

영상문법&분석

문과생을 위한 AI 활용 파이썬 코딩_5주차

영상 구성의 단위

1. 프레임(frame):

- 프레임은 영상을 구성하는 가장 작은 단위로, 정지된 하나의 이미지를 의미.
- 영상은 초당 수십 개의 프레임(예: 24fps, 30fps 등)으로 이루어지며, 이 프레임들이 연속적으로 재생되면서 움직임이 있는 영상처럼 보이게 됨.

2. 컷(cut, shot):

- 컷 또는 샷은 카메라가 시작부터 끝까지 녹화한 연속적인 영상.
- 카메라가 꺼지지 않고 하나의 흐름을 유지하는 동안 기록된 하나의 연속적인 장면을 의미.
- 컷은 영상의 기본 구성 요소로, 편집자는 서로 다른 컷을 이어 붙여 스토리를 완성.
- 편집에서 “컷“은 한 샷에서 다른 샷으로 넘어가는 전환점을 뜻하기도 함.

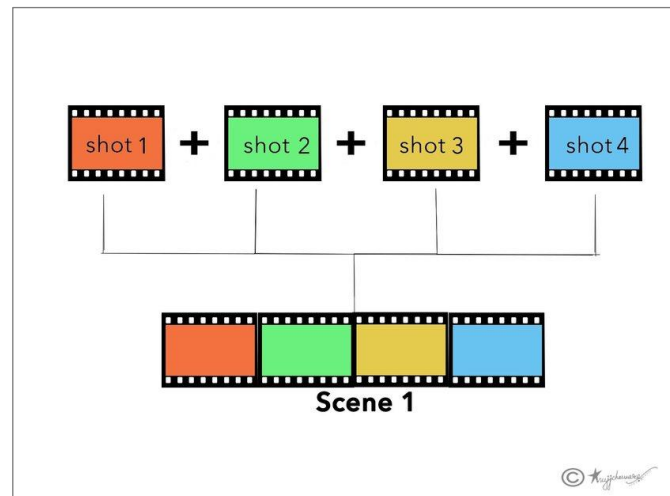
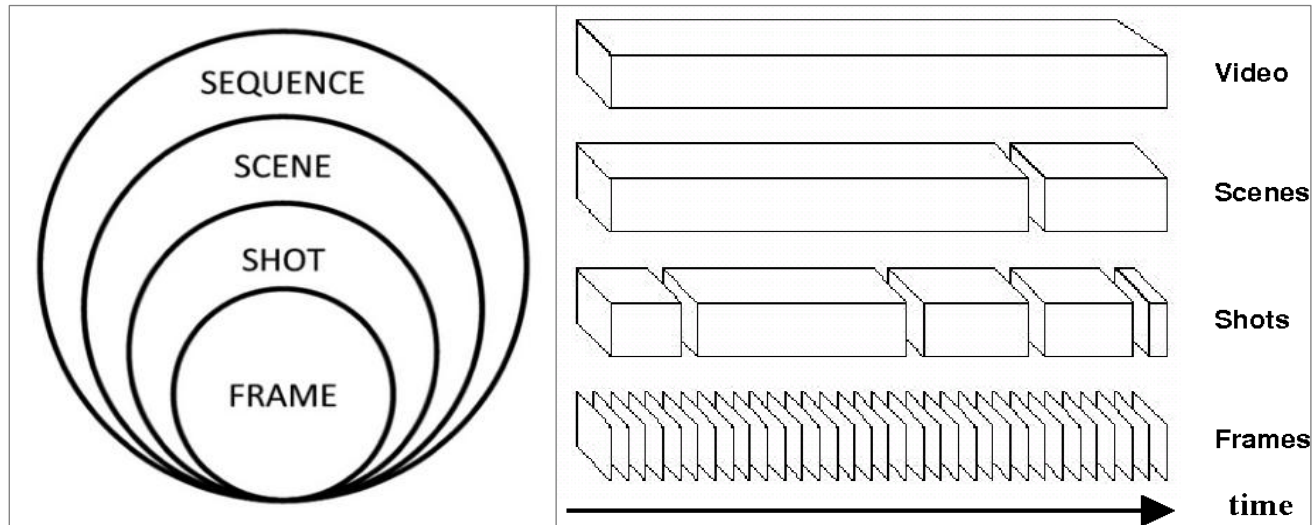
3. 시퀀스(Sequence):

- 시퀀스는 여러 컷을 모아 특정 사건이나 주제를 중심으로 연관된 장면을 구성한 단위.
- 여러 샷이 연속적으로 이어져 사건의 흐름을 보여주는 단위로, 시각적 및 이야기적 연속성을 지님.

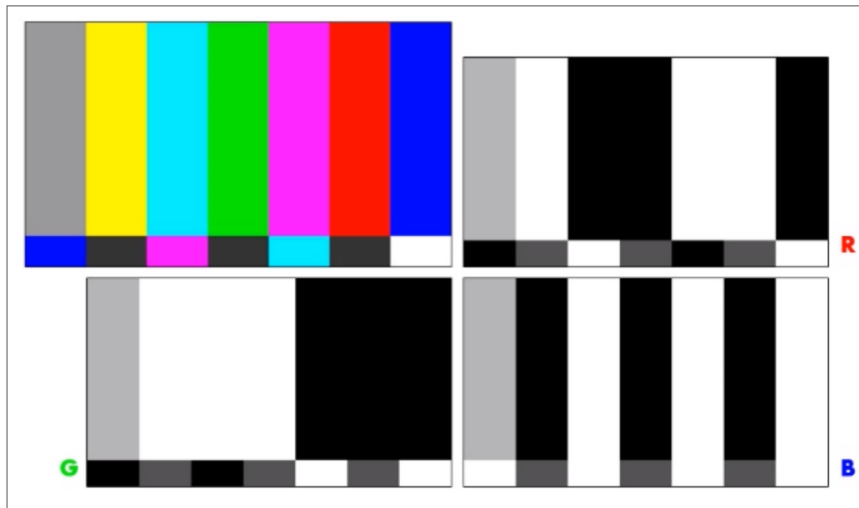
4. 기타 형식

- 신은 시퀀스보다 더 큰 단위로, 시간과 장소의 변화가 없는 특정 사건이나 상황을 묘사하는 장면.

Basic Principle of Video



Basic Principle of Video



R	G	B	색상	
0	0	0	검은색	
255	255	255	흰색	
255	0	0	빨간색	
0	255	0	초록색	
0	0	255	파란색	
255	255	0	노란색	
128	0	0	어두운 빨간색	
69	69	69	어두운 회색	



(a) 64 x 64 영상



(b) 128 x 128 영상



(c) 256 x 256 영상

Basic Principle of Video

아이레벨(Eye Level)



로우 앵글(Low Angle)



하이 앵글(High Angle)



버드 아이 뷰(Bird's Eye View)



Basic Principle of Video



Extreme Close-Up



Big Close-Up / Choker



Close-Up



Medium Close-Up



Medium / Mid-Shot



Medium Long Shot



Long / Wide Shot



Very Long / Wide Shot



Extreme Long / Wide Shot



2-Shot



Over-The-Shoulder

※ 이미지 출처: Thompson, R., & Bowen, C. J. (2009). Grammar of the Edit (Vol. 13). Taylor & Francis.

Video Shooting&Editing

영상 분석 예시

연구 대상의 수집과 선정

- 'K-POP MV'를 키워드로 유튜브에서 검색된 총 579개의 영상 중에서 (2024년 8월 22일 0시 30분 기준), 뮤직비디오와 관련 없는 영상을 제외한 총 527개의 영상을 최종 분석 대상으로 선정
- 본 연구에서 대상으로 삼은 K-POP 뮤직비디오 또한 한국 대중 음악 장르 전반을 망라하는 뮤직비디오 일반이 아니라 YouTube를 통해 국내외로 유통되는 특정 한 'K-POP 뮤직비디오'를 대상으로 하는 것이며, K-POP 뮤직비디오에 대한 전수 조사가 아닌 의미 있는 분석을 수행 가능한 정도의 데이터의 수집이 목적.

<표 1 > 분석대상 뮤직비디오 연도별/성별 건수(N=527)

뮤직비디오 수(건)																		
성별	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	총합
여성	1	2	1	5	2	6	9	11	8	15	24	26	34	31	58	32	28	293
남성	0	1	3	4	5	3	7	10	13	16	13	15	32	29	32	18	24	225
혼성	0	0	0	1	0	0	0	0	0	1	1	1	0	0	1	2	2	9
	1	3	4	10	7	9	16	21	21	32	38	42	66	60	91	52	54	527

영상 분석 예시

I-frame 추출

- 총 527개의 K-POP 뮤직비디오에서 파이썬 코드를 활용하여 총 85,554개의 I-frame을 추출.
- I-frame은 영상에서 장면 전환이나 큰 화면 변화 시 새로운 이미지 데이터를 저장하는 핵심 프레임으로(백지연 · 윤호영, 2021), 일정 수준 이상의 화면 변화가 있음을 나타냄.



<그림 3> 뮤직비디오에서 추출된 I-frame 예시(BTS 'Dynamite' M/V, N=147)

영상 분석 예시

Image Vector로 변환

- ViT-B/32 CLIP 모델을 사용해서 각각의 영상 이미지를 512차원 벡터로 변환.
- ViT-B/32는 Transformer 구조를 사용하여 이미지를 32x32 패치로 나누고, 이를 종합하여 고차원 벡터로 변환하는 역할을 수행할 수 있어서 대규모 이미지 데이터 처리에 적합한 CLIP 모델.
- 위 과정을 통해서 총 82,801개의 I-frame을 총 512차원 82,801개의 이미지 벡터로 변환
- CLIP 모델의 ViT-B/32는 Transformer 구조를 기반으로 이미지를 32x32 패치로 나누어 시퀀스 데이터를 생성한 뒤, 임베딩, 멀티 헤드 셀프 어텐션, 층 정규화, 잔차 연결 등 Vision Transformer 아키텍처의 과정을 통해 가중치를 연산하며, 최종적으로 다층퍼셉트론(MLP)을 통해 이미지 분류를 수행함.

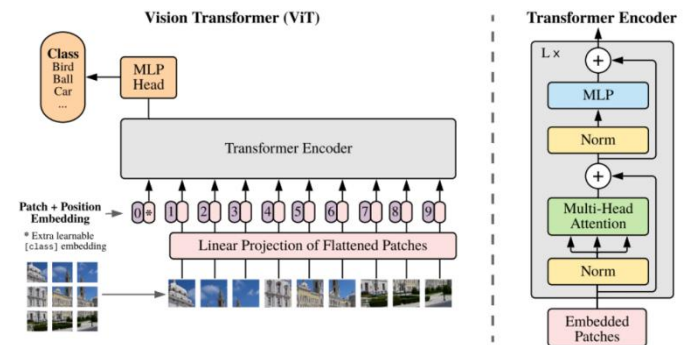
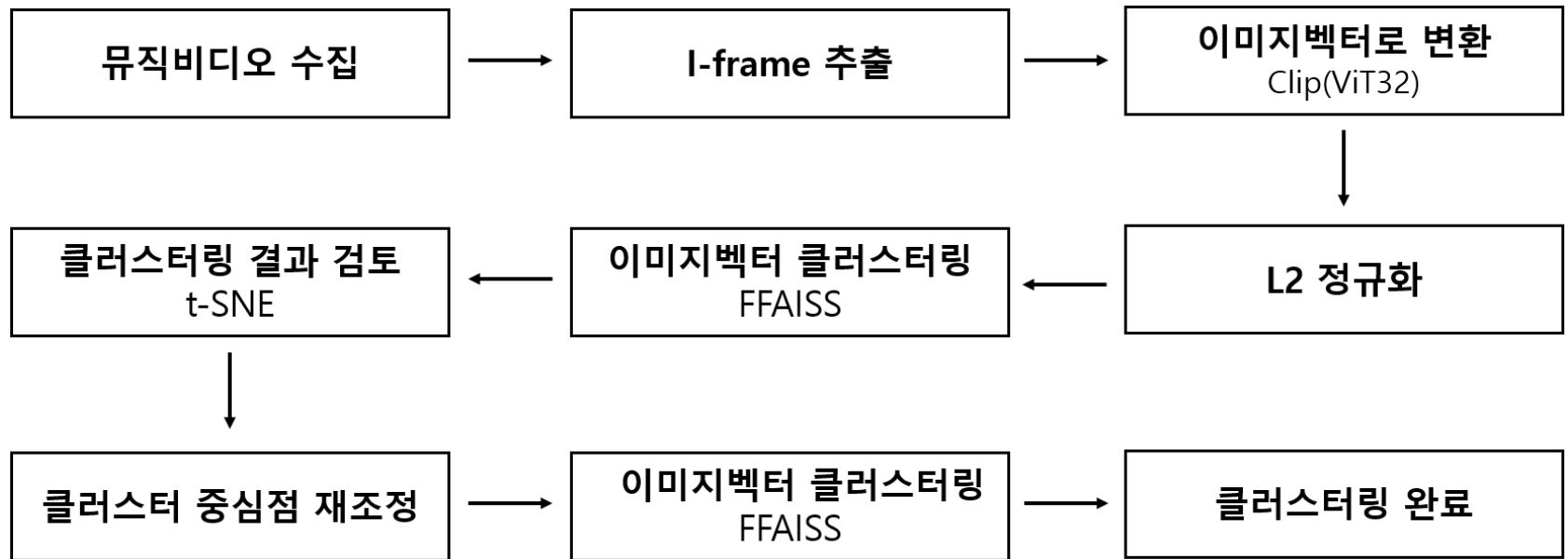


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

<그림> ViT(Vision Transformer)의 구조
(Dosovitskiy, et al., 2020, 3쪽)

영상 분석 예시



<그림> 이미지 벡터 클러스터링 과정

영상 분석 예시

4.1. K-POP 뮤직비디오 이미지 클러스터링

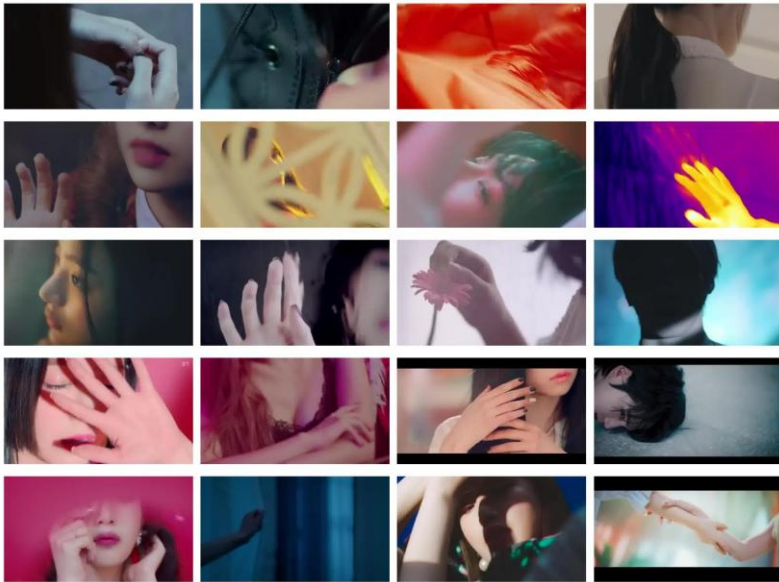
4.1.1. 전체 뮤직비디오 이미지 클러스터링 결과

- 각 클러스터는 샷 크기, 출연진 성별, 사물 표현 여부, 동작 지시 여부, 그리고 인물 중심 이미지인지 배경 중심 이미지인지에 따라 명명됨.

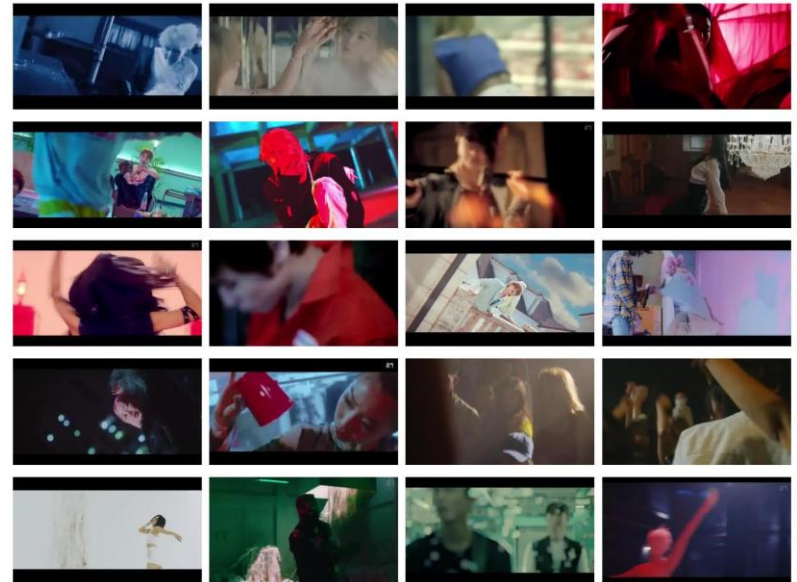
<표 2 > 전체 이미지 유형 분류 결과 (N=85,544)

			이미지 수(건)	
Cluster			Images	
Number	Name	Description	Count	Proportion
0	Body Image Shot	신체 이미지	4189	4.9%
1	Human Image Shot	인물 이미지	11177	13.1%
2	Female Dance Full Shot	여성 그룹 댄스 동작	7200	8.4%
3	Female Close-Up Shot	여성 가창과 표정	9049	10.6%
4	Object Image Shot	사물 이미지	7375	8.6%
5	Male Close-Up Shot	남성 가창과 표정	5567	6.5%
6	Female Bust Shot	여성 가창과 상반신 동작	8885	10.4%
7	Male Bust Shot	남성 가창과 상반신 동작	8693	10.2%
8	Mixed Image Shot	혼합 이미지	4487	5.2%
9	Male Dance Full Shot	남성 그룹 댄스 동작	5074	5.9%
10	Mixed Dance Long Full Shot	혼합 군무와 배경	5224	6.1%
11	Back Ground Image Shot	배경 이미지	4096	4.8%
12	Female Dance Long Full Shot	여성 댄스와 배경	4528	5.3%
			85544	100.0%

영상 분석 예시



<그림 10> Cluster 0: Body Image Shot/신체 이미지



<그림 11> Cluster 1: Human Image Shot/인물 이미지

영상 분석 예시

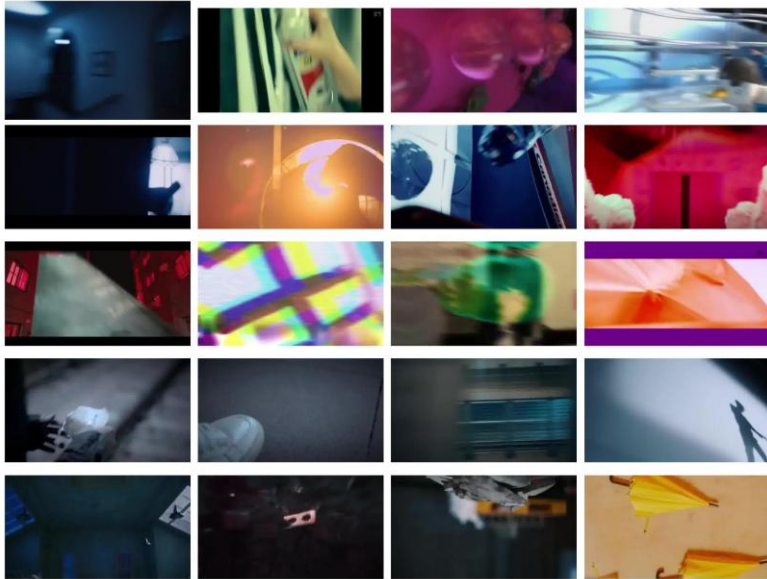


<그림 12> Cluster 2: Female Dance Full Shot
/여성 그룹 댄스 동작

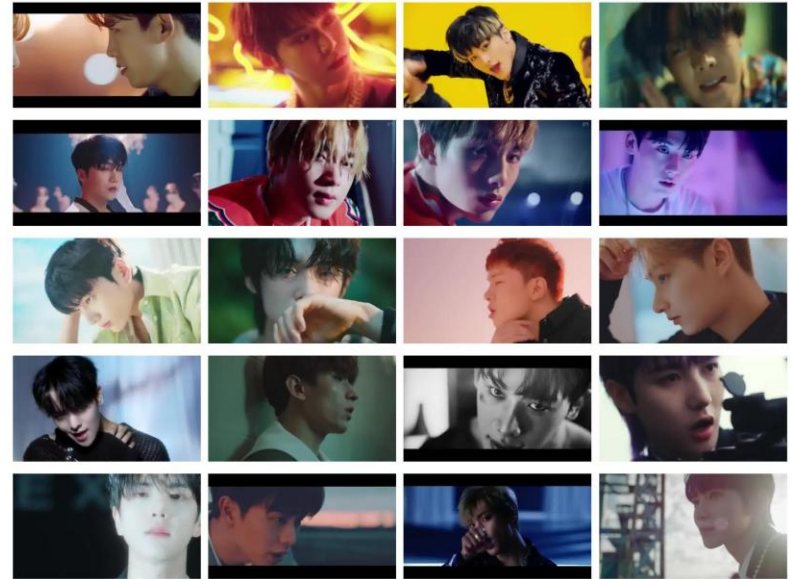


<그림 13> Cluster 3: Female Close-Up Shot
/여성 가창과 표정

영상 분석 예시

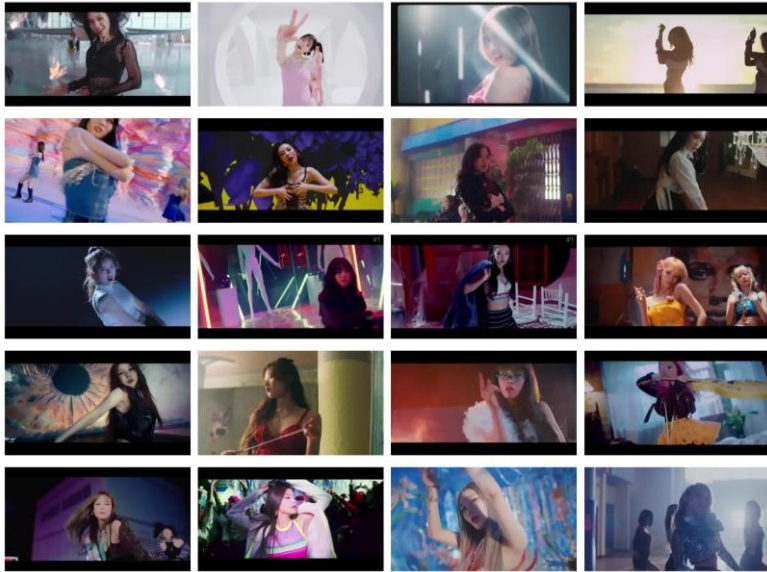


<그림 14> Cluster 4: Object Image Shot/사물 이미지

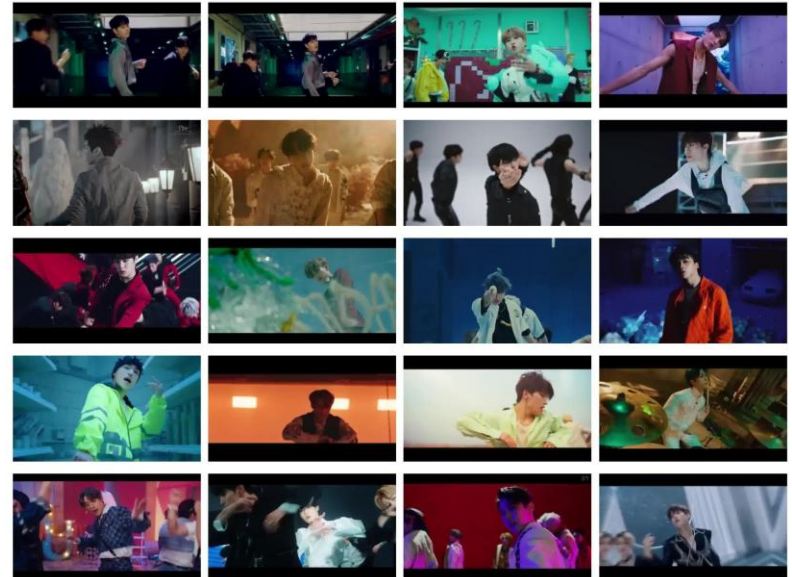


<그림 15> Cluster 5: Male Close-Up Shot
/남성 가창과 표정

영상 분석 예시

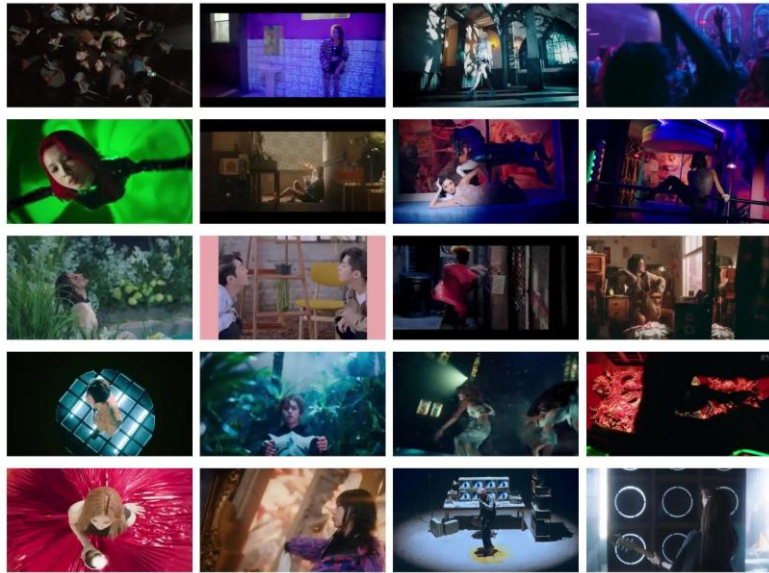


<그림 16> Cluster 6: Female Bust Shot
/여성 가창과 상반신 동작

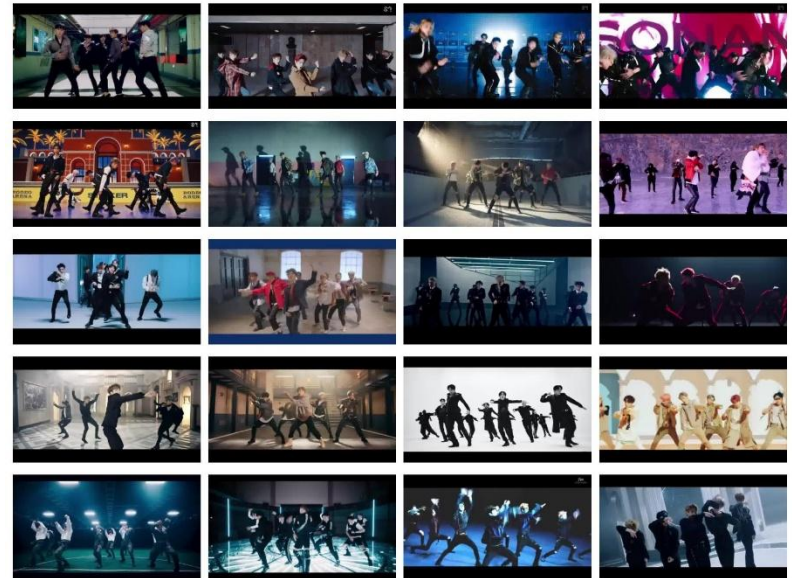


<그림 17> Cluster 7: Male Bust Shot
/남성 가창과 상반신 동작

영상 분석 예시

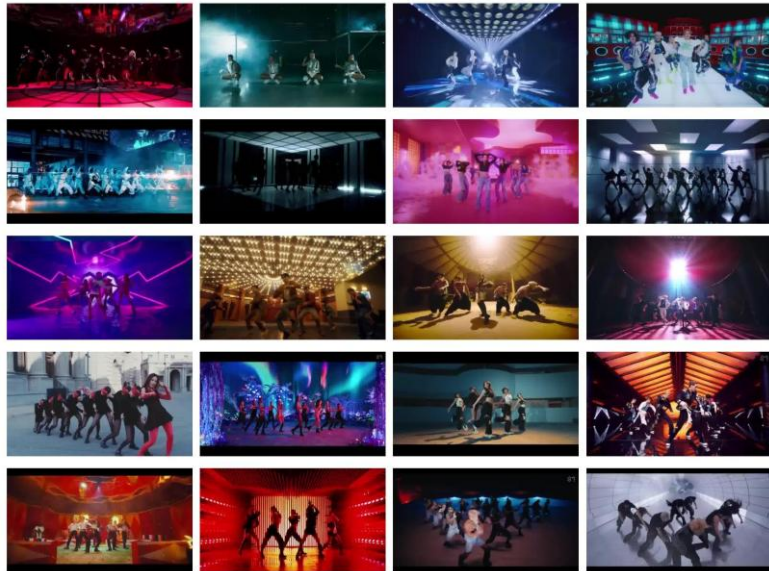


<그림 18> Cluster 8: Mixed Image Shot/혼합 이미지

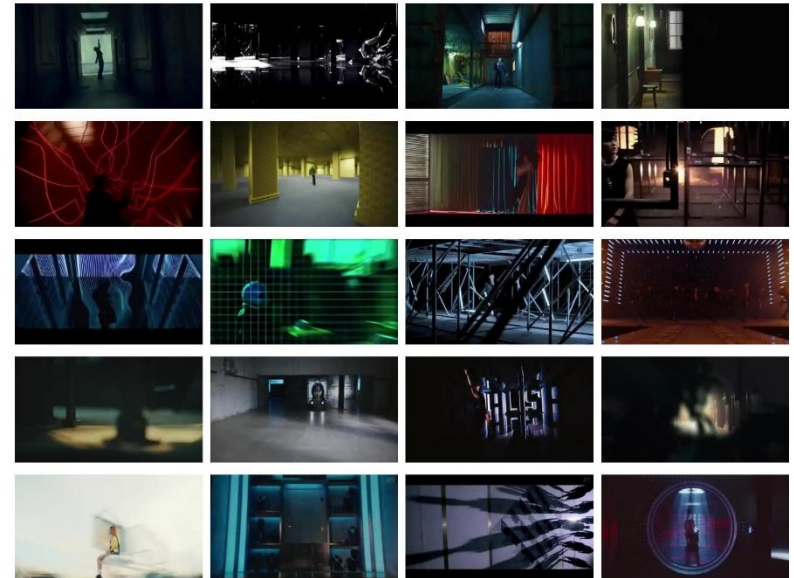


<그림 19> Cluster 9: Male Dance Full Shot
/남성 그룹 댄스 동작

4. 연구결과

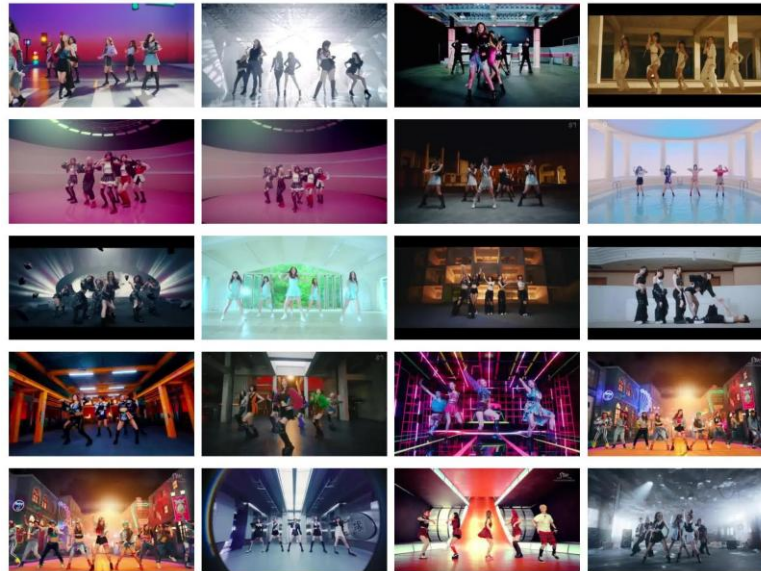


<그림 20> Cluster 10: Mixde Dance Long Full Shot
/혼합 군무와 배경



<그림 21> Cluster 11: Back Ground Image Shot
/배경 이미지

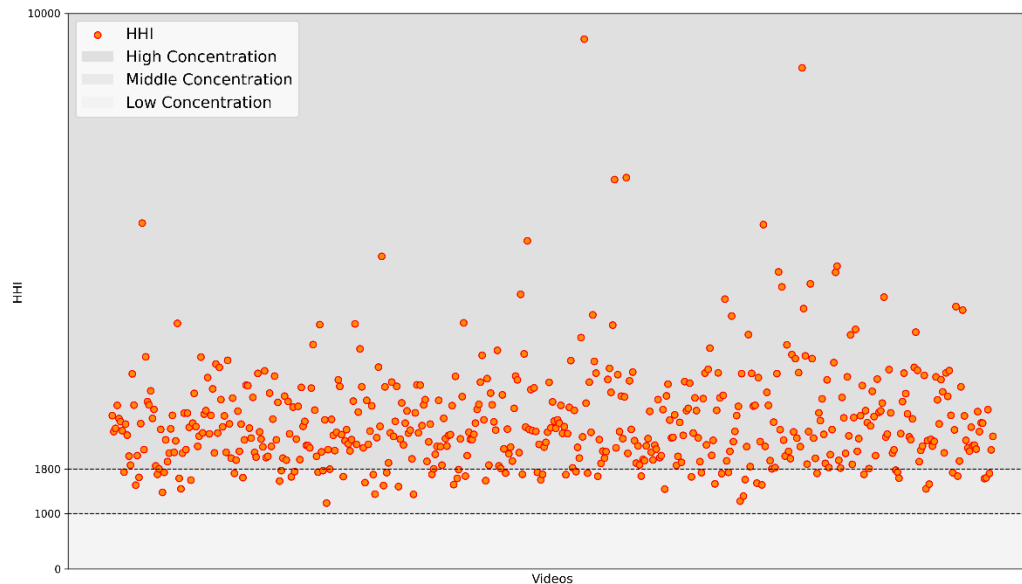
영상 분석 예시



<그림 22> Cluster 12: Female Dance Long Full Shot
/여성 댄스와 배경

영상 분석 예시

- K-POP 뮤직비디오의 HHI 지수 평균은 2678.20, 최대값은 9537.72, 최소값은 939.74로 나타나, 이미지의 다양성이 낮고 특정 유형에 집중되어 있음을 시사함. 대부분의 뮤직비디오가 HHI 지수 1800 이상을 기록하여 특정 이미지 유형에 집중된 경향이 있음을 보여줌.
- K-POP 뮤직비디오의 시각적 이미지 구성은 특정 유형에 집중되어 있음. 아티스트 외모 묘사와 군무에 대한 이미지로 집중도가 더욱 증가하였으며, 이는 소수 제작사가 시장을 지배할 경우 다양성이 저하될 수 있고(Litman, 1979; Rothenbuhler & Dimmick, 1982), 경쟁이 오히려 콘텐츠의 다양성을 제한하는 요인으로 작용할 수 있음을 시사함(Van der Wurff & Culienburg, 2001)..



<그림 30> 전체 뮤직비디오 HHI 지수 (N=527)

영상 분석 예시

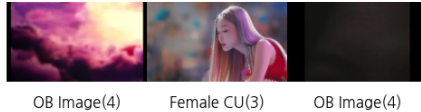
전체 뮤직비디오의 대표 이미지 패턴

- 그룹화된 전체 뮤직비디오에서 각 그룹별 대표 패턴을 도출함. 대표 패턴은 그룹 내 각 뮤직비디오의 이미지 등장 순서 중에서 가장 자주 반복되는 이미지 순서를 의미하며, 4개의 그룹별로 각각 가장 많이 반복된 패턴이 선정

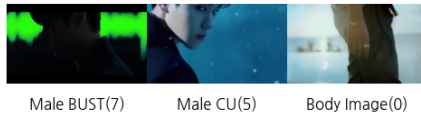
Video Cluster 1 (263) - "[남]사물-가창-군무-가창-군무"



Video Cluster 2 (101) - "[여]사물-가창-사물"



Video Cluster 3 (100) - "[남]가창-가창-신체"



Video Cluster 0 (63) - "[남]인물-가창-인물-군무-인물-가창-군무"



<그림 41> 전체 뮤직비디오의 그룹별 대표 이미지 패턴 (N=527)

Q&A