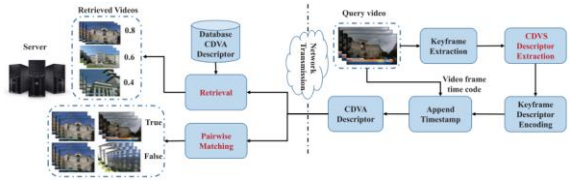


(그림 6) ViSiL 에서 유사도 매트릭스를 통해 비디오 유사도를 계산하는 과정

[출처] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019). ViSiL: Fine-Grained Spatio-Temporal Video Similarity Learning. 6350-6359. https://doi.org/10.1109/ICCV.2019.00645

나. CDVA

CDVA[9]는 Compact Descriptors for Video Analysis 의 약자로, 2019 년도에 ISO./IEC MPEG (Moving Picture Experts Group)에서 비디오 검색해 표준으로 제정한 기술이다. CDVA 는 프레임 기술자 기반의 방법으로, (그림 7)에서와 같이 비디오 프레임에 존재하는 높은 중복성을 피해 검색 효율성을 높이고자 칼라 히스토그램(color histogram) 기반 키프레임(keyframe) 선정 방식을 도입하였다. 또한, 선정된 키프레임 마다 전통적인 시각적 불변성 기술자 기반 표준 CDVS (Compact Descriptors for Visual Search)[10]와 회전 강인성을 포함한 딥러닝 기반의 기술자 NIP (Nested Invariance Pooling)[11]를 추출하여 강인한 비디오 검색이 가능하도록 설계된 방법론이다.



(그림 7) CDVA 프레임워크

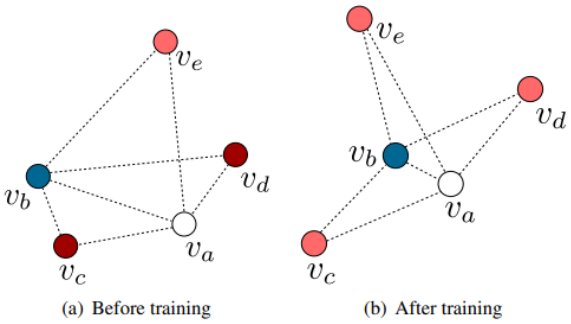
[출처] Lou, Y., Bai, Y., Lin, J., Wang, S., Chen, J., Chandrasekhar, V., Duan, L.-Y., Huang, T., Kot, A. C., & Gao, W. (2017). Compact Deep Invariant Descriptors for Video Retrieval. 2017 Data Compression Conference (DCC), 420–429. https://doi.org/10.1109/DCC.2017.31

2. 비디오 기술자 기반 방법론

비디오 기술자 기반 방법론은 두 비디오 간의 유사도 연산 시, 비디오 기술자(video-level feature)를 활용하는 방식이다. 이때, 각 비디오에서 일부 프레임을 선정해 기술된 단 하나의 기술자를 비디오 기술자라고 한다. 해당 방법론들은 비디오 전체를 하나의 벡터 내에 표현할 수 있어야하기 때문에 비디오 길이가 길어질수록 검색 성능이 떨어진다는 단점이 있다. 그러나, 속도 측면에서 프레임 기술자 기반 방법론과 비교하였을 때, 두 비디오 간의 유사도는 한번의 연산으로만 결정되기에 빠른 검색 속도를 보장할 수 있다. 이 같은 장점을 활용하여 실제 검색 상황에 사용되기 위해서는 앞으로 검색 성능을 높이기 위한 연구들이 요구된다.

가. DML

DML[12]은 Deep Metric Learning 의 약자로, 2017 년도 ITI (Information Technologies Institute) 소속 MKLab 에서 ICCVW (International Conference on Computer Vision Workshop)에 제안한 비디오 기술자 기반 방법론이다. DML 은 첫번째로 near-duplicate 상황을 다룬 딥러닝 기반의 비디오 기술자 기반 방법론이며, 트리플렛 (triplet)을 통한 기술자 거리 학습으로 하여금 (그림 8)과 같이 서로 연관된 비디오 기술자의 표현력을 높이고자 한 방법론이다.

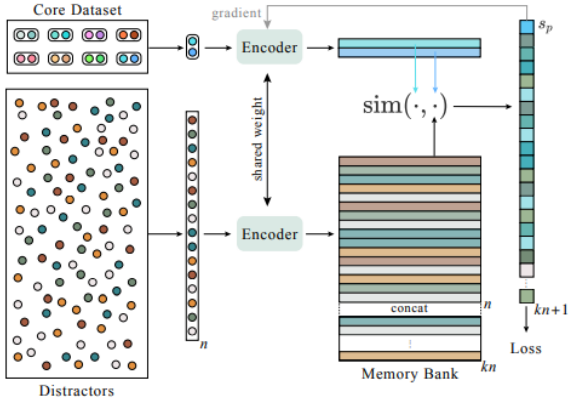


(그림 8) DML 에서의 학습 전후 기술자, 청색-쿼리, 백색-NEAR-DUPLICATE 비디오, 적색-관련 없는 비디오

[출처] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-Duplicate Video Retrieval with Deep Metric Learning. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 347–356. https://doi.org/10.1109/ICCVW.2017.49

나. TCA

TCA[13]는 Temporal Context Aggregation 의 약자로, 2021 년도 ByteDance AI Lab 에서 WACV(Winter Conference on Applications of Computer Vision)에 제안한 비디오 기술자 기반 방법론이다. 해당 방법론은 프레임 기술자들 간의 자가 어텐션(self-attention)을 통해 장 기간의 시간적 맥락을 비디오 기술자에 포함시키고자 한 딥러닝 기반 방법론이다. 뿐만 아니라, (그림 9)와 같이 학습 시 메모리 뱅크 (memory bank) 개념을 포함하여, 구별하기 어려운 하드 네거티브 (hard negative) 데이터에 대응 가능하도록 설계된 방법론이다.



(그림 9) TCA 에서 메모리 뱅크를 활용한 학습 과정
[출처] Shao, J., Wen, X., Zhao, B., & Xue, X. (2021). Temporal Context Aggregation for Video Retrieval with Contrastive Learning. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 3267–3277. https://doi.org/10.1109/WACV48630.2021.00331

IV. 결론

비디오 스트리밍 시장의 성장과 함께 다양한 비디오 검색 연구가 요구되고 있다. 이 같은 요구의 힘입어 해당 연구를 뒷받침해줄 여러 종류의 데이터 셋도 공개되어 있고 다양한 비디오 검색 기술도 제안되어 왔다. 그러나, 현존하는 비디오 검색 기술들은 각각의 장단점을 보이고 있기에, 실제 스트리밍 시장에서 활용되기 위한 앞으로의 비디오 검색 기술은 프레임 기술자 기반의 장점인 높은 성능과 비디오 기술자 기반의 장점인 빠른 검색 속도를 동시에 포함하는 방향으로 나아가야할 것이다.

Reference

[1] Acosta, T., Acosta-Vargas, P., Zambrano-Miranda, J., & Lujan-Mora, S. (2020). Web Accessibility evaluation of videos published on YouTube by worldwide top-ranking universities. IEEE Access, 8, 110994–111011.

[2] Shen, H., Liu, J., Huang, Z., Ngo, C.-W., & Wang, W. (2013). Near-Duplicate Video Retrieval: Current Research and Future Trends. Multimedia, IEEE, 45, 1–1. https://doi.org/10.1109/MMUL.2011.39

[3] Wu, X., Hauptmann, A., & Ngo, C.-W. (2007). Practical elimination of near-duplicates from web video search. 218–227. https://doi.org/10.1145/1291233.1291280

[4] Jiang, Y.-G., Jiang, Y., & Wang, J. (2014). VCDB: A Large-Scale Database for Partial Copy Detection in Videos. ECCV.

[5] Fabian Caba Heilbron, B. G., Victor Escorcia, & Niebles, J. C. (2015). ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 961–970.

[6] Yang Feng et al. (2018). Video re-localizati on. Proceedings of European Conference on Computer Vision (ECCV), 51–66.

[7] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2018). FIVR: Fine-grained Incident Video Retrieval. CoRR, abs/1809.04094.http://arxiv.org/abs/1809.04094

[8] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, I. (2019). ViSiL: Fine-Grained Spatio-Temporal Video Similarity Learning. 6350–6359. https://doi.org/10.1109/ICCV .2019.00645

[9] ISO/IEC 15938-15:2019 Information technology - Multimedia content description interface - Part 15: Compact descriptors for video analysis, 2019.

[10] ISO/IEC 15938-13:2015 Information technology - Multimedia content description interface - Part 13: Compact descriptors for visual search, 2015.

[11] Lou, Y., Bai, Y., Lin, J., Wang, S., Chen, J., Chandrasekhar, V., Duan, L.-Y., Huang, T., Kot, A. C., & Gao, W. (2017). Compact Deep Invariant Descriptors for Video Retrieval. 2017 Data Compression Conference (DCC), 420–429. https://doi.org/10.1109/DCC.2017.31

[12] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., & Kompatsiaris, Y. (2017). Near-Duplicate Video Retrieval with Deep Metric Learning. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 347–356. https://doi.org/10.1109/ICCVW.2017.49

[13] Shao, J., Wen, X., Zhao, B., & Xue, X. (2021). Temporal Context Aggregation for Video Retrieval with Contrastive Learning. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 3267–3277. https://doi.org/10.1109/WACV48630.2021.00331