Music Video Information Retrieval

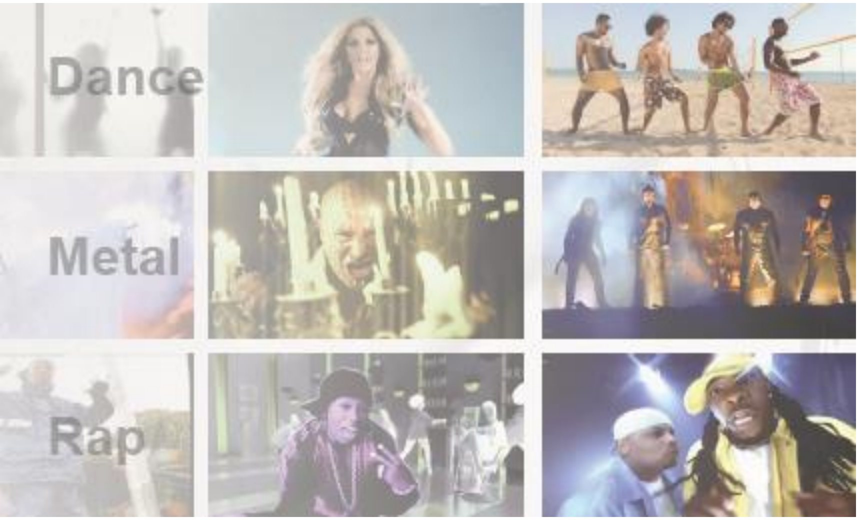
Alexander Schindler

Department of Softwaretechnology and Interactive Systems
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~schindler

MVIR Objectives

- Multimodal Approach to MIR Problems
  - Classification / Tagging
  - Mood estimation
  - Music Similarity Retrieval

- Hypothesis:
  - visual layer of music videos conta
    related information



- Research Aims
  - Can this information be used?
    - Improve MIR solutions
    - Use images as queries

Music Video Dataset

MVD-VIS

Genre	Videos	Artists
Bollywood	100	32
Country	100	70
Dance	100	84
Latin	100	72
Metal	100	76
Opera	100	NA
Rap	100	81
Reggae	100	75

MVD-MM

Genre	Videos	Artists
80s	100	72
Dubstep	100	78
Folk	100	66
Hard Rock	100	69
Indie	100	64
Pop Rock	100	65
Reggaeton	100	69
RnB	100	67

MVD-MIX

MVD-VIS + MVD-MM 16 Genres	1600	1040

MVD-Themes

Christmas	56	42
K-Pop	50	39
Broken Heart	56	48
Protest Songs	50	42

MVD-Artists

Artist Name	Videos	Artist Name	Videos	Artist Name	Videos
Aerosmith	23	Jennifer Lopez	23	Nickelback	18
Avril Lavigne	20	Justin Timberlake	12	P!nk	23
Beyonce	26	Katy Perry	12	Rihanna	25
Bon Jovi	27	Madonna	30	Shakira	24
Britney Spears	25	Maroon 5	14	Taylor Swift	20
Christina Aguilera	15	Matchbox Twenty	13	Train	11
Foo Fighters	23	Nelly Furtado	16		

MVD-Complete

MVD-VIS + MVD-MM + MVD-THEMES + MVD-ARTISTS	2212

Audio Classification Baseline Results

MVD-Results

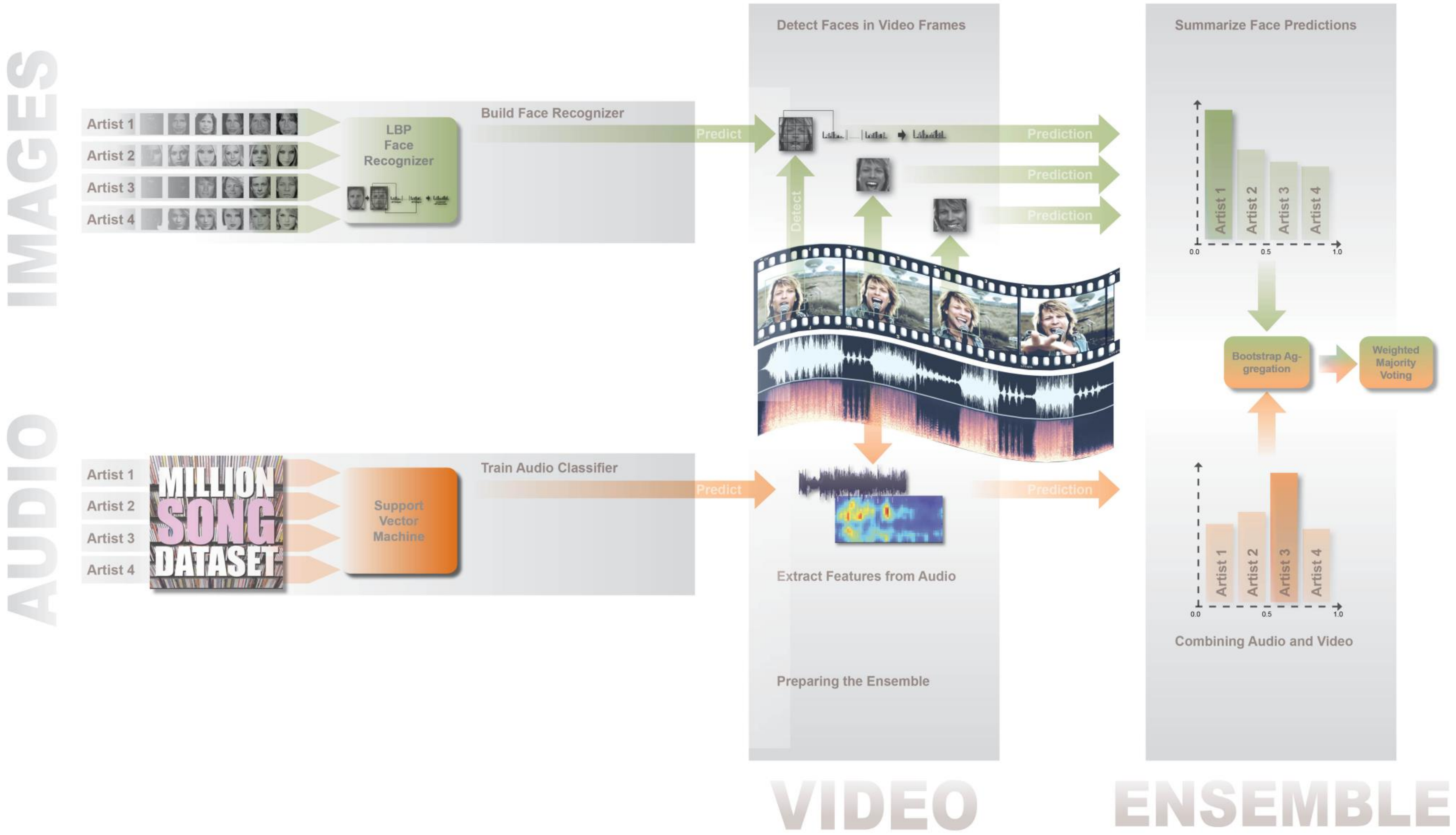
(a) Content Based Audio Features											
$a1$	Chroma	48	36.34	28.09	23.03	25.26	20.11	19.41	19.64	14.68	12.08
$a2$	MFCC	52	62.28	48.58	46.95	42.14	29.16	34.17	37.02	26.60	27.11
$a3$	SSD	168	85.78	73.18	58.81	68.74	50.28	48.41	65.11	44.64	38.92
$a4$	RP	1440	87.26	69.81	64.04	60.35	42.38	41.63	63.19	43.06	41.39
$a5$	TRH	420	71.04	55.83	53.86	49.50	38.28	39.66	46.61	33.02	35.70
$a6$	TSSD	1176	86.81	72.58	62.61	69.97	53.33	**53.65**	66.19	47.40	44.22
$a7$	$a4$+$a6$	2616	93.08	79.47	**71.88**	74.44	54.00	51.03	74.64	53.06	**48.54**
$a8$	$a4$+$a3$+$a5$	2028	92.19	75.93	67.45	71.00	50.26	44.85	72.73	49.88	43.65
$a9$	$a4$+$a3$	1608	92.55	77.74	67.36	71.64	52.44	44.40	74.38	51.60	43.52
$a10$	$a4$+$a5$+$a6$	3036	**93.79**	**80.85**	71.46	**74.76**	**55.00**	52.20	**75.91**	**54.16**	48.32

Baseline

Comparison with MIR Datasets

Classifiers	chro	spfe	timb	mfcc	rp	rh	trh	ssd	tssd	EN0	EN3	EN4	EN5	TEN
ISMIR Genre Dataset														
SVM Poly	50.3	54.9	67.7	62.1	75.1	64.0	66.5	78.8	80.9	67.0	67.2	78.5	80.4	**81.1**
Latin Music Database														
SVM Poly	39.4	38.2	68.6	60.4	86.3	59.9	62.8	86.2	87.3	70.5	69.6	82.9	87.1	**89.0**
GTZAN														
SVM Poly	41.1	43.1	75.2	67.8	64.9	45.5	38.9	73.2	66.2	56.4	53.6	63.9	65.2	**66.9**
ISMIR Rhythm														
SVM Poly	38.1	41.4	60.7	54.5	88.0	82.6	73.7	58.6	56.0	55.1	51.7	62.7	63.7	**67.3**

Artist Identification



Artist Identification

MVD-Artists Results

Ensemble		Audio		Video		
Prec.	Recall	Prec.	Recall	Prec.	Recall	Artist
0.36	0.57	0.33	0.52	0.14	0.33	Aerosmith
0.64	0.45	0.50	0.45	0.62	0.25	Avril Lavigne
0.55	0.32	0.33	0.26	0.28	0.42	Beyonce
0.24	0.27	0.28	0.36	0.20	0.04	Bon Jovi
0.34	0.42	0.32	0.33	0.16	0.17	Britney Spears
0.33	0.50	0.48	0.71	0.18	0.43	Christina Aguilera
0.62	0.53	0.41	0.47	0.00	0.00	Foo Fighters
0.27	0.19	0.22	0.24	0.33	0.14	Jennifer Lopez
0.30	0.24	0.27	0.28	0.50	0.12	Madonna
0.35	0.70	0.20	0.10	0.12	0.80	Maroon 5
0.58	0.44	0.55	0.38	1.00	0.18	Nickelback
0.75	0.14	0.29	0.19	0.40	0.10	Rihanna
0.28	0.65	0.44	0.40	0.25	0.21	Shakira
1.00	0.16	0.60	0.32	0.50	0.06	Taylor Swift
0.47	**0.38**	**0.37**	**0.36**	**0.34**	**0.21**	**avg**

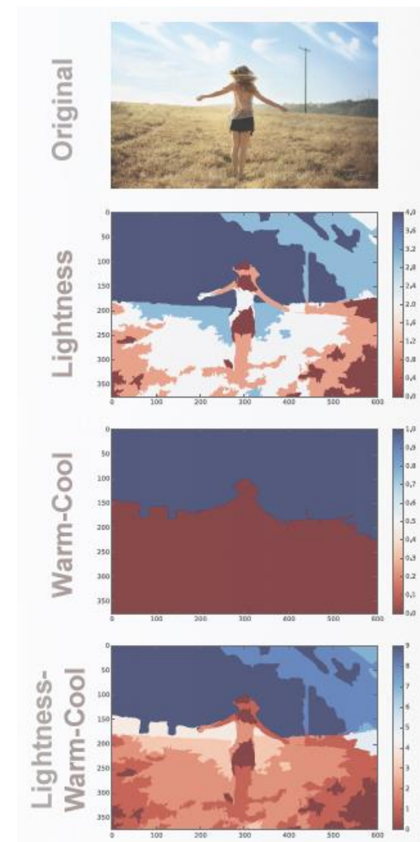
Low-Level image processing features

- **7 feature sets**
- **360 descriptors**

- **Colorfulness**
- **Itten Contrasts**
- **Lightness Fluctuation Patterns**
- **Etc.**

Color Names

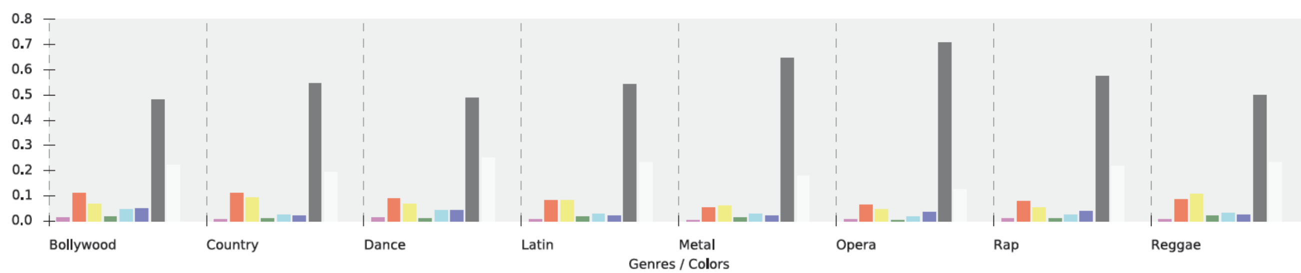
Affective Contrasts

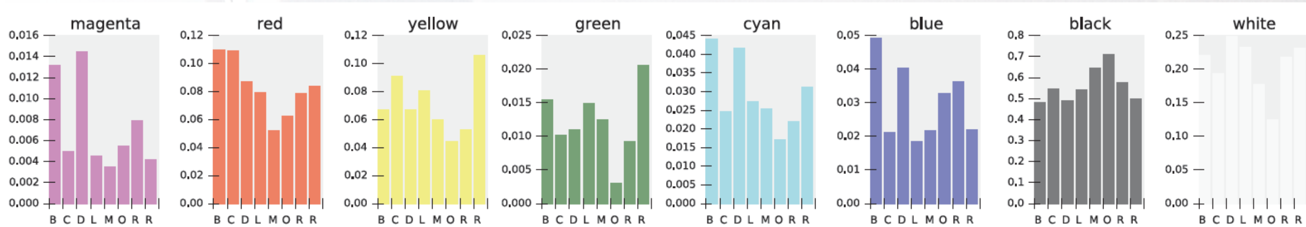


Low-Level features Results

(b) Low-level Color and Affect related Image Features

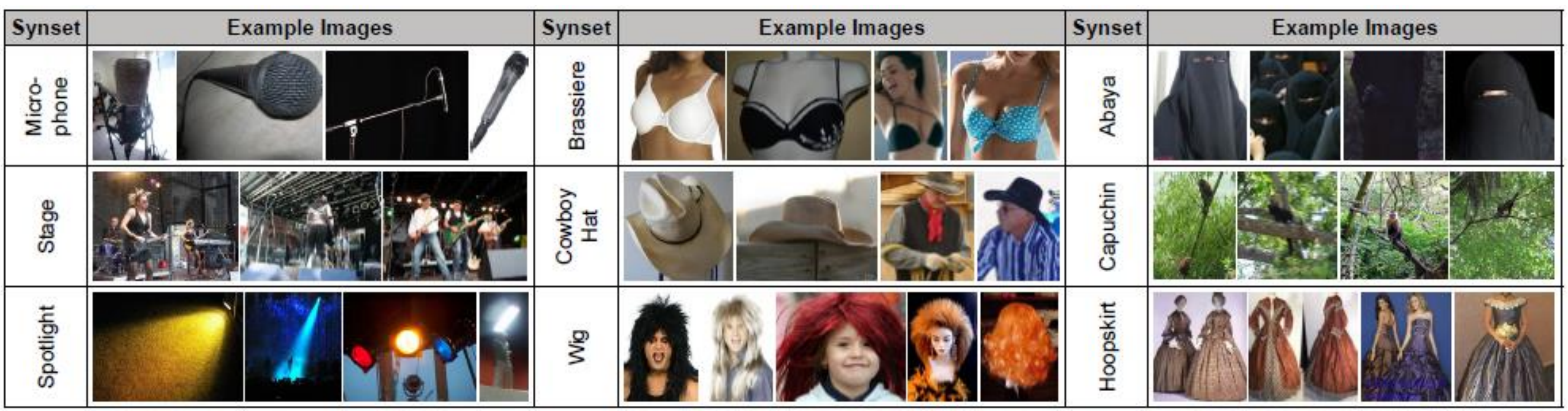
$v_{co}1$	LFP	60	33.21	23.59	25.45	20.38	16.74	16.46	16.93	11.71	13.36
$v_{co}2$	CF	7	34.89	25.49	31.50	21.84	17.06	20.41	18.53	11.92	16.49
$v_{co}3$	IC	28	36.80	27.55	27.51	24.83	**19.43**	19.68	21.44	13.54	12.66
$v_{co}4$	GEV	21	39.45	**29.84**	**34.15**	20.81	17.04	18.51	20.27	14.47	**17.89**
$v_{co}5$	GCS	42	40.55	29.76	33.91	24.08	17.29	18.15	**23.72**	**15.40**	17.34
$v_{co}6$	WAF	126	41.01	26.43	29.86	26.01	19.08	**21.38**	22.86	13.90	16.60
$v_{co}7$	CN	56	**43.68**	29.04	32.23	**26.74**	19.13	18.77	23.48	14.76	15.99
$v_{co}8$	Combi	360	**50.13**	**34.04**	**39.38**	**31.69**	**21.16**	**23.38**	**32.22**	**17.89**	**21.16**





High-Level Visual Concepts

- Convolutional Neural Networks
- Applied Model
  - 1000 concepts of the Large Scale ImageNet classification campaing
  - Wide range of different semantic concepts



Top concepts of music video frames examples

stage	0.3162
electric guitar	0.1169
bassoon	0.0649
accordion	0.0611
drumstick	0.0386
microphone	0.0313
marimba	0.0276

mosquito net	0.0932
wardrobe	0.0857
brassiere	0.0815
shower curtain	0.0471
candle	0.0400
plastic bag	0.0204
hoopskirt	0.0187

maillot	0.2745
bolo tie	0.0732
Windsor tie	0.0550
letter opener	0.0486
brassiere	0.0390
bikini	0.0384
bassoon	0.0364

lumbermill	0.1925
tow truck	0.1215
harvester	0.1152
thresher	0.0513
jeep	0.0484
half track	0.0473
pickup truck	0.0460

wig	0.4399
neck brace	0.0577
chimpanzee	0.0418
hair spray	0.0375
orangutan	0.0366
cloak	0.0267
Windsor tie	0.0236

Classification results (visual concepts only)

(c) High-level Visual Concepts											
$v_{in}1$	MEAN	1000	66.86	42.09	53.69	51.26	31.23	37.05	46.87	23.90	**33.07**
$v_{in}2$	STD	1000	69.78	46.76	50.08	51.95	29.99	32.88	48.29	26.83	29.63
$v_{in}3$	MAX	1000	73.15	44.26	46.41	54.60	33.05	31.94	50.07	26.93	27.49
$v_{in}4$	$v_{in}3+v_{in}2$	2000	73.61	46.53	**51.21**	55.04	31.48	34.00	51.30	27.03	31.04
$v_{in}5$	$v_{in}3+v_{in}1$	2000	**74.36**	47.70	53.65	55.99	**33.70**	**37.83**	**51.58**	**28.88**	33.83

Multimodal Improvements?

- **Improvements**
  - MVD-VIS => 2.94%
  - MVD-MM => 6.84%
  - MVD-MIX => 10.82%
- **Largest Improvement:** TSSD (MVD-MIX) => 16.43%

(a) Content Based Audio Features											
$a6$	TSSD	1176	86.81	72.58	62.61	69.97	53.33	**53.65**	66.19	47.40	44.22
$a10$	$a4+a5+a6$	3036	**93.79**	**80.85**	71.46	**74.76**	**55.00**	52.20	**75.91**	**54.16**	48.32

(e) Audio-Visual Combinations											
$av1$	$a10+v_{in}5$	5036	**96.73**	81.13	65.00	**81.60**	**55.73**	49.31	**86.73**	**59.01**	47.48
$av2$	$a9+v_{in}5$	3608	**95.63**	77.05	64.16	**77.83**	49.54	46.58	**79.44**	51.31	43.71
$av3$	$a9+v_{pl}5$	2118	94.50	79.95	68.08	72.96	53.29	45.99	77.40	53.73	45.51
$av4$	$av2+v_{pl}5$	4118	95.76	75.76	61.00	77.55	50.31	44.59	80.16	52.43	41.79
$av4$	$a6+v_{in}5$	3176	**94.65**	68.61	63.64	**78.49**	53.01	50.41	**82.62**	48.94	48.53
$av5$	$a4+v_{in}5$	3440	**91.24**	68.80	63.40	**71.95**	**43.78**	**44.86**	**74.14**	**45.53**	42.69
$av6$	$a3+v_{in}5$	2168	**89.85**	62.11	57.89	**70.13**	43.16	42.93	**70.30**	37.98	38.88

Music Tagging / MVD-Themes Classification

- **Non-Audible Themes**
  - Cross-genre problem
  - Aligned to MusiClef music tagging task
- **Improvements**
  - Christmas => 45.5%
  - K-Pop => 14.1%
  - Protest Song => 22.6%
  - Broken Heart => No improvement
- Video features outperform audio-content descriptors

	Audio-Only				Visual-Only				Audio-Visual			
Theme	**VIS**	**MM**	**MIX**	**TH**	**VIS**	**MM**	**MIX**	**TH**	**VIS**	**MM**	**MIX**	**TH**
Christmas	67.6	36.7	29.5	52.9	71.7	**65.5**	**64.0**	88.9	**87.5**	**70.8**	75.0	**90.4**
K-Pop	86.0	65.4	68.6	86.0	**88.4**	**81.6**	**80.4**	**91.7**	**95.5**	**88.2**	82.7	**90.0**
Protest Song	50.0	21.7	7.7	47.5	23.7	**33.3**	**16.7**	75.5	44.4	**57.1**	30.3	**77.5**
Broken Heart	75.0	28.6	28.6	54.9	51.2	21.9	16.7	**70.2**	61.0	**31.9**	25.5	**68.6**

Salient Visual Concepts

Country	Dance	Metal	Opera	Reggae
1. cowboy hat	1. brassiere	1. spotlight	1. theater curtain	1. seashore coast
5. drumstick	3. maillot	2. electric-guitar	3. hoopskirt	2. academic gown
8. restaurant	4. lipstick	4. drumstick	5. stage	3. capuchin
9. tobacco shop	9. seashore coast	6. matchstick	11. flute	5. black stork
10. pickup truck	10. bikini	7. drum	19. harmonica	7. sunglasses
11. acoustic guitar	15. sarong	8. barn spider	21. marimba	8. orangutan
13. violin fiddle	16. perfume	10. radiator	25. oboe	9. titi monkey
16. jeep landrover	17. trunks	12. chain	26. french horn	10. lakeshore
18. tractor trailer	18. ice lolly	14. grand piano	27. panpipe	11. cliff drop
19. tow truck	19. pole	23. spider web	30. grand piano	17. elephant
21. minibus	20. bubble	24. nail	31. cello	23. steel drum
23. electric guitar	30. miniskirt	28. brassiere	48. pipe organ	24. macaw
33. thresher	42. feather boa	37. loudspeaker	55. harp	25. coonhound

Summary

- Visual Layer of Music Video contains music relevant information
- Can be harnessed using visual concept detection
- Significantly improves performance of multi-modal approaches
- Outperforms audio-only approaches in audio-tagging tasks

Thank You!