

영상 문법과 영상 데이터

영상은 이미지의 연속이고, 이미지는 숫자임.
그 두 문장 사이를 잇는 것이 오늘 하는 일.

10 오늘 쓰는 말	4 영상 구성의 단위	4 카메라 앵글	15 오늘의 절
----------------------	-----------------------	--------------------	--------------------

강의	일시	주제	담당
제5회 문과코딩 워크숍 · 5일차	2026. 8. 28.(금)	영상 문법과 영상 데이터	이문혁 (경희대 인문사회과학데이터연구소)

오늘 코드에 나오는 말

어제 쓰던 말에 넷이 더해짐. 코드를 읽다 막히는 자리가 대개 여기임.

API <code>application programming interface</code> 프로그램이 다른 프로그램에게 정해진 형식으로 물어보는 창구 . 어제는 화면을 받아서 뜯었으나, 오늘은 물어보면 값이 옴	API 키 <code>key</code> 그 창구에서 누가 물었는지를 알리는 문자열. 비밀번호와 같음. 각자 하나씩 발급받아 씀
할당량 <code>quota</code> 하루에 얼마까지 물어볼 수 있는지 . 유튜브는 하루 10,000 유닛. 다 쓰면 다음 날까지 기다려야 함	JSON <code>javascript object notation</code> 값이 돌아오는 형식 . 이름표와 값이 짝지어진 꾸러미라 태그를 뜯을 필요가 없음
인코딩 <code>encoding</code> 사람의 글자를 컴퓨터의 숫자로 바꾸는 규칙 . 보내는 쪽과 읽는 쪽이 다른 규칙을 쓰면 글자가 깨짐. 요즘 표준은 UTF-8	설치 <code>!pip install</code> 코랩에 없는 도구를 그 자리에서 받아 오는 것 . 느낌표로 시작하는 줄은 파이썬이 아니라 컴퓨터에게 직접 내리는 명령임

어제와 오늘의 차이는 여기 하나. 어제는 «화면을 받아 뜯는 일»이었고, 오늘은 «창구에 물어보는 일»임. 뜯을 것이 없으니 선택자도 필요 없음.

영상 파일을 이루는 말

영상 파일을 다루는 코드에서 나오는 말. 넷이면 충분함.

컨테이너 · 코덱 container, codec 컨테이너는 담는 상자(mp4·mkv), 코덱은 안의 내용을 압축한 방식(H.264 등). 확장자가 같아도 코덱이 다를 수 있음	해상도 resolution 가로 × 세로 화소 수. 1920×1080처럼 씬. 숫자가 클수록 표가 촘촘해지고 파일이 커짐
프레임레이트 fps 1초에 지나가는 프레임 수. 24fps면 1초에 정지 이미지 24장. 2분짜리 30fps 영상에는 프레임이 3,600장 있음	I-프레임 I-frame 화면이 크게 바뀌는 자리에 통째로 저장되는 프레임. 모든 프레임을 뽑으면 너무 많으므로 여기만 골라 뽑음

주. 어제 배운 말(라이브러리·임포트·객체·주석)은 오늘도 그대로 쓰임. 새로 나오는 것만 위에 적어 둬.

Part 01

영상의 기본 원리

영상이 무엇으로 이루어져 있는지, 그리고 그 이미지를 무엇이라 부르는지. 뒤에서 이미지가 유형으로 갈릴 때 그 유형의 이름이 여기서 나옴.

영상 구성의 단위

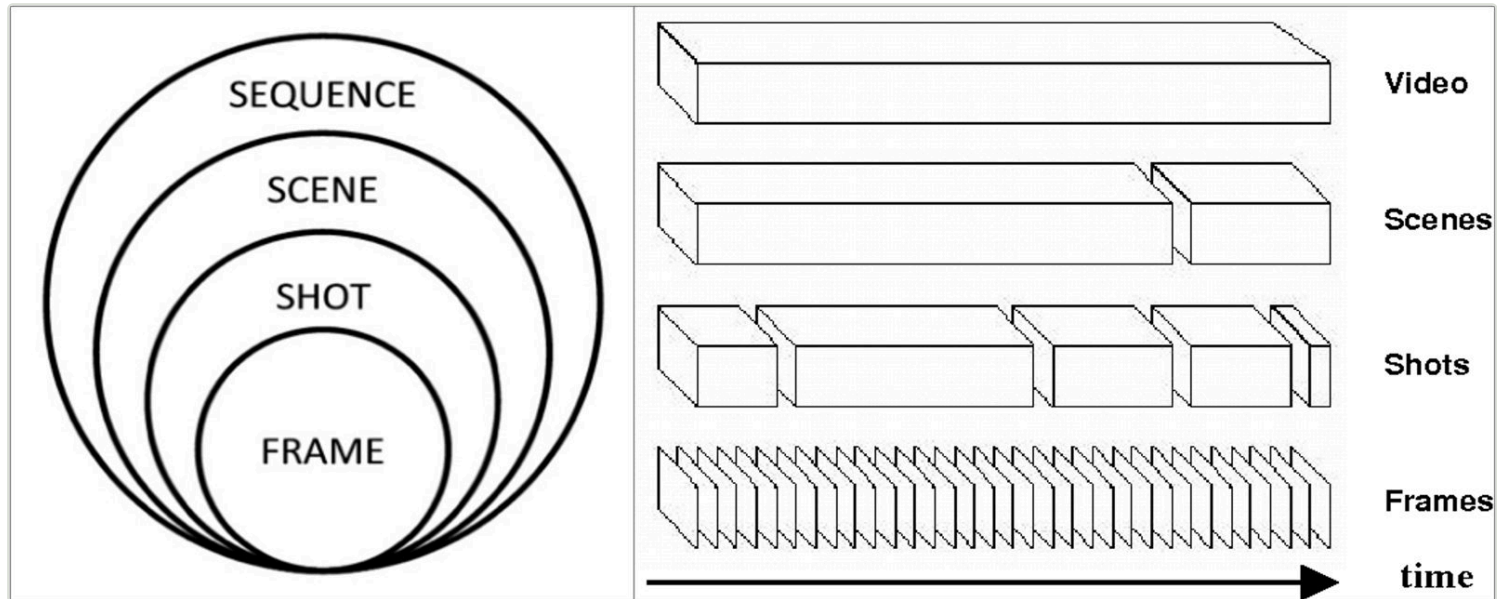
영상은 통째로 하나가 아니라 **작은 단위가 쌓여 만들어진 것**. 가장 작은 단위부터 올라가면 네 층이 나옴.

프레임 frame 영상을 구성하는 가장 작은 단위로 정지된 하나의 이미지. 영상은 초당 수십 개의 프레임 (24fps·30fps 등)으로 이루어지며, 이 프레임들이 연속적으로 재생되면서 움직임이 있는 영상처럼 보이게 됨	컷 · 샷 cut, shot 카메라가 시작부터 끝까지 녹화한 연속적인 영상. 카메라가 꺼지지 않고 하나의 흐름을 유지하는 동안 기록된 하나의 연속적인 장면. 편집에서 «컷»은 한 샷에서 다른 샷으로 넘어가는 전환점을 뜻하기도 함
신 scene 시간과 장소의 변화가 없는 특정 사건이나 상황을 묘사하는 장면. 여러 샷이 모여 하나의 신을 이룸	시퀀스 sequence 여러 신을 모아 특정 사건이나 주제를 중심으로 연관된 장면을 구성한 단위. 시각적·이야기적 연속성을 지님

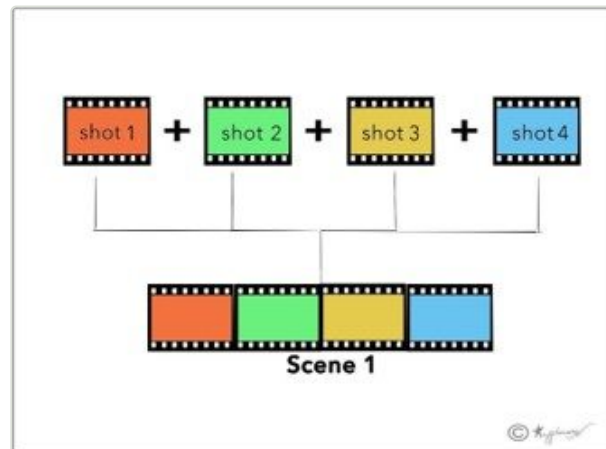
편집자는 서로 다른 컷을 이어 붙여 이야기를 완성함. 오늘 다루는 층은 그 아래, 프레임임.

프레임에서 시퀀스까지

같은 층위를 두 가지로 그린 것. 왼쪽은 포함 관계, 오른쪽은 그것을 시간축으로 편 모습.



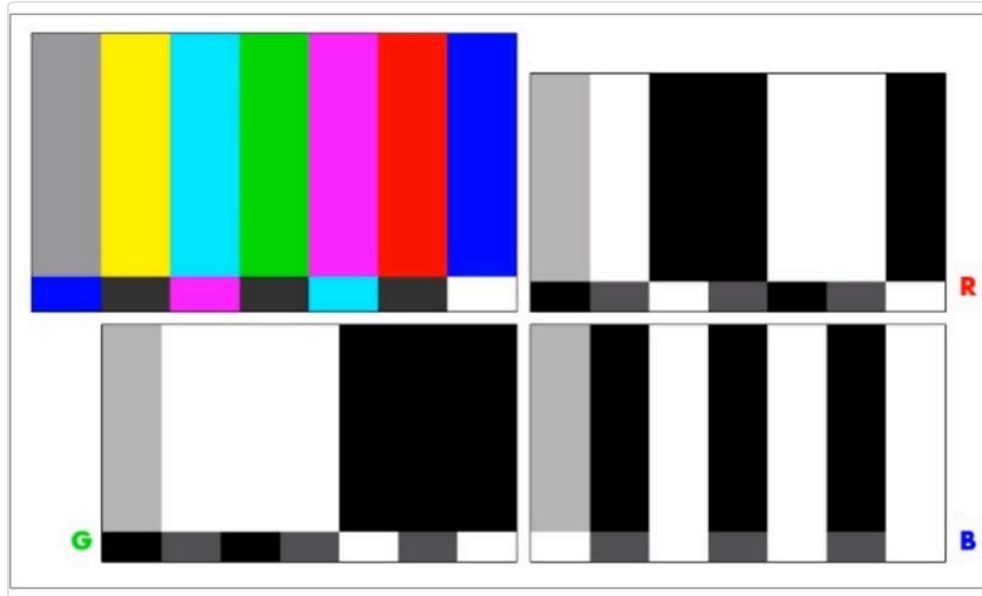
안쪽부터 프레임 · 샷 · 신 · 시퀀스. 오른쪽은 같은 층위를 시간 순으로 늘어놓은 것



샷 넷을 이어 붙여 신 하나를 만드는 모습

이미지는 숫자다

이미지는 작은 사각형(화소)의 격자이고, 화소마다 색이 숫자로 적혀 있음. 컴퓨터가 이미지를 «본다»는 것은 이 숫자 표를 읽는 일임.



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	

화소마다 빨강 · 초록 · 파랑의 세기를 0~255로 적어 둬. 세 숫자의 조합이 색 하나

주. 색이 없는 흑백 이미지는 화소마다 숫자가 하나뿐임. 아래 예시가 그런 경우.

I 화소 수를 바꾸면 — 같은 그림, 다른 해상도



(a) 64 × 64



(b) 128 × 128



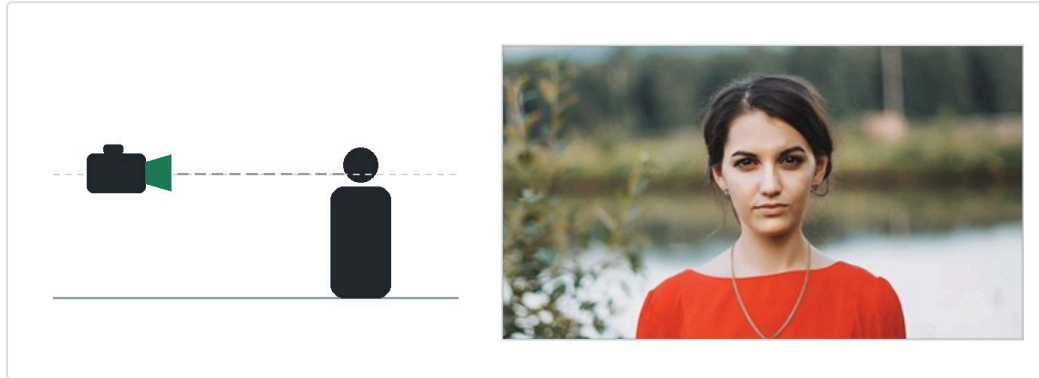
(c) 256 × 256

주. 세 장 모두 같은 장면이고 **화면에서 차지하는 크기도 같음**. 다른 것은 «몇 개의 화소로 나누어 담았는가»뿐임. 이미지 출처 — 배포 예제 영상 *Big Buck Bunny* (Blender Foundation, CC BY 3.0)

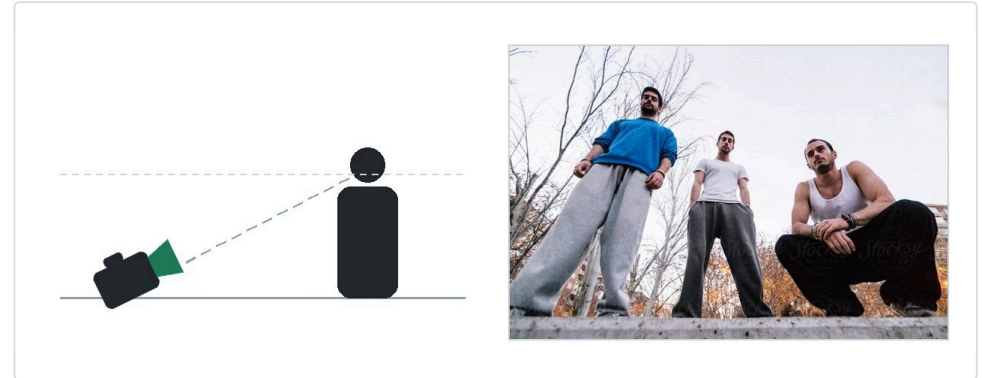
화소가 적으면 계단처럼 보임. 뒤에서 이미지를 숫자 줄로 바꿀 때 «몇 화소짜리로 맞출지»를 정하는데, 그것이 이 이야기임.

카메라 앵글 — 어디서 보는가

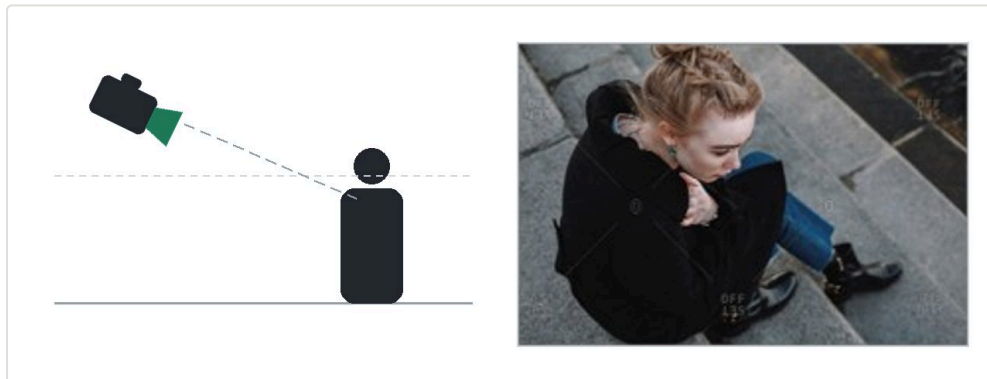
같은 인물을 찍어도 카메라가 놓인 높이에 따라 다른 인상이 만들어짐. 네 가지 이름을 알아 두면 충분함.



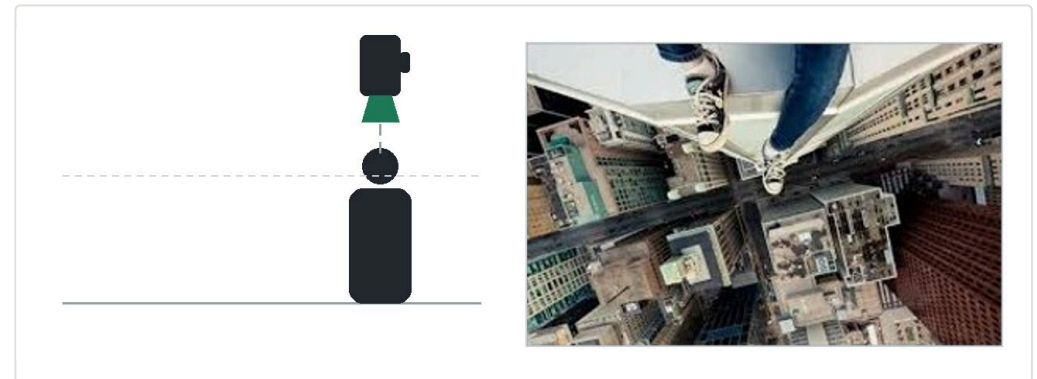
아이레벨(Eye Level) — 눈높이에서. 가장 중립적임



로우 앵글(Low Angle) — 아래에서 올려다봄. 대상이 커 보임



하이 앵글(High Angle) — 위에서 내려다봄. 대상이 작아 보임

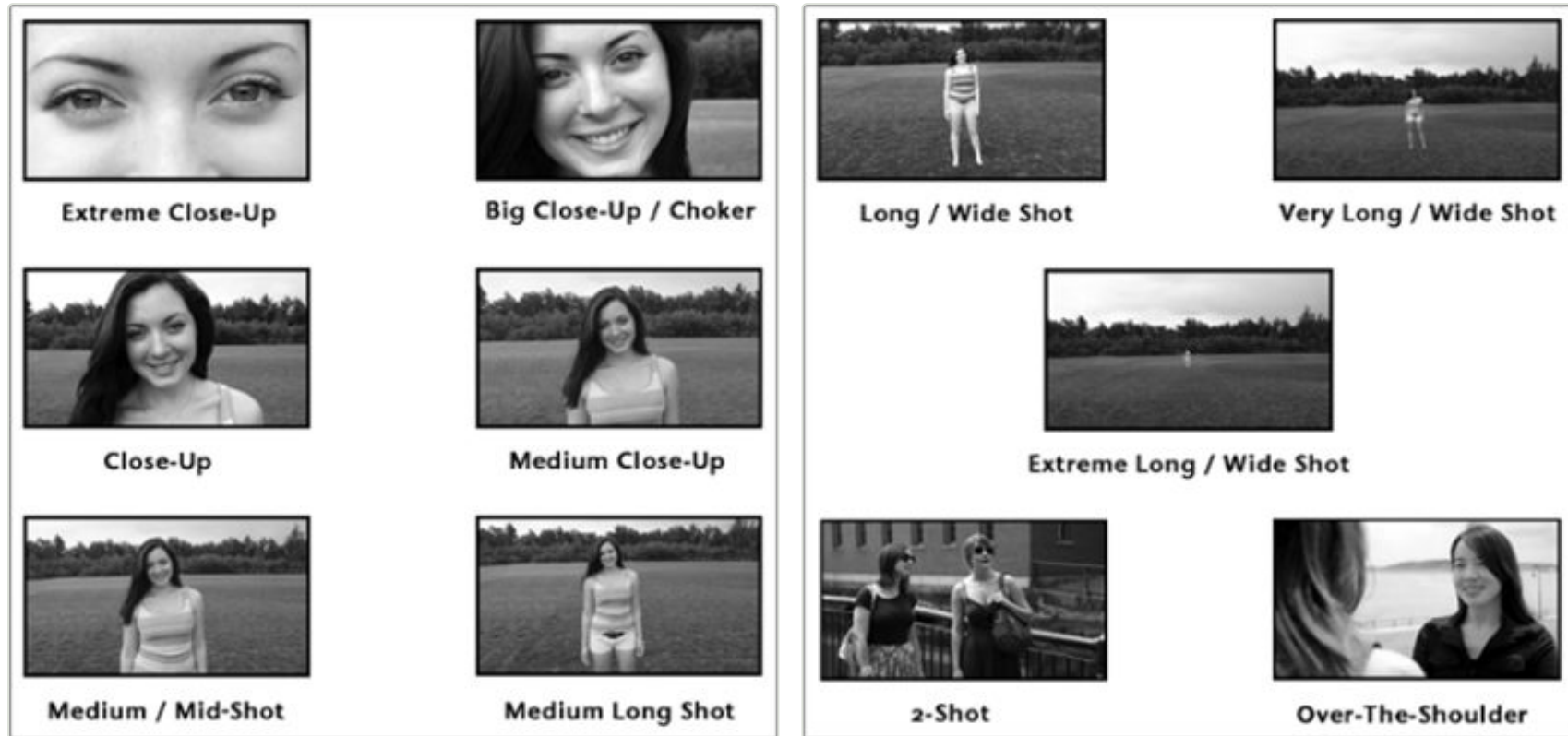


버드 아이 뷰(Bird's Eye View) — 거의 수직으로 내려다봄

주. 왼쪽이 카메라가 놓인 자리, 오른쪽이 그때 찍히는 화면임. 점선은 눈높이이고, 카메라가 그 선의 아래에 있으면 로우, 위에 있으면 하이임.

숏 크기 — 얼마나 가까이

카메라가 대상에서 얼마나 떨어져 있는지. 가까울수록 표정이, 멀수록 상황이 보임.



가까운 쪽 — 익스트림 클로즈업부터 미디엄 롱까지. 표정이 보임

먼 쪽 — 롱부터 익스트림 롱까지. 상황이 보임. 아래 둘은 인물 수로 나뉘는 이름

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

Part 02

영상 분석 예시

앞에서 배운 말이 실제 연구에서 어떻게 쓰이는지. K-POP 뮤직비디오 527편을 모아 이미지 유형을 갈라낸 사례임.

연구 대상의 수집과 선정

무엇을 얼마나 모았는지를 먼저 적음. 이것이 없으면 뒤의 숫자를 읽을 수 없음.

- «K-POP MV»를 키워드로 유튜브에서 검색된 총 **579개**의 영상 중에서(2024년 8월 22일 0시 30분 기준), 뮤직비디오와 관련 없는 영상을 제외한 총 **527개**의 영상을 최종 분석 대상으로 선정
- 여기서 대상으로 삼은 것은 한국 대중음악 장르 전반을 망라하는 뮤직비디오 일반이 아니라, **유튜브를 통해 국내외로 유통되는 K-POP 뮤직비디오**임
- 전수 조사가 아니라 **의미 있는 분석을 수행할 수 있는 정도의** 자료를 모으는 것이 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

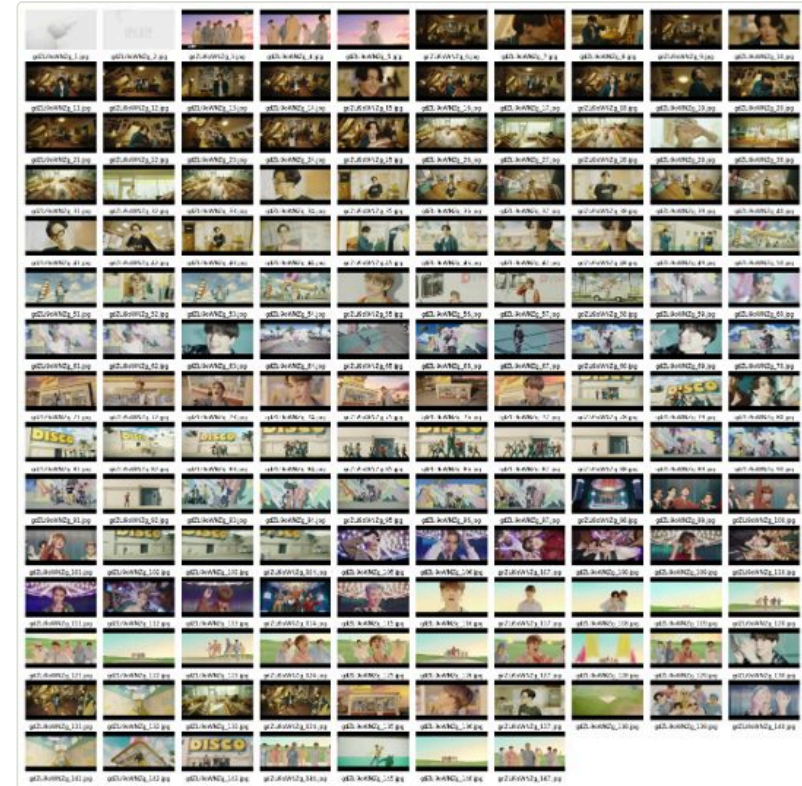
<표 1> 분석 대상 뮤직비디오의 연도별·성별 건수 (N=527) — 여성 293 · 남성 225 · 혼성 9

«어떻게 골랐는가»가 곧 연구의 범위임. 오늘 실습에서 검색어와 편수를 정하는 자리가 바로 이 단계에 해당함.

I-frame 추출

영상 527편을 그대로 볼 수는 없음. 이미지로 바꾸어 놓는 것이 첫 손질임.

- 총 527개의 K-POP 뮤직비디오에서 파이썬 코드를 활용하여 총 **85,544**개의 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 있을 나타냄
- 모든 프레임을 뽑으면 같은 장면이 수십 장씩 겹침. **I-frame만 뽑으면 «장면이 바뀐 자리»만 남음**



뮤직비디오 한 편에서 뽑은 I-frame 예시 (BTS 'Dynamite' M/V, N=147)

이미지를 숫자 줄로 바꾸기

이미지를 512개의 숫자로 바꿈. 그래야 «비슷하다»를 계산할 수 있음.

- ViT-B/32 CLIP 모델을 사용해서 각각의 영상 이미지를 512차원 벡터로 변환
- ViT-B/32는 Transformer 구조를 사용하여 이미지를 32×32 패치로 나누고, 이를 종합하여 고차원 벡터로 변환하는 역할을 수행할 수 있어서 대규모 이미지 데이터 처리에 적합한 CLIP 모델
- 패치로 나눈 뒤 임베딩 · 멀티 헤드 셀프어텐션 · 층 정규화 · 잔차 연결 등 Vision Transformer 아키텍처의 과정을 거쳐 가중치를 연산하며, 최종적으로 다층퍼셉트론(MLP)을 통해 이미지 분류를 수행함

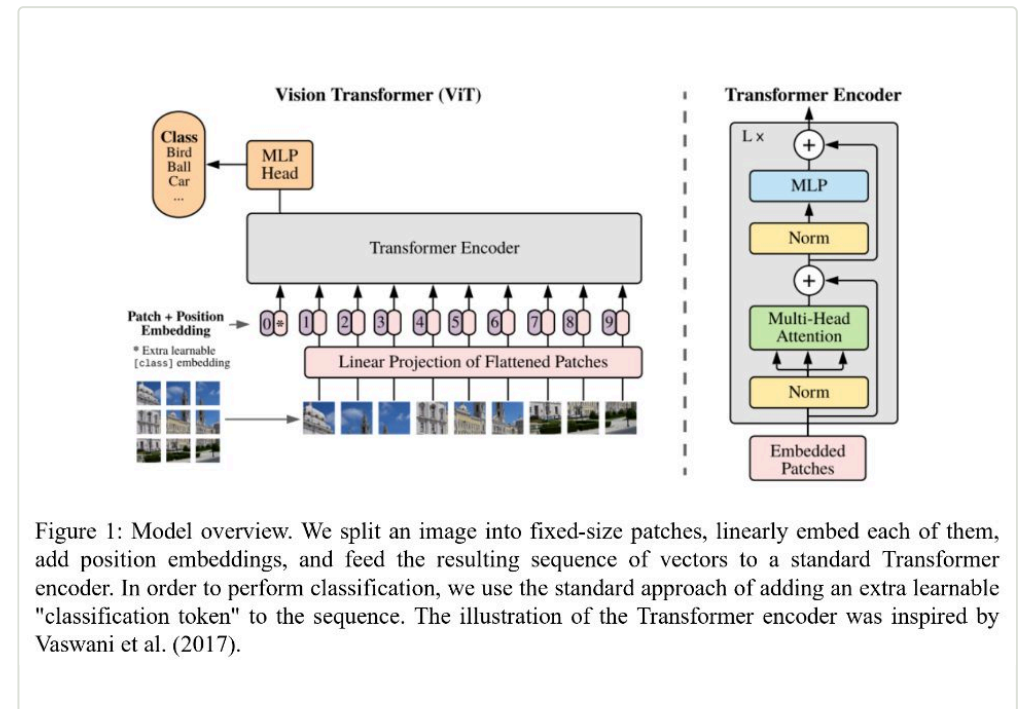


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification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

ViT(Vision Transformer)의 구조

Dosovitskiy, et al. (2020), 3쪽

I «512차원»이 무슨 뜻인가 – 자가 512개 있는 셈

사람을 숫자로 적는다고 하면 **키와 몸무게** 둘로 적을 수 있음. 그러면 사람 하나가 «숫자 두 개»가 되고, 이것을 **2차원**이라 부름. 이미지를 512개의 숫자로 적으면 512차원임.

01

자를 사람이 정하지 않음

«밝기»·«얼굴 크기» 같은 자를 사람이 고르는 것이 아님.
수많은 이미지를 미리 본 모델이 스스로 만들어 둔 자를 씀

02

가까우면 비슷한 것

두 이미지의 숫자 줄이 **가까우면** 비슷한 이미지임. «비슷하다»가 그제야 **계산할 수 있는** 말이 됨

03

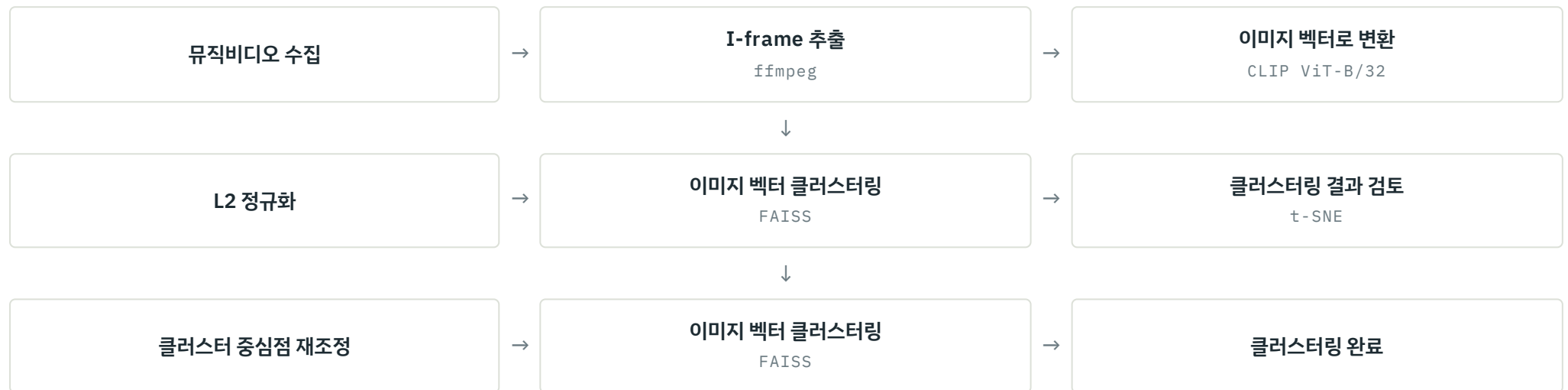
사람은 못 읽음

512개 숫자를 눈으로 읽어 뜻을 알 수는 없음. **서로 비교할 때만** 쓸모가 있음

여기가 오늘 강의의 갈림길임. 이미지를 숫자로 바꾸기 전에는 «비슷한 영상 찾기»가 사람 눈의 일이었고, 바꾼 뒤에는 **계산의 일**이 됨.

이미지 벡터 클러스터링 과정

앞의 두 단계를 포함한 전체 흐름. 오늘 실습에서 하는 것은 앞의 두 칸임.



주. 원본 도판에 FFAISS·Clip(ViT32)로 잘못 적힌 것을 FAISS·CLIP(ViT-B/32)로 바로잡아 다시 그림.

오늘 코드로 하는 것은 앞의 두 칸 — «수집»과 «I-frame 추출»까지. 그 뒤는 같은 자료를 놓고 무엇을 물을지의 문제임.

이미지 유형이 갈린 결과

85,544개의 이미지를 비슷한 것끼리 묶었더니 **13가지 유형**으로 갈림.

- 각 클러스터는 **샷 크기, 출연진 성별, 사물 표현 여부, 동작 지시 여부**, 그리고 인물 중심 이미지인지 배경 중심 이미지인지에 따라 명명됨
- 이름에 쓰인 말이 앞에서 본 **샷 크기·앵글** 그대로임 — 클로즈업 · 버스트 샷 · 풀 샷 · 롱 풀 샷

<표 2 > 전체 이미지 유형 분류 결과 (N=85,544)

Cluster			이미지 수(건)	
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%

13가지 유형과 각 유형에 속한 이미지 수

I «뭉는다»는 것은 무슨 일인가

01 답을 미리 주지 않음

«이건 클로즈업, 저건 롱 샷»이라고 가르쳐 주지 않음. 숫자 줄이 가까운 것끼리 모으라고만 시킴

02 몇 덩어리로 나눌지는 사람이 정함

기계가 «13개»를 알아서 고르는 것이 아님. 여러 수로 나뉘 본 뒤 사람이 읽어 낼 수 있는 수를 고름

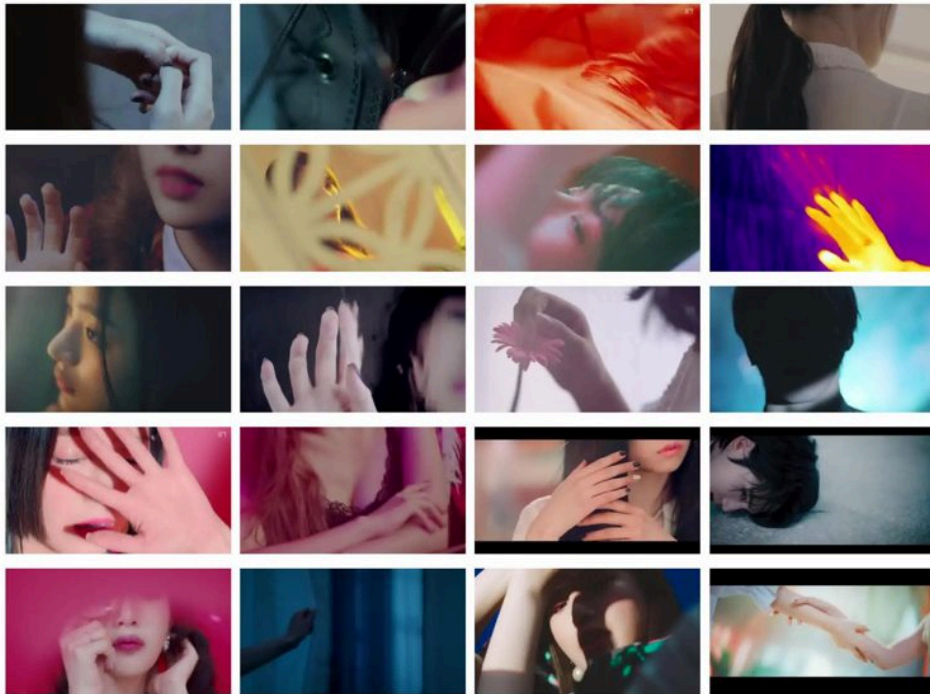
03 이름은 사람이 붙임

뭉인 덩어리를 열어 보고 «여성 가창과 표정»이라고 이름을 붙이는 것은 사람임. 기계는 «0번 덩어리»라고만 부름

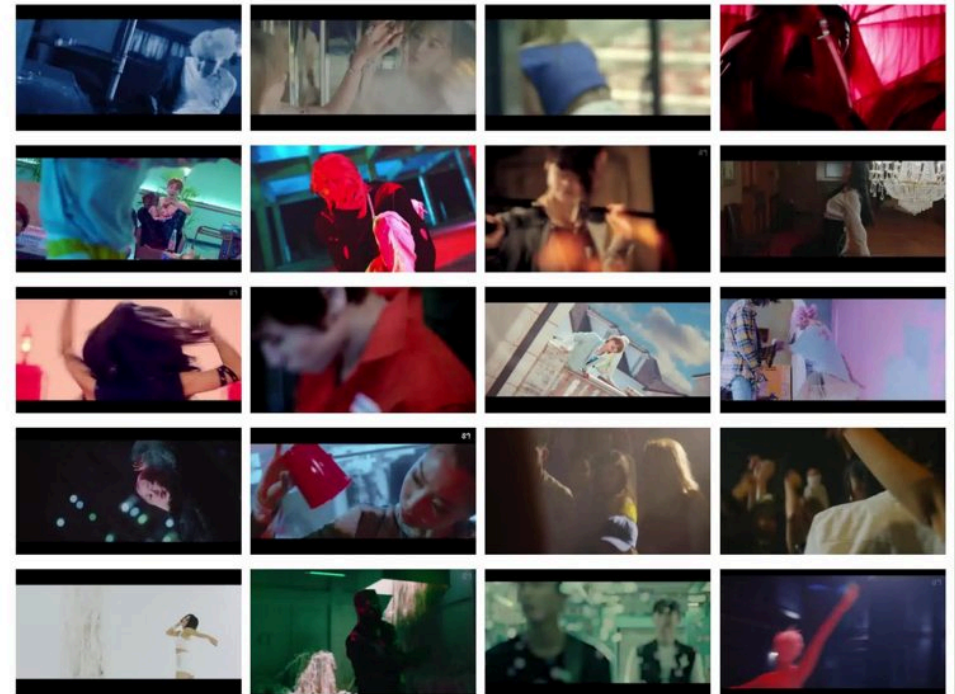
그래서 결과를 그대로 믿으면 안 됨. 뭉는 것은 기계가 하고 읽는 것은 사람이 함. 앞 쪽에서 배운 «샷 크기·앵글»이 여기서 읽어 내는 말이 됨.

이미지 유형 13가지

묵인 결과를 사람이 눈으로 확인하고 이름을 붙인 것. 기계가 이름을 지어 주지는 않음.



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



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

왼쪽은 신체 이미지, 오른쪽은 인물 이미지 — 화면을 채우는 것이 무엇인가로 갈림

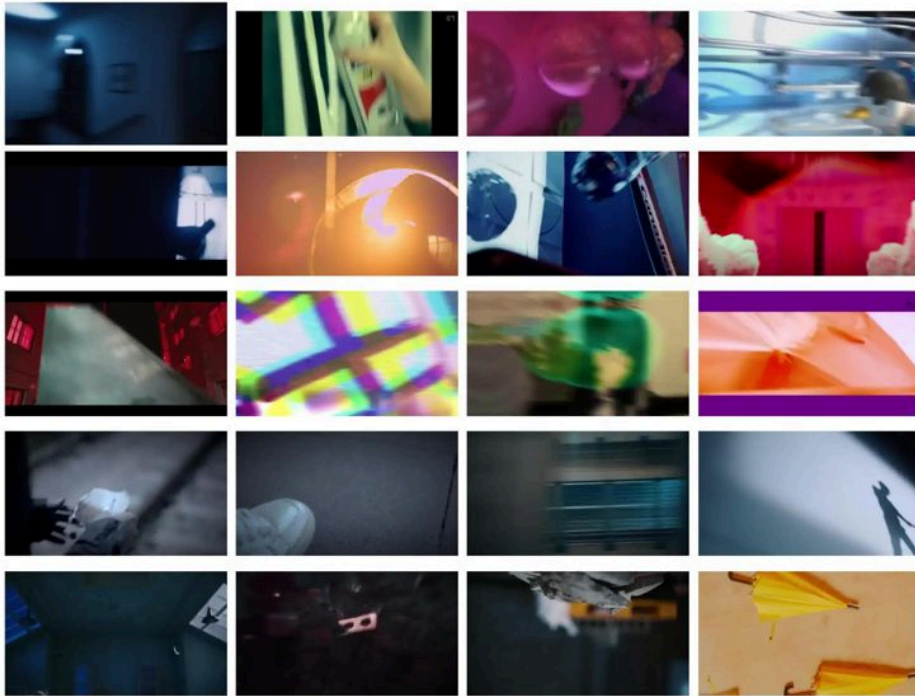


<그림 12> Cluster 2: Female Dance Full Shot
/여성 그룹 댄스 동작

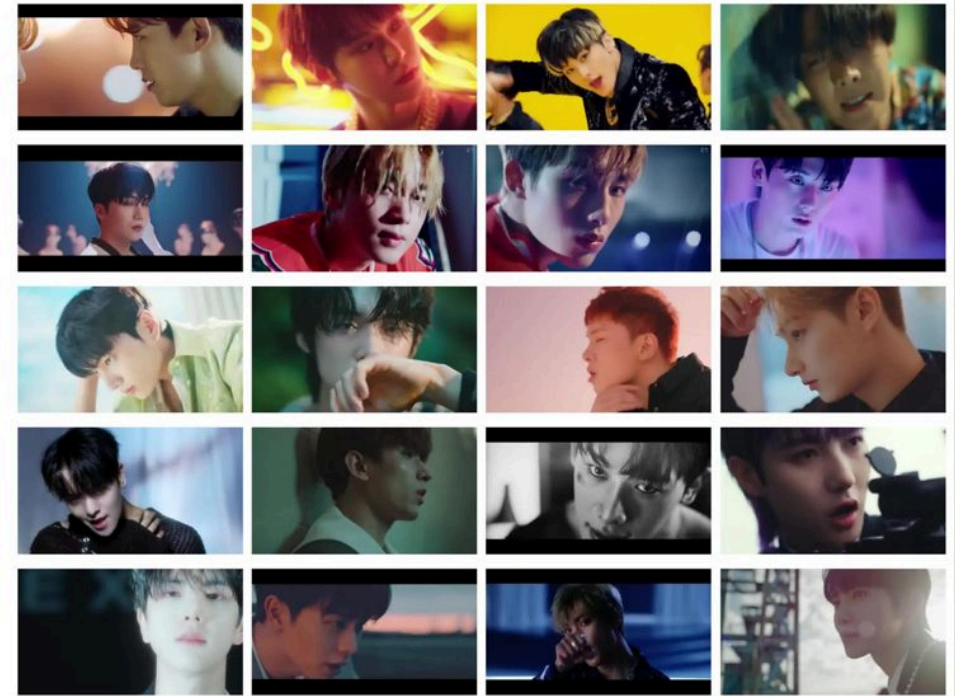


<그림 13> Cluster 3: Female Close-Up Shot
/여성 가창과 표정

여성 그룹 댄스는 풀 샷, 여성 가창은 클로즈업 — 샷 크기가 유형을 가름

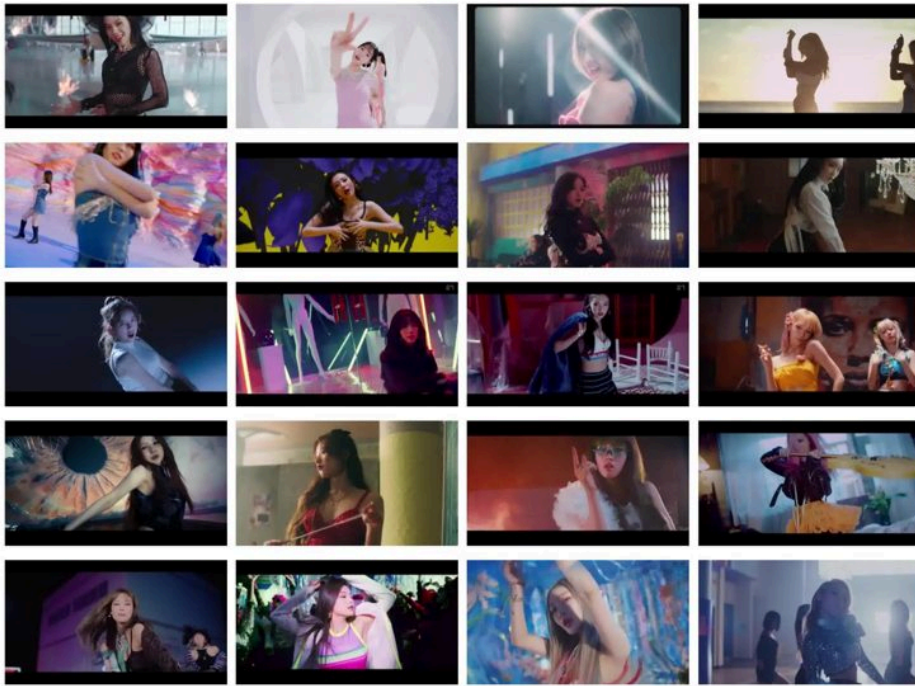


<그림 14> Cluster 4: Object Image Shot/사물 이미지

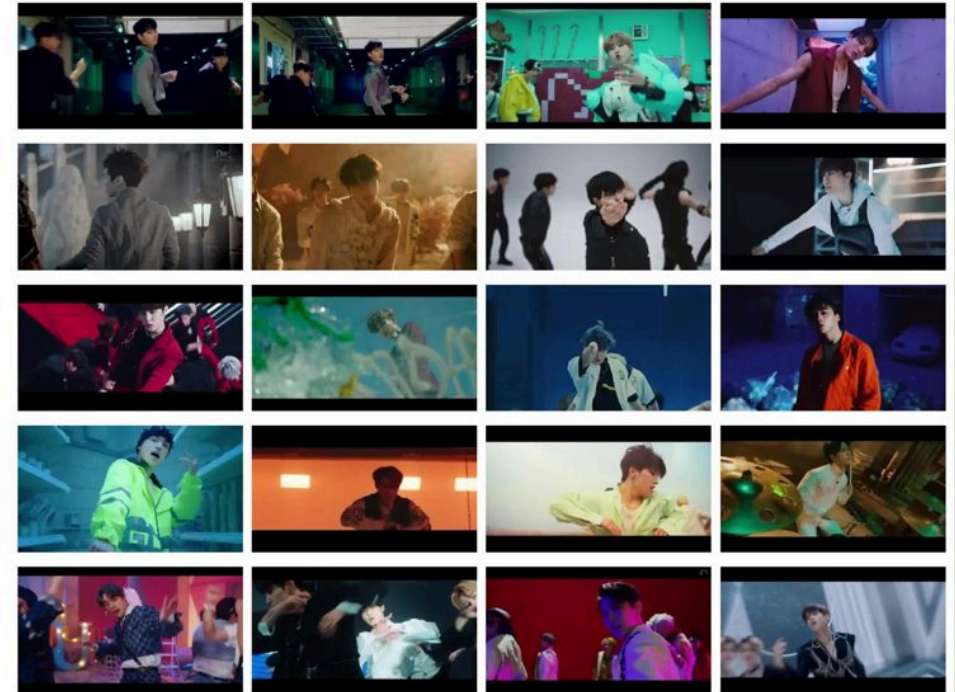


<그림 15> Cluster 5: Male Close-Up Shot
/남성 가창과 표정

사물 이미지와 남성 가창 클로즈업 — 인물이 없는 유형도 하나의 묶음이 됨

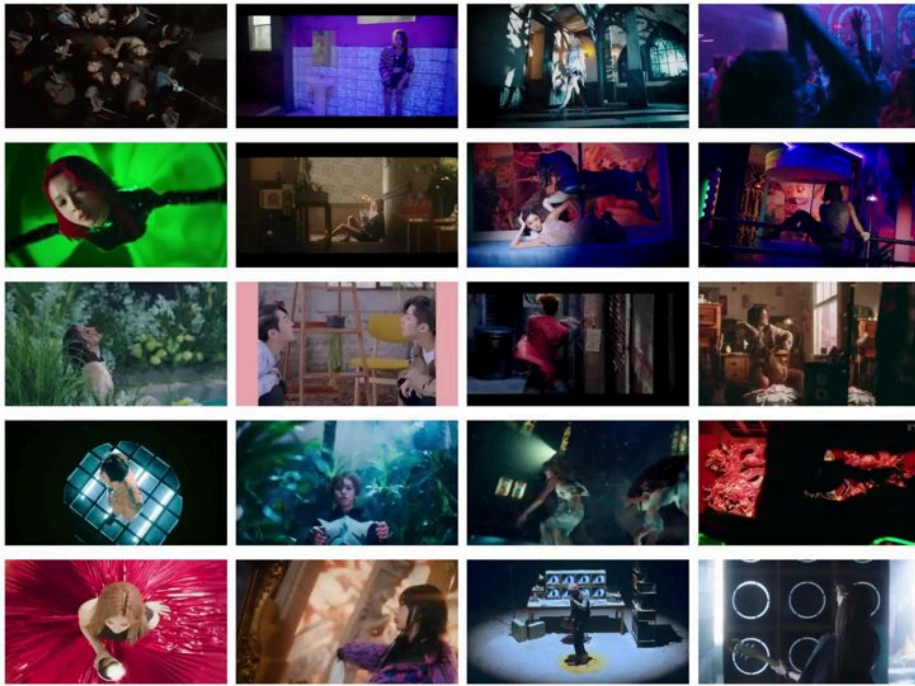


<그림 16> Cluster 6: Female Bust Shot
/여성 가창과 상반신 동작

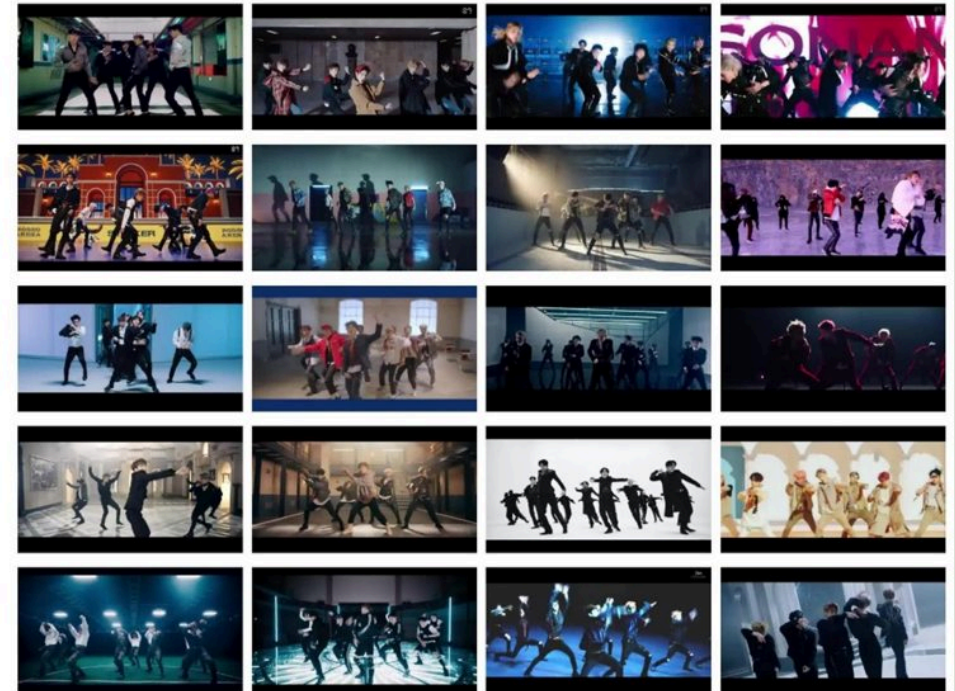


<그림 17> Cluster 7: Male Bust Shot
/남성 가창과 상반신 동작

여성·남성 상반신(버스트 샷) — 같은 샷 크기가 성별로 갈린 자리

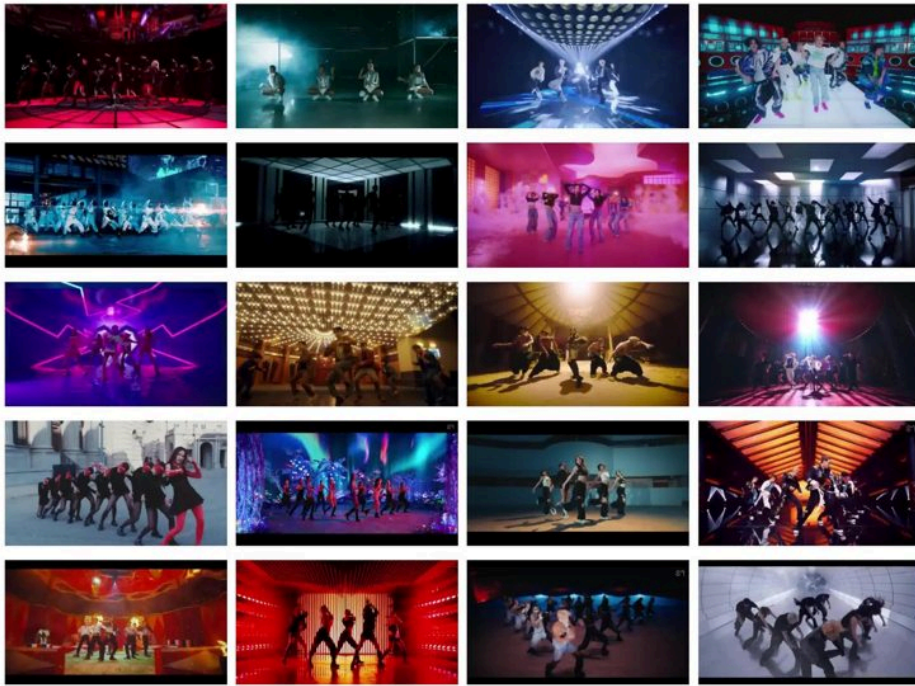


<그림 18> Cluster 8: Mixed Image Shot/혼합 이미지

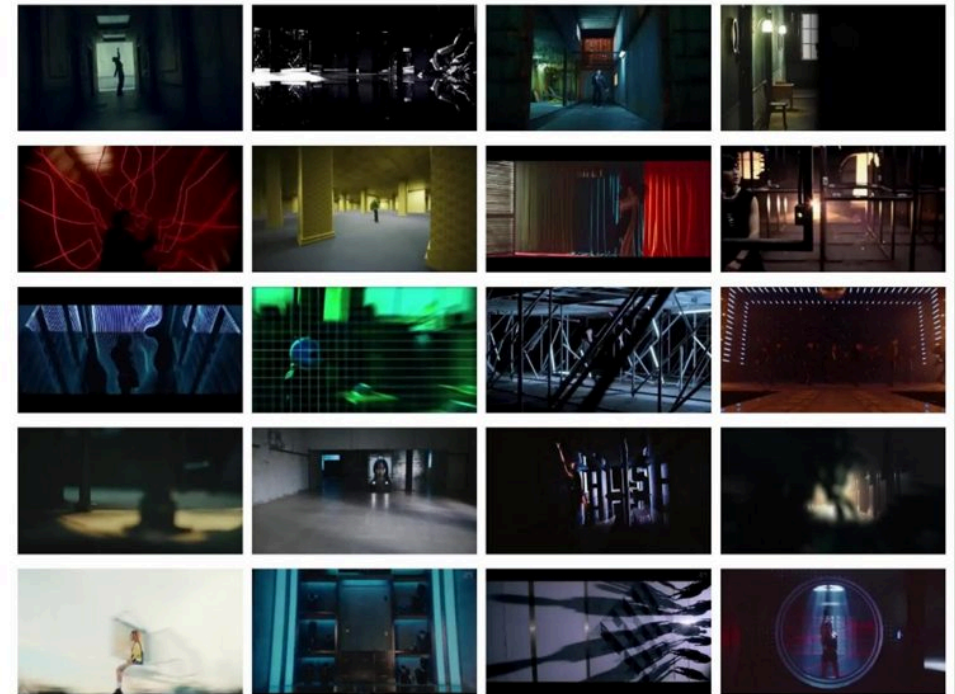


<그림 19> Cluster 9: Male Dance Full Shot
/남성 그룹 댄스 동작

혼합 이미지와 남성 그룹 댄스 — 여럿이 섞이면 별도 유형으로 묶임

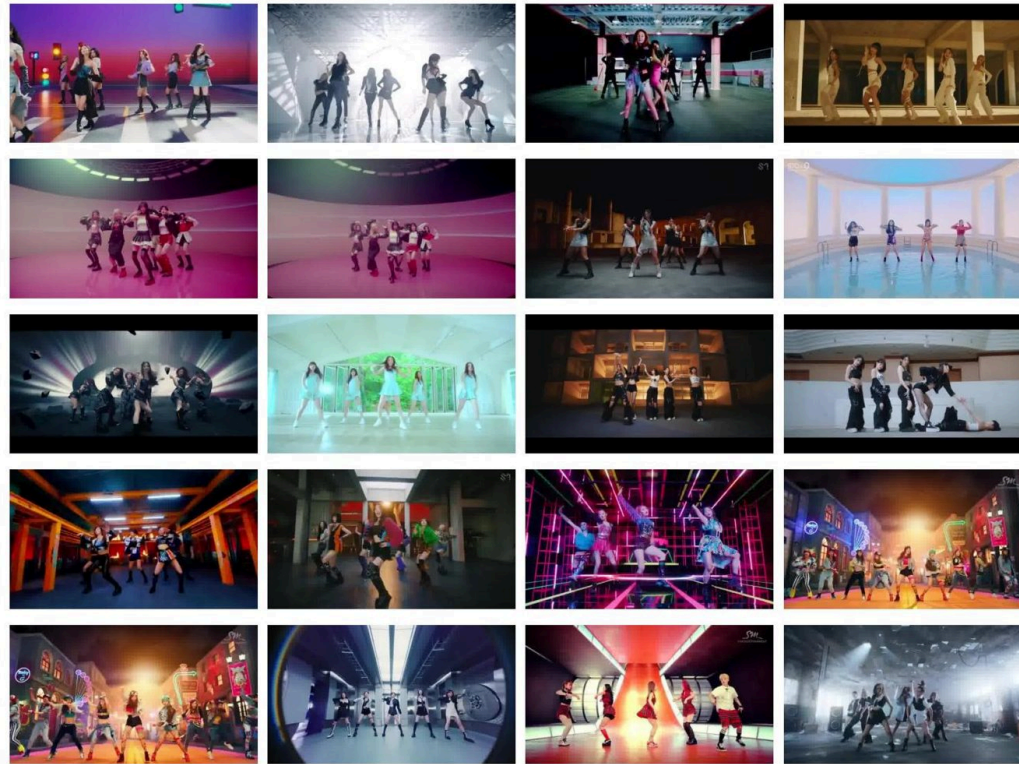


<그림 20> Cluster 10: Mixde Dance Long Full Shot
/혼합 군무와 배경



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

혼합 군무와 배경 — 인물이 작아지고 배경이 커지는 쪽



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

여성 댄스와 배경 — 인물이 가장 작게 잡히는 롱 풀 샷

유형의 이름을 다시 보면 — 클로즈업 · 버스트 · 풀 샷 · 롱 풀 샷. 07절에서 본 샷 크기가 그대로 분석의 언어가 됨.

얼마나 쏠려 있는가 — HHI

유형이 13가지로 갈렸어도, 한 편 안에서 몇 가지가 주로 쓰이는지는 또 다른 물음임.

I HHI는 원래 «시장이 한쪽으로 쏠렸는가»를 재는 자임

공정거래에서 «몇몇 회사가 시장을 다 가졌는가»를 잴 때 쓰던 지수임. 여기서는 회사 대신 이미지 유형을 넣음 — 한 영상이 몇 가지 유형을 고루 쓰는가, 한두 가지에 몰리는가.

어떻게 재나 계산

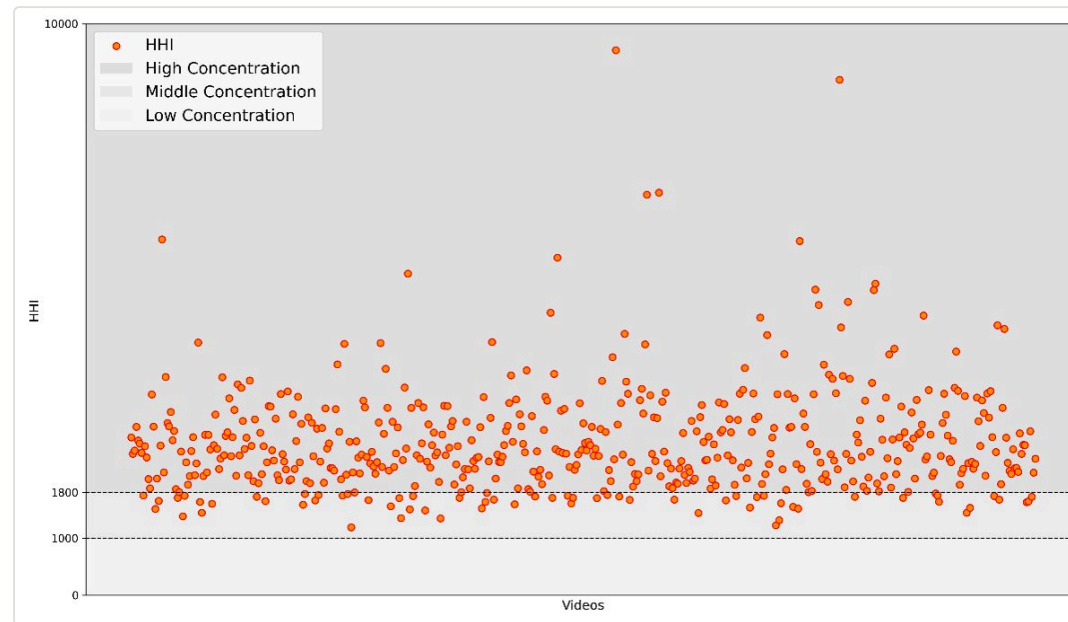
영상 한 편에서 각 유형이 차지하는 비율(%)을 **제공해서 모두 더함**. 한쪽에 몰릴수록 제공이 커지므로 값이 커짐

값의 폭 0 ~ 10,000

한 유형만 쓰면 $100^2=10,000$ (가장 쏠림). 열 가지를 고루 쓰면 $10 \times 10^2=1,000$ (가장 고루)

주. 이 연구는 **1,800**을 «쏠려 있다»고 보는 기준선으로 삼음.

- K-POP 뮤직비디오의 HHI 지수 평균은 **2,678**, 최대값은 9,537.72, 최소값은 939.74로 나타나, 이미지 유형의 다양성이 낮고 **특정 유형에 집중되어** 있음을 시사. 대부분의 뮤직비디오가 HHI 지수 1,800 이상을 기록
- 시각적 이미지 구성이 특정 유형에 집중되어 있으며, **아티스트 외모 묘사와 군무**에 대한 이미지로 집중도가 더욱 증가하였음. 이는 소수 제작사가 시장을 지배할 경우 다양성이 저하될 수 있고(Litman, 1979; Rothenbuhler & Dimmick, 1982), 경쟁이 오히려 콘텐츠의 다양성을 제한하는 요인으로 작용할 수 있음을 시사(Van der Wurff & Cuilenburg, 2001)

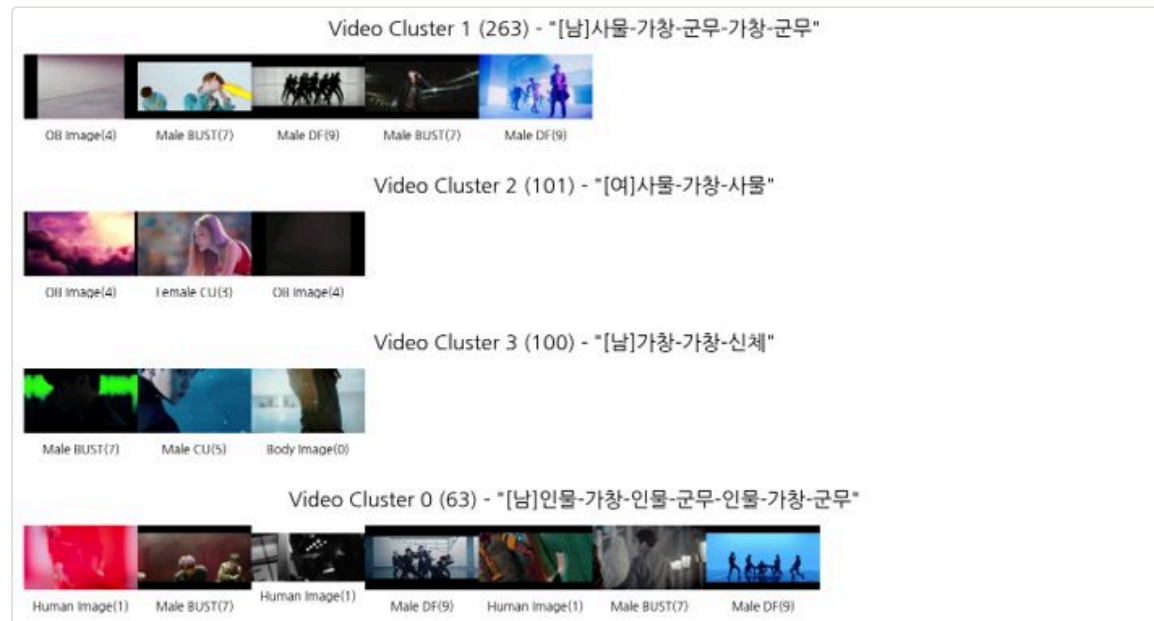


뮤직비디오별 HHI 지수 — 점 하나가 영상 한 편. 가로선 위쪽이 «솔림이 큰» 쪽 (N=527)

전체 뮤직비디오의 대표 이미지 패턴

유형이 어떤 순서로 이어지는지까지 보면, 뮤직비디오의 «짜임»이 드러남.

- 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



그룹별 대표 이미지 순서 (N=527)

여기까지가 «수집 다음»의 이야기임. 오늘 코드로 얻는 것은 이 그림의 맨 앞 두 칸이고, 그 뒤는 물음에 따라 갈림.

오늘의 마무리

Q & A

문의 — 이문혁 (경희대학교 인문사회과학데이터연구소)

`dalvitt@khu.ac.kr`