Multi-Modal Music Information Retrieval:

Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis

Rigorosum: Alexander Schindler
Supervisor: Ao.univ.Prof. Dr. Andreas Rauber
Reviewer: Univ. Prof. Mag. Dipl.-Ing. Dr. Markus Schedl
Univ. Prof. Dr. Allan Hanbury

Agenda

- Problem statement
- Structure of research activity
- Approaches and contributions
  - Artist Identification
  - Color of Music Videos
  - Visual Concept Detection
- Conclusions and Future work

Problem Statement

- Music Information Retrieval
- Multi-Modality
- Audio-Visual Augmentation
- Motivation
- Research Questions

Thesis Title

Multi-Modal Music Information Retrieval

Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis

- Music Information Retrieval
- Multi-Modality
- Augmenting Audio-Analysis with Visual Computing

Music Information Retrieval

Multi-Modal **Music Information Retrieval**: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis

- Multidisciplinary Research Field
  - Information Retrieval, Musicology, Psychology, Digital Signal Processing, etc.
- Research Tasks
  - Genre classification,Artist identification, Mood classification
- Approaches
  - Symbolic: e.g. music scores
  - Content: recorded audio

➔ Context of this thesis based on audio content analysis

Audio Content based Approaches

- **Idea:**
  - Task specific relevant information is provided by the audio
- **Aim:**
  - Extract information in an expressive representation
  - Also referred to as *feature extraction*
  - Common Features
    - Timbre (MFCC), Pitch (Chroma), Rhythm (Rhythm Patterns)
- **Problems:**
  - Semantic gap: Inability of features to describe abstract concepts (e.g. mood, genre)
  - Glass ceiling reached with audio only

Multi-Modal Approaches to MIR

Multi-Modal Music Information Retrieval: Augmenting Audio-Analysis with Visual Computing for Improved Music Video Analysis

- **Idea:**
  - harness information from different modalities
  - bridging the semantic gap
- **Modalities Examples:**
  - **Text**
    - Lyrics, Web pages, social media tagging
    - Genre-, Mood-Recognition
  - **Symbolic Music**
    - Midi + audio features ➔ Genre-Recognition

Augmenting Audio-Analysis with Visual Computing

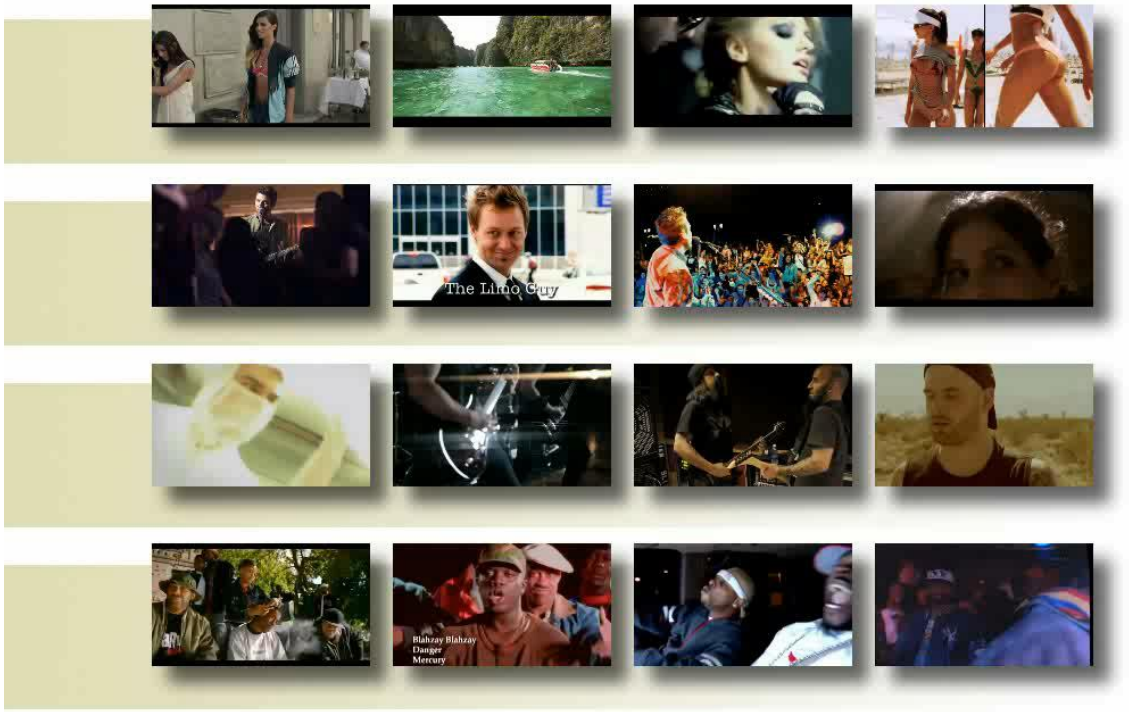
Multi-Modal Music Information Retrieval: **Augmenting Audio-Analysis with Visual Computing** for Improved Music Video Analysis

- Image-Analysis for Music-Retrieval
  - Album-art images
    - music genre recognition, emotion recognition
  - Artist promotional images
    - music genre recognition
- Music-Analysis for Video-Retrieval
  - Estimate video/movie genre
- Analysis of the visual layer of music videos
  - using visual layer to make predictions about audio
    - nearly unexplored

  ➔ Here this thesis is positioned!

Motivation

- **Observation:** Music characteristics that can be identified by the visual content
  - Genre, Style
    - Clothing, instruments
  - Mood
    - Candles, facial expression
  - Themes
    - Christmas tree
  - Tempo, Rhythm
    - Nodding, playing instrument
  - Artist
    - Face-ID



- **Hypothesis**
  - Existence of music related visual information

Research Questions

1. Which visual features are able to capture task related information?
   - Appropriate state-of-the-art image/video processing features
   - develop custom features
   - extract music/task related information from visual sources.
   - Chapters 6 - 9
2. How can the different modalities be combined to improve performance?
   - appropriate methods to combine multi-modal information
   - fuse different modalities in feature space or ensemble methods.
   - Chapters 6,8,9

Research Questions

3. To which extent can these visual patterns be used to derive further music characteristics?
   - predict further music related properties
   - genres, themes, emotions, instruments, …
   - Chapter 6,8,9
4. Is it possible to verify concepts within the production process?
   - verify visual patterns reported in literature (see RQ1) through automated analysis
   - Chapter 9

Structure of Research Activities

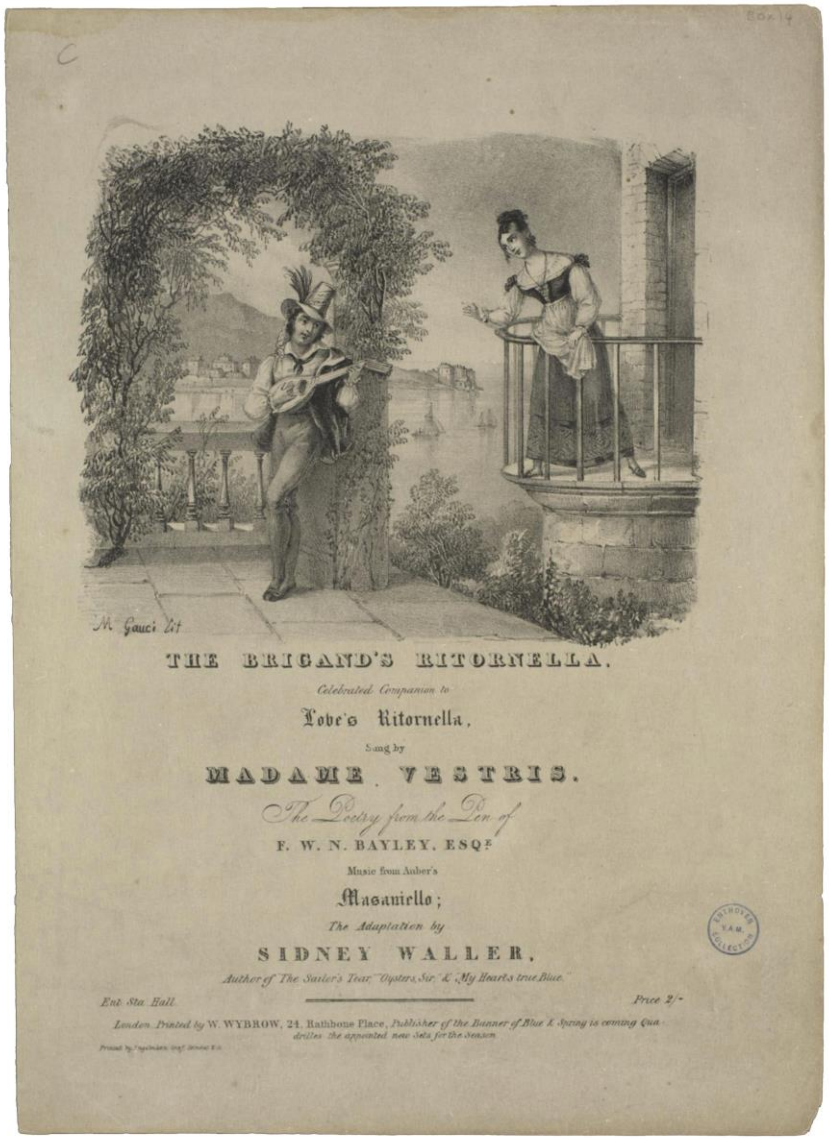
- Overall structure
- Awards
- Exploitation
- Timeline

Overall Structure

- Types of visual media
  - Album cover arts
  - Music Videos
- MIR Tasks evaluated
  - Artist Identification
  - Genre Classification
  - Mood/Themes Classification
- Publications
  - Music Videos: 5
  - Towards Album arts: 7
- Datasets
  - Music Videos: 1 - Music Video Dataset (4 subsets)
  - Towards Album arts:
    - 2 published - MSD Benchmarking collection, MSD multi-label tag-set
    - 3 pending - MSD Album-arts Dataset, MSD Lyrics, MSD Album Reviews

Approaches and Contributions

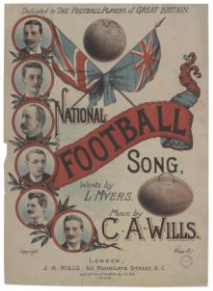
Visual Media in Music Distribution and Sales



Historic View

- 1820 - Artists became famous for cover design of songbooks
- 1896 - Thomas Edison establishes the National Phonograph Company
  - Wax cylinders, Company Name, Title
- 1901 - Victor Talking Machine Company
  - Flat disc phonographs, Sleeves, name of store, title, later portraits of the performing artists/composer
- 1940 - Self-service stores
  - Album cover as marketing tool
- 1948 - Long Playing records (LP)
  - more fragile, new packaging with more space
- 1950 - Rock and Roll era
  - Portraits of artists, LPs only collections of hit singles

(A) Music sheet cover 'The Brigand's Ritornella' from Auber's opera Masaniello, showing the image of the singer Madam Vestris who was a hugely popular star of burlesque at the Paris Olympic Theatre, 1835.



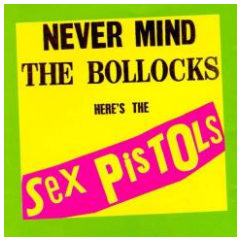
(B) Music sheet cover 'The National Football Song' designed by H. G. Banks depicts popular football players, about 1880.

(A) Edison Gold Moulded record made of relatively hard black wax, 1904. © Wikipedia

(B) 78rpm record with sleeve by Victor Talking Machine Company, 1928. © Jeff Crompton

Historic View

- 1960 - Complete Albums
  - Artists got involved in album design
  - Album as holistic art concept; famous designers
  - Visual design as part of the marketing campaign
- 1975 - Punk Rock
  - Opposition to established aesthetics
  - Simple visual composition
- 1981 - MTV
- 1983 - Compact Disc (CD)
  - "Jewel Box"
  - Smaller space
  - Medium lost its haptic nature
- 2000 - MP3, IPod
- 2005 – Youtube
- 2006 - Spotify



Examples of Music Distribution

- Typical music selection process in the 1980s
  - Go to record store
  - Decide from a great offer of records
  - Pre-listen at provided record players
- Time consuming process
- Often no devices for pre-listening available
- Only reliable information => visual representation

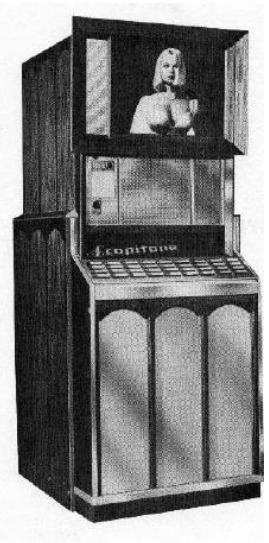
Spotify (1990)

Music Recommender System (1990)

Music Videos

- History of Music Videos
  - Early attempts already 1940
  - MTV 1981
  - Youtube 2005
- Expressive Types
  - Illustration
  - Amplification
  - Contradiction
- Music Video Types
  - Performance Video
  - Concept Video
  - Narrative Video
  - Dance Video
  - Animated Video
  - Lyrics Video

(A) The *Movie Machine* or *Panoram Soundie* by Mills Novelty Company, 1940.



(B) Scopitone machine, 1960.

(A) **Performance Videos**, *Van Halen - Jump, 1970.* The band is filmed while performing in synch to the music.

(B) **Concept Videos**, *Red Hot Chili Peppers - Californication, 1999.* Video is realized as a fictional 3D video game.

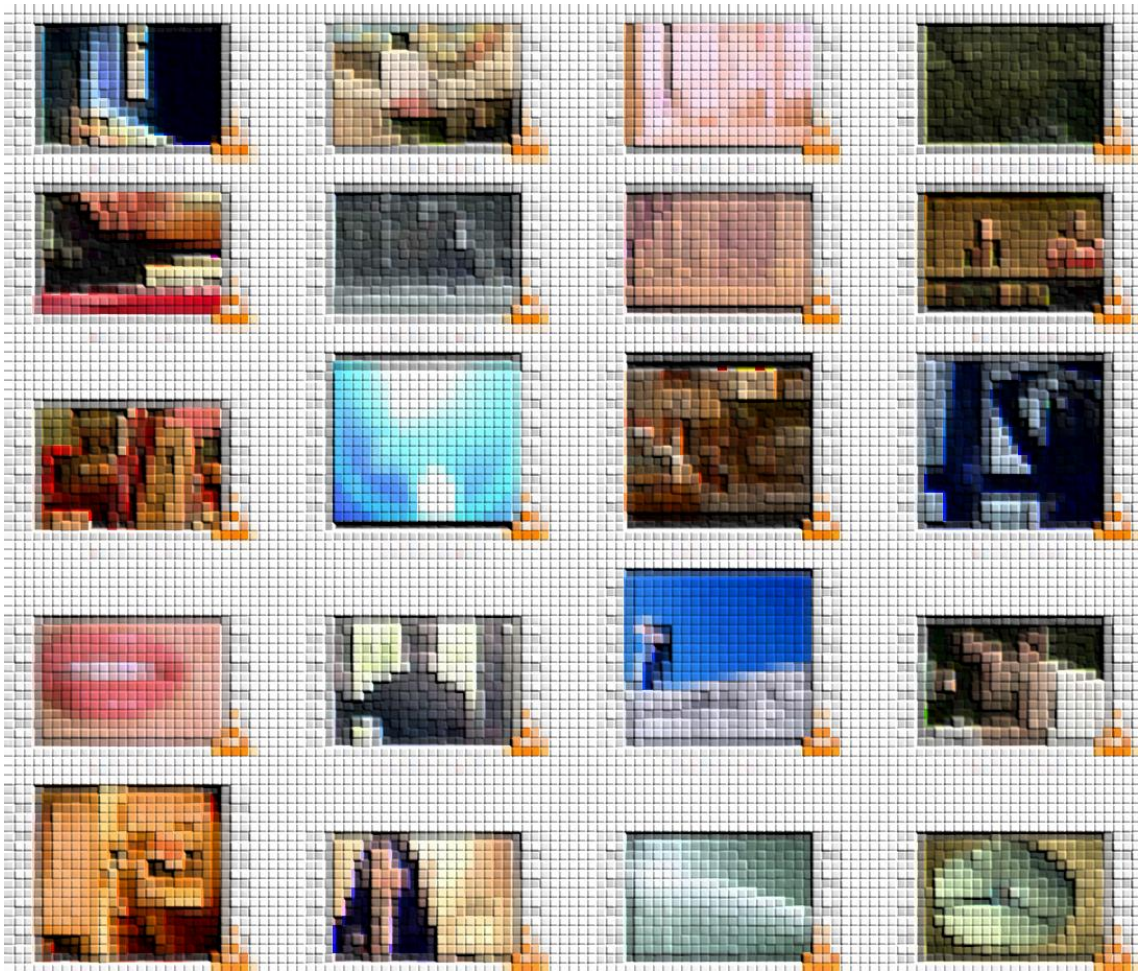
(C) **Narrative Videos**, *Beyoncé - If I Were A Boy, 2008.* The video plays with gender stereotypes by switching the roles.

(D) **Dance Videos**, *Michael Jackson - Bad, 1987.* Choreagraphed dance in reference to the musical 'West Side Story'.

(E) **Animated Videos**, *Gorillaz - Clint Eastwood, 2001.* Animated band members are performing with dancing zombie gorillas.

(F) **Lyrics Videos**, *Neil Young - My Pledge, 2016.* Lyrics appear on array of postcards.

Datasets



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

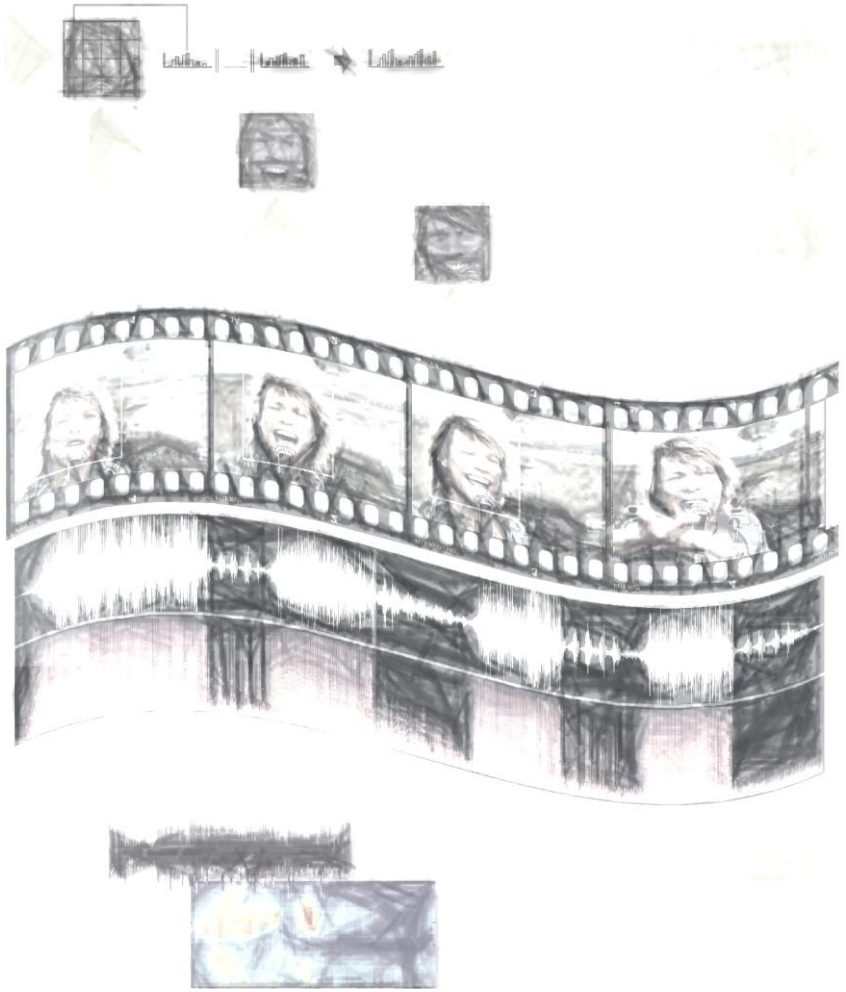
(a) Content Based Audio Features											
$a1$	Chroma	48	36.34	28.09	23.03	25.26	20.11	19.41	19.64	14.68	12.08
$a2$	MFCC	52	62.28	48.58	46.95	42.14	29.16	34.17	37.02	26.60	27.11
$a3$	SSD	168	85.78	73.18	58.81	68.74	50.28	48.41	65.11	44.64	38.92
$a4$	RP	1440	87.26	69.81	64.04	60.35	42.38	41.63	63.19	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

ISMIR Genre Dataset														
Classifiers	chro	spfe	timb	mfcc	rp	rh	trh	ssd	tssd	EN0	EN3	EN4	EN5	TEN
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**

A music video information retrieval approach to *Artist Identification*



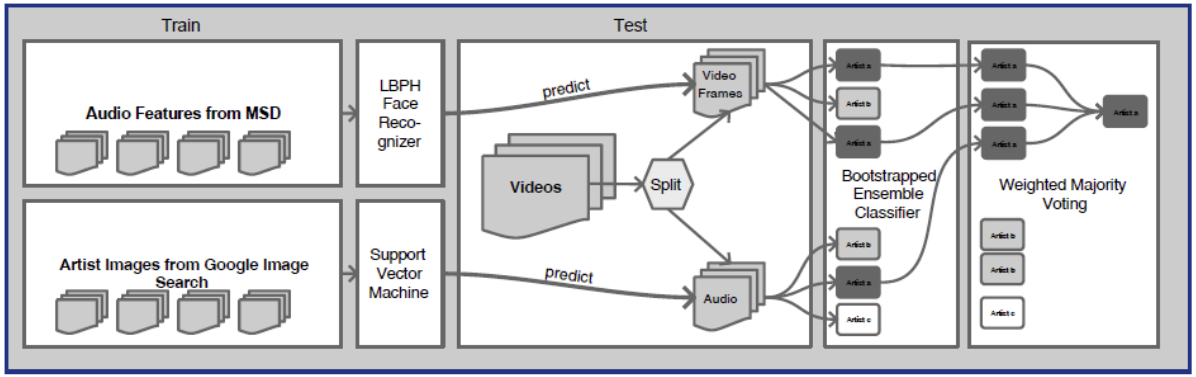
Alexander Schindler and **Andreas Rauber**. A music video information retrieval approach to artist identification. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR2013) to appear*, Marseille, France, October 14-18 2013.

Problem statement

- Aim: Identify performing artist
- State-of-the-art
  - Classification with Block level features (e.g. MFCC, Chroma)
  - Extracting Singers voice
  - Vocal segmentation / singer identification
- Challenges
  - Features more related to genre/style
  - Abstraction/Generalization vs. Identification
  - Artists progress, develop, change style/genre
- Methodology
  - Harness facial information of the performing artist

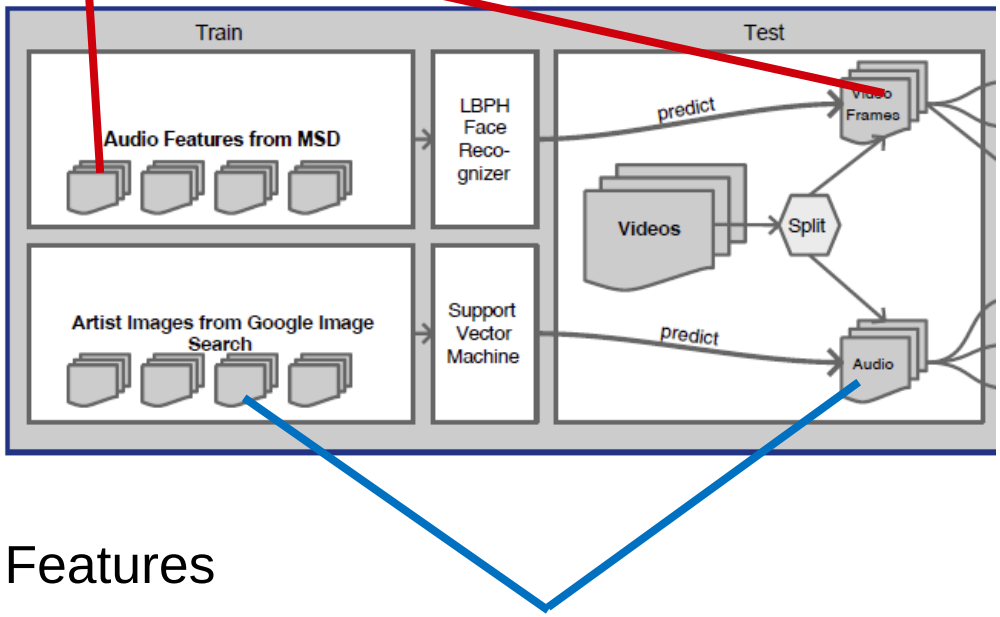
Audio-Visual Approach

- Architecture
  - 3-tiered classification approach
- Datasets
  - **Training Data**
    - Artist tracks
      - from MSD not in MVD-Artists
    - Artist Images
      - Google Images
      - If band ➔ only lead singer
  - **Test Data**
    - MVD-Artists Dataset
    - Constraint: official music video, artists have to appear in video



Audio-Visual Features

- Visual Features
  - Face Detection boosted cascades of Haar-like features
  - Local Binary Patterns (LBP)



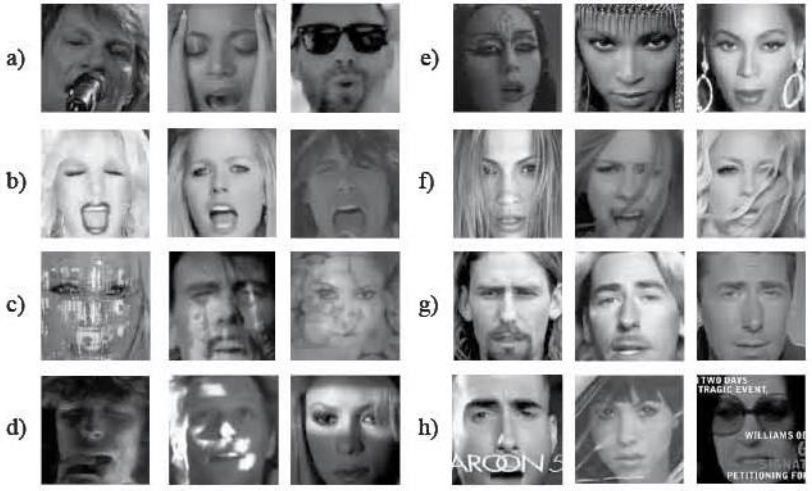
- Music Features
  - Temporal Echonest Features (TEN)
    - Aggregated Subset of MSD Features (Pitch, Timbre, Loudness, Segment length / 224 Dims)
    - Train Data ➔ MSD
    - Test Data ➔ Echonest API

Alexander Schindler and Andreas Rauber. **Capturing the temporal domain in echonest features for improved classification effectiveness**. In *Adaptive Multimedia Retrieval*, Lecture Notes in Computer Science, Copenhagen, Denmark, October 24-25 2012. Springer

Face Detection

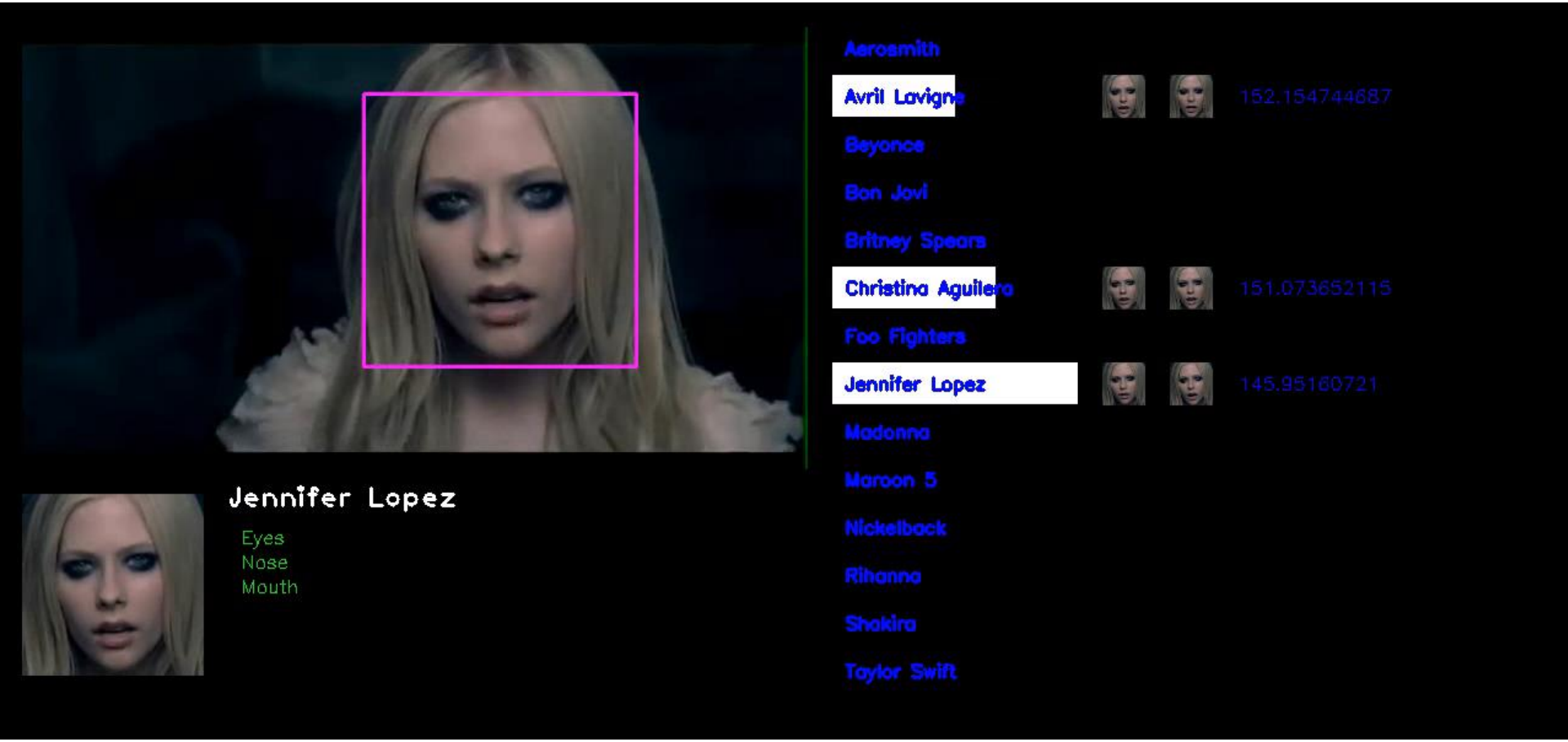
- Boosted cascades of Haar-like features
  - Computationally efficient with high prediction accuracy
  - simple features, neighboring pixel statistics
  - AdaBoost classifier, combined in a cascading structure

- Obstacles
  - Occlusions (a)
  - Distortions (b)
  - Blending Effects (c)
  - Illumination (d)
  - Make up / Jewelry (e)
  - Facial Hair (f,g)
  - Overlays (h)
  - Pose



- Additional detectors for eyes, nose and mouth
  - Automatic verification / reduction of false positives

Detection / Prediction Example



Ensemble Classifier

- Bootstrap Aggregation (Bagging)
  - Multiple versions of the same classifier
    - Using random sub-sampling for bootstrapping
  - Classifier: Support Vector Machine (SVM)
  - 10x Bootstrapping
    - 90% training / 10% assess prediction confidence $conf_{audio}$
  - Prediction of each classifier weighted by ist confidence $\overline{pred}_{audio} = pred_{audio} * conf_{audio}$
  - Final prediction

$$pred_{ensemble} = argmax[\sum_{0}^{i} \overline{pred}_{audio} + \sum_{0}^{i} \overline{pred}_{video}]$$

Results

Artist Name	Audio			Video			Audio-Visual Ensemble		
	Prec	Recall	f1	Prec	Recall	f1	Prec	Recall	f1
Aerosmith	0,33	0,52	0,39	0,14	0,33	0,20	0,36	0,57	0,44
Avril Lavigne	0,50	0,45	0,47	0,62	0,25	0,36	0,64	0,45	0,53
Beyonce	0,33	0,26	0,29	0,28	0,42	0,33	0,55	0,32	0,40
Bon Jovi	0,28	0,36	0,32	0,20	0,04	0,07	0,24	0,27	0,25
Britney Spea	0,32	0,33	0,33	0,16	0,17	0,16	0,34	0,42	0,38
Christina Ag	0,48	0,71	0,57	0,18	0,43	0,26	0,33	0,50	0,40
Foo Fighters	0,41	0,47	0,44	0,00	0,00	0,00	0,62	0,53	0,57
Jennifer Lop	0,22	0,24	0,22	0,33	0,14	0,20	0,27	0,19	0,22
Madonna	0,27	0,28	0,24	0,50	0,12	0,19	0,30	0,24	0,27
Maroon 5	0,20	0,10	0,13	0,12	0,80	0,20	0,35	0,70	0,47
Nickelback	0,55	0,38	0,44	1,00	0,18	0,30	0,58	0,44	0,50
Rihanna	0,29	0,19	0,23	0,40	0,10	0,15	0,75	0,14	0,24
Shakira	0,44	0,40	0,41	0,25	0,21	0,23	0,28	0,65	0,39
Taylor Swift	0,60	0,32	0,41	0,50	0,06	0,10	1,00	0,16	0,27
	0,37	0,36	0,35	0,34	0,21	0,20	0,47	0,38	0,37

- Audio-baseline improved by
  - Precision: 27%
  - Recall / f1: 6%

Conclusions

- First successful demonstration
  - Potential of harnessing visual layer of music videos to improve the MIR task
  - Artist Identification Precision improved by 27%, Recall by 6%

- Underperforming Face Recognition approach
  - Not invariant to head pose changes, illumination changes.
  - Performance could be improved with DNN based recognizers

Analyzing Color in Music Videos

Alexander Schindler and **Andreas Rauber**. An audio-visual approach to music genre classification through affective color features. In Proceedings of the 37th European Conference on Information Retrieval (ECIR'15), Vienna, Austria, March 29 - April 02 2015.

Problem statement

- No substantial research on audio-visual relationships in music videos
- Hypothesis:
  - Correlation between visual aesthetics and characteristics of the song
  - Music videos with expressive type *Illustration, Amplification*
- Aim:
  - Provide a bottom-up evaluation of affective visual features applied to MIR tasks
  - Focus on low-level color features
  - Analyze audio-visual relationships used in common language
    - Black Metal, Dark/bright music
- Evaluation:
  - Music Genre Classification
  - Dataset: MVD-VIS, MVD-MM and MVD-MIX

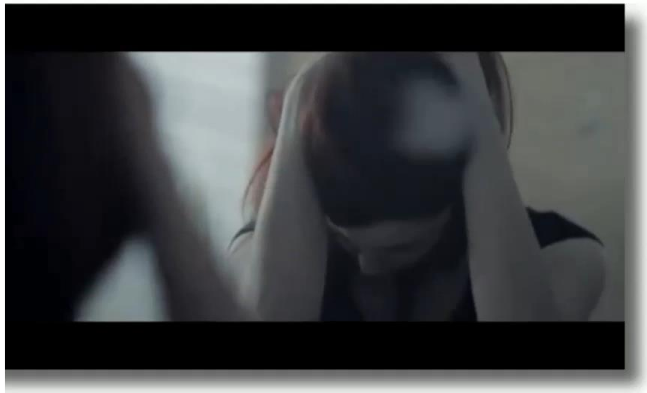
Acoustic & Visual Features

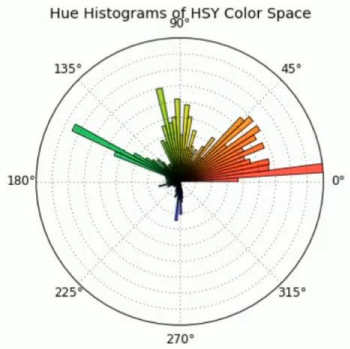
- 7 Audio Feature-sets
- 7 Visual Feature-sets
  - Low-level image processing features
  - Color features derived from art-theory
  - based on psychological studies

	Short Name	#	Description
Audio	Statistical Spectrum Descriptors (SSD)	168	Statistical description of a psycho-accoustic transformed audio spectrum
	Rhythm Patterns (RP)	1024	Description of spectral fluctuations
	Rhythm Histograms (RH)	60	Aggregated Rhythm Patterns
	Temporal SSD and RH		Temporal variants of RH (TRH #420), SSD (TSSD #1176)
	MFCC	12	Mel Frequency Cepstral Coefficients
	Chroma	12	12 distinct semitones of the musical octave
Visual	Global Color Statistics	6	mean saturation and brightness, mean angular hue, angular deviation, with/without saturation weighting
	Colorfulness	1	colorfulness measure based on Earth Movers Distance
	Color Names	8	Magenta, Red,Yellow,Green,Cyan Blue, Black, White
	Pleasure, Arousal, Dominance	3	approx. emotional values based on brightness and saturation
	Itten Contrasts	4	Contrast of Light and Dark, Contrast of Saturation, Contrast of Hue and Contrast of Warm and Cold
	Wang Emotional Factors	18	Features for the 3 affective factors by Wang et al. [281]
	Lightness Fluctuation Patterns	80	Rhythmic fluctuations in video lightness

Global Color Statistics

- Improved Hue, Luminance and Saturation (IHLS) color space
  - The saturation of achromatic pixels is always low
  - independence of saturation from the brightness function





- Color Statistics
  - Mean Saturation/Brightness
  - Hue in IHLS is angular ➔ circular statistics
  - Angular mean/std Hue
  - Saturation weighted mean/std Hue

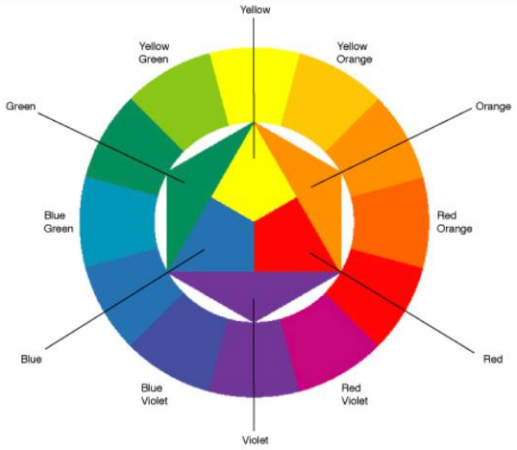
Global Emotional Values

- Perceptual / Affective Features
  - Relationship between saturation (S) and brightness (B)
  - Evaluated in psychological studies

$$Pleasure = 0.69 \cdot B + 0.22 \cdot S$$
$$Arousal = -0.31 \cdot B + 0.60 \cdot S$$
$$Dominance = 0.76 \cdot B + 0.32 \cdot S$$

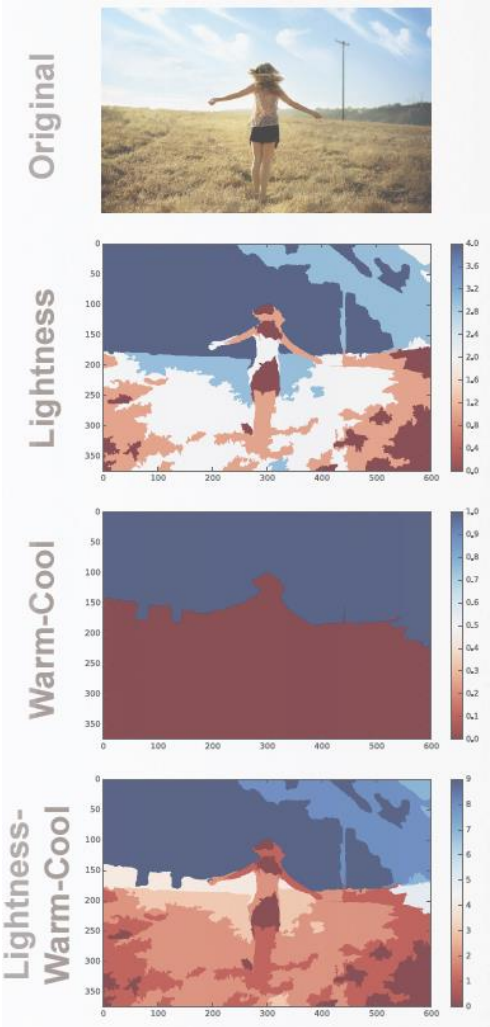
Itten‘s Contrasts

- **Art-theory concepts** defined by Johannes Itten
- **Opponent Color model**
- Contrasts extracted
  - Contrast of Light and Dark
  - Contrast of Saturation
  - Contrast of Hue
  - Contrast of Warm and Cold.



Wang Emotional Factors

- **lightness** description of a segmented image ranging from ***very dark*** to ***very bright***, combined with the hue labels *cold* and *warm*.
- description of **warm** or **cool** regions with respect to **different saturations** as well as a description of contrast.
- combines **lightness** contrast with an estimation of **sharpness**. A no-reference perceptual blur measure was used.



Color Names

- Map colorspace to Named Colors

a) Original image with **Letterboxing**

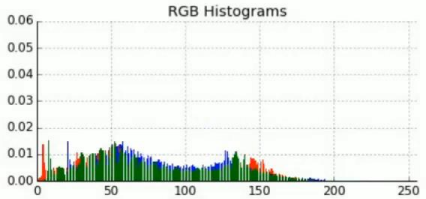
b) Pre-processing: remove letterboxing

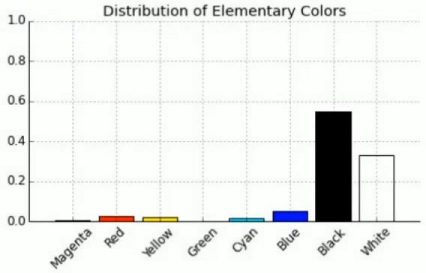
c) Color Quantization: Naive nearest neighbor match

d) Ordered Dithering (OD)

e) OD + **enhanced brightness (CLAHE)**

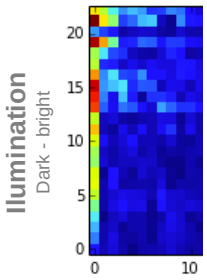
f) OD + **enhanced brightness, saturation**

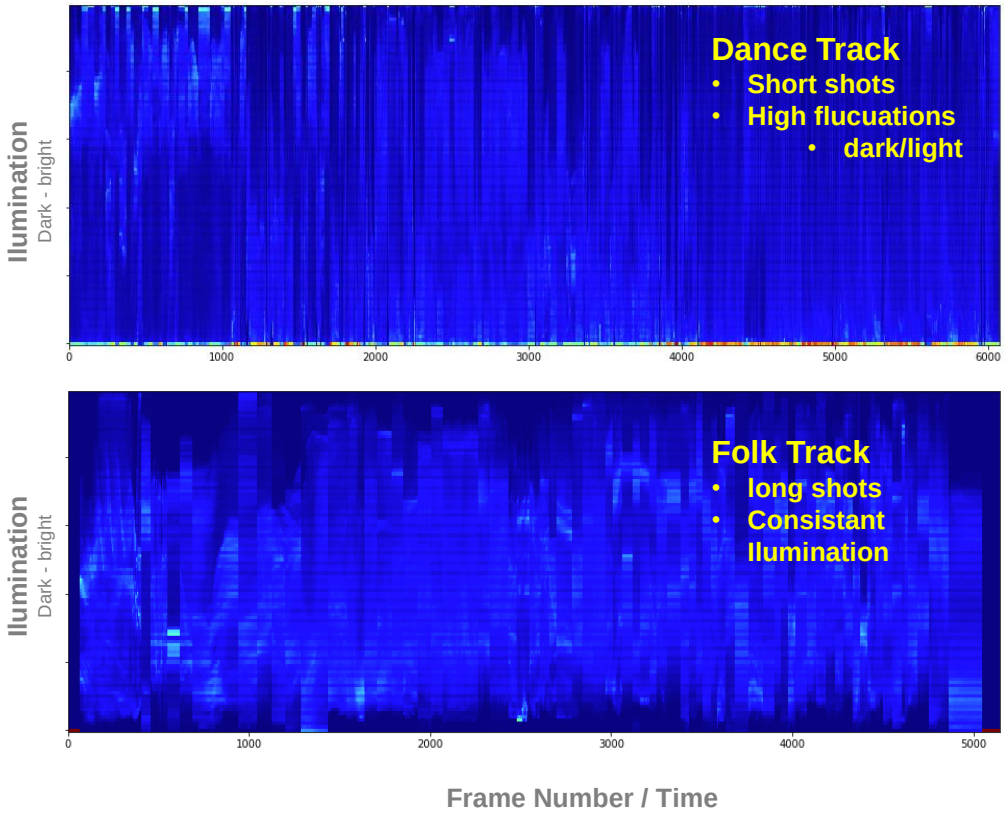




Lightness Fluctuation Patterns

- Calculated analogous to Rhythm Patterns
  1. Transform frame to LAB color space
  2. 24-bin histogram of the „L“uminance channel
  3. Fast Fourier Transform (FFT) over time-axis for each Lumnance-bin
  4. Amplitutde Modulation <= 10Hz
- Result
  - Time-invariant luminance fluctuation patterns
  - Capture tempo/rhythm of visual change in music videos





Results

		MVD-VIS			MVD-MM			MVD-MIX		
		SVM	KNN	NB	SVM	KNN	NB	SVM	KNN	NB
Audio	TSSD-RP-TRH	93.79	80.85	71.46	74.76	55.00	52.20	75.91	54.16	48.32
	TSSD	86.81	72.58	62.61	69.97	53.33	53.65	66.19	47.40	44.22
	RP	87.26	69.81	64.04	60.35	42.38	41.63	63.19	43.06	41.39
	SSD	85.78	73.18	58.81	68.74	50.28	48.41	65.11	44.64	38.92
	TRH	71.04	55.83	53.86	49.50	38.28	39.66	46.61	33.02	35.70
	MFCC	62.28	48.58	46.95	42.14	29.16	34.17	37.82	26.60	27.11
	Chroma	36.34	28.09	23.03	25.26	20.11	19.41	19.64	14.68	12.08
Visual	LFP	33.21	23.59	25.45	20.38	16.74	16.46	16.93	11.71	13.36
	CF	34.89	25.49	31.50	21.84	17.06	20.41	18.53	11.92	16.49
	IC	36.80	27.55	27.51	24.83	**19.43**	19.68	21.44	13.54	12.66
	GEV	39.45	**29.84**	**34.15**	20.81	17.04	18.51	20.27	14.47	**17.89**
	GCS	40.55	29.76	33.91	24.08	17.29	18.15	**23.72**	**15.40**	17.34
	WAF	41.01	26.43	29.86	26.01	19.08	**21.38**	22.86	13.90	16.60
	CN	43.68	29.04	32.23	**26.74**	19.13	18.77	23.48	14.76	15.99
	Visual Features combined	50.13	34.04	39.38	31.69	21.16	23.38	32.22	17.89	21.16
Audio-Visual	TSSD-RP-TRH	**94.86**	**81.38**	**71.65**	**75.69**	**55.78**	51.36	**76.53**	**55.76**	**49.08**
	TSSD	**88.45**	71.65	**64.75**	**70.55**	52.60	52.25	**69.46**	46.15	**45.16**
	RP	**89.80**	**71.99**	**65.78**	**62.79**	**43.93**	41.61	**66.59**	**44.47**	**41.68**
	SSD	85.25	62.05	57.80	65.34	42.28	44.24	65.21	36.13	**38.76**
	TRH	77.84	**55.98**	**59.71**	**58.50**	32.79	**41.40**	**56.31**	31.39	40.09
	MFCC	**63.71**	41.53	46.28	42.88	24.38	27.35	43.11	22.33	**25.62**
	Chroma	**55.70**	**39.28**	**43.13**	**35.29**	**24.16**	**25.51**	**35.43**	**20.10**	**24.14**

Conclusions

- Color = Low-Level Feature
  - Insufficient for predicting high-level concepts (objects, scenes, etc.) or genres
- Accuracy of 50.1% / 32.2%
  - Exceeds baseline significantly: 12.5% / 6.25%
  - Successfully proofs genre dependand color distributions in music videos
- Linguistic references: „dark“ / “light“ sounds
  - Analytically not confirmed

Further Visual Features

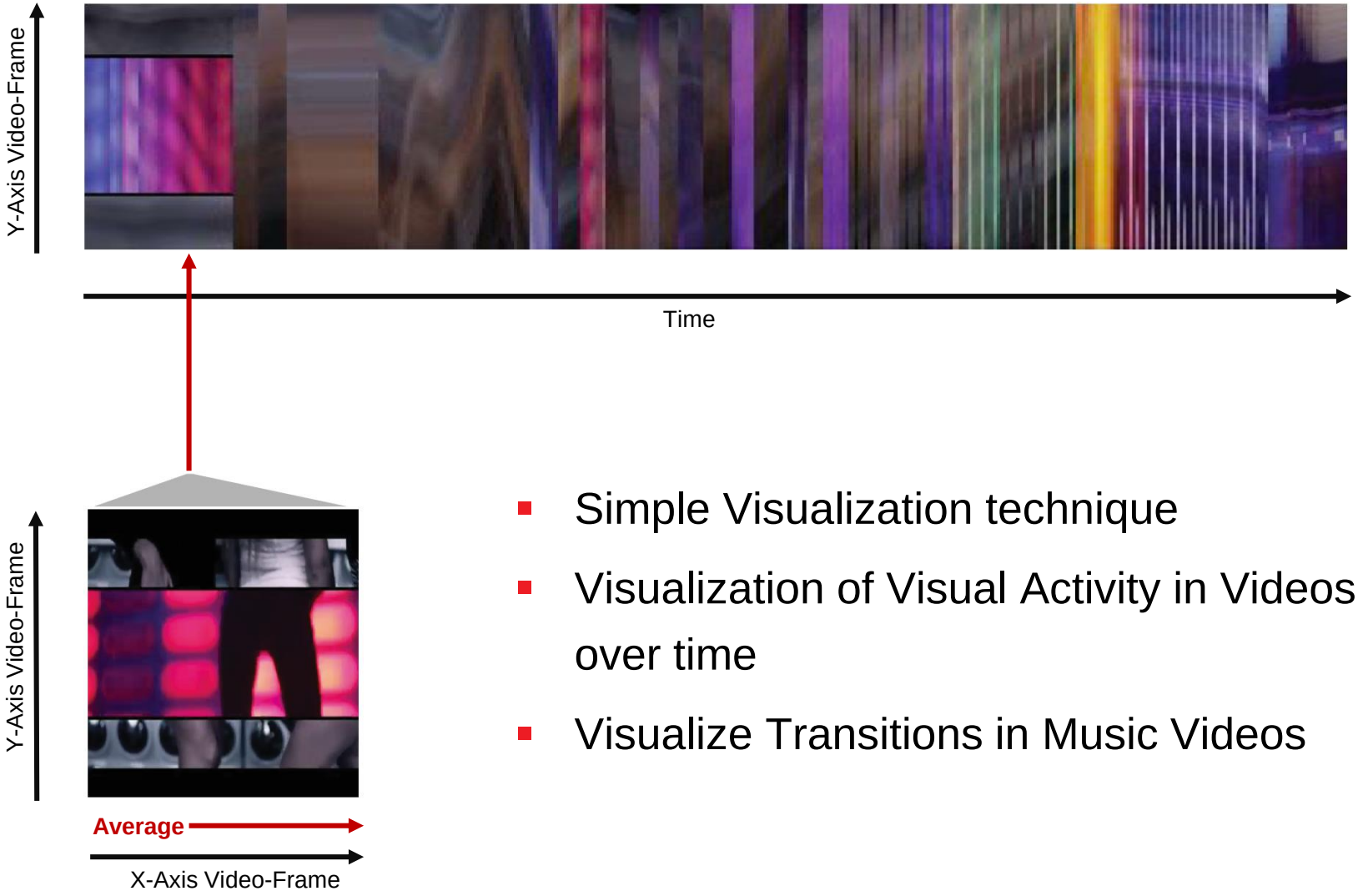
Shot Detection for Music Videos

Alexander Schindler and **Andreas Rauber**. On the unsolved problem of Shot Boundary Detection for Music Videos. In Proceedings of the 25th International Conference on MultiMedia Modeling (MMM2019), January 8-11, 2019, in Thessaloniki, Greece

Problemstatement

- Intention: Feature development
  - Shots per Minute (SPM)
    - Simlar to Beats per Minute (BPM)
    - Distinguish different genres
  - Average Shot-length / statistics
    - Distinguish different genres, epochs, themes
  - Transition type
    - Distinguish different genres
- Segmentation
  - Per-segment analysis (color, objects, etc.)
  - Music Segmentation vs. Music Video Segmentation

Average-Color Bar



- Simple Visualization technique
- Visualization of Visual Activity in Videos over time
- Visualize Transitions in Music Videos

Common Transition Types in Movies / TV

- Fade-In

- Object/Text overlay

- Zoom-In

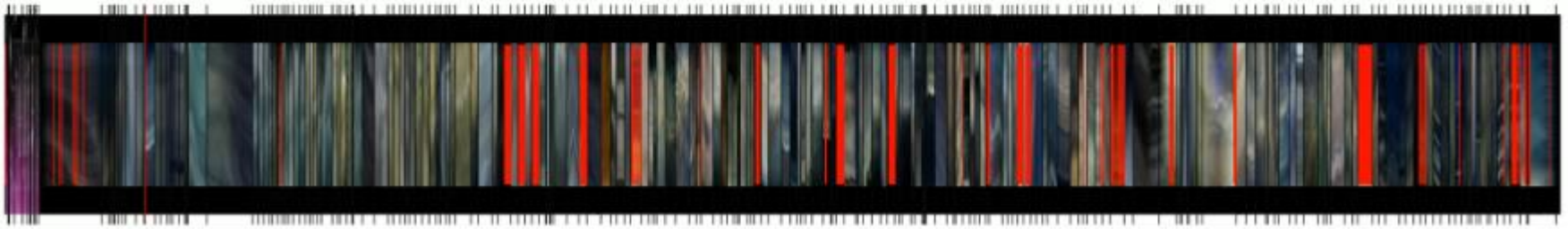
- Sharp-Cuts

- Moving Camera Focus

- Gradual Transitions / Dissolving

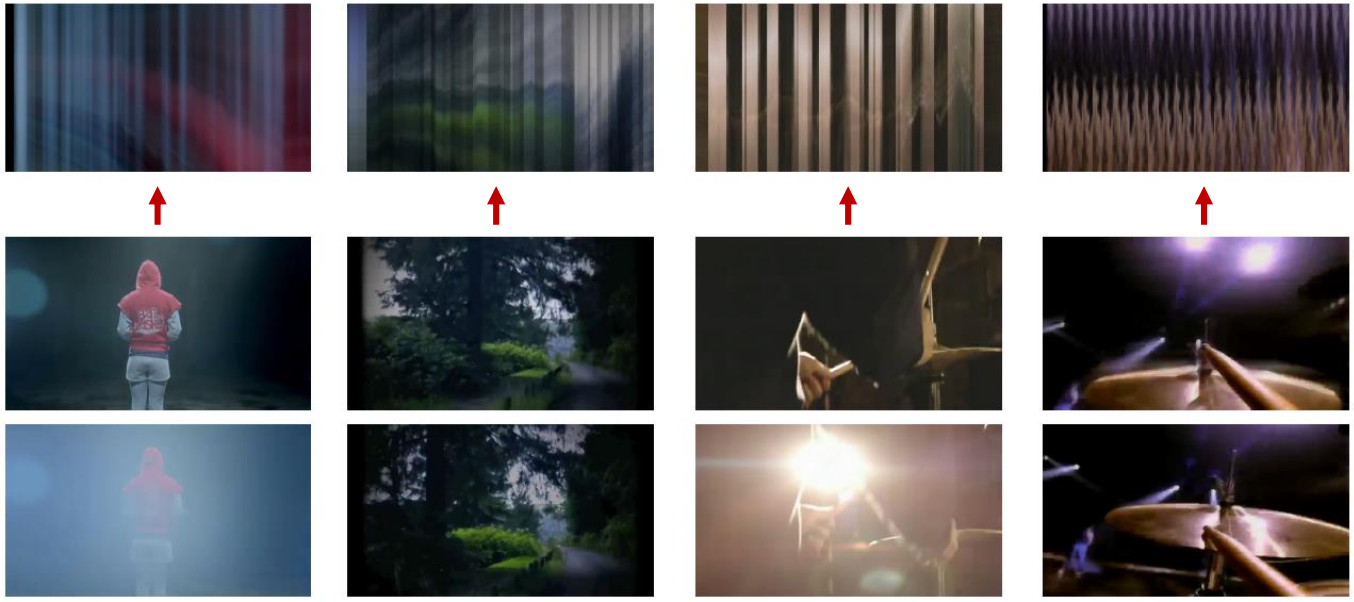
Detecting Standard Transitions

Frame IDX: 481



Uncommon Transition Types

- Split Screen
- Distortion
- Overlays
- Effects



Rapid Blending

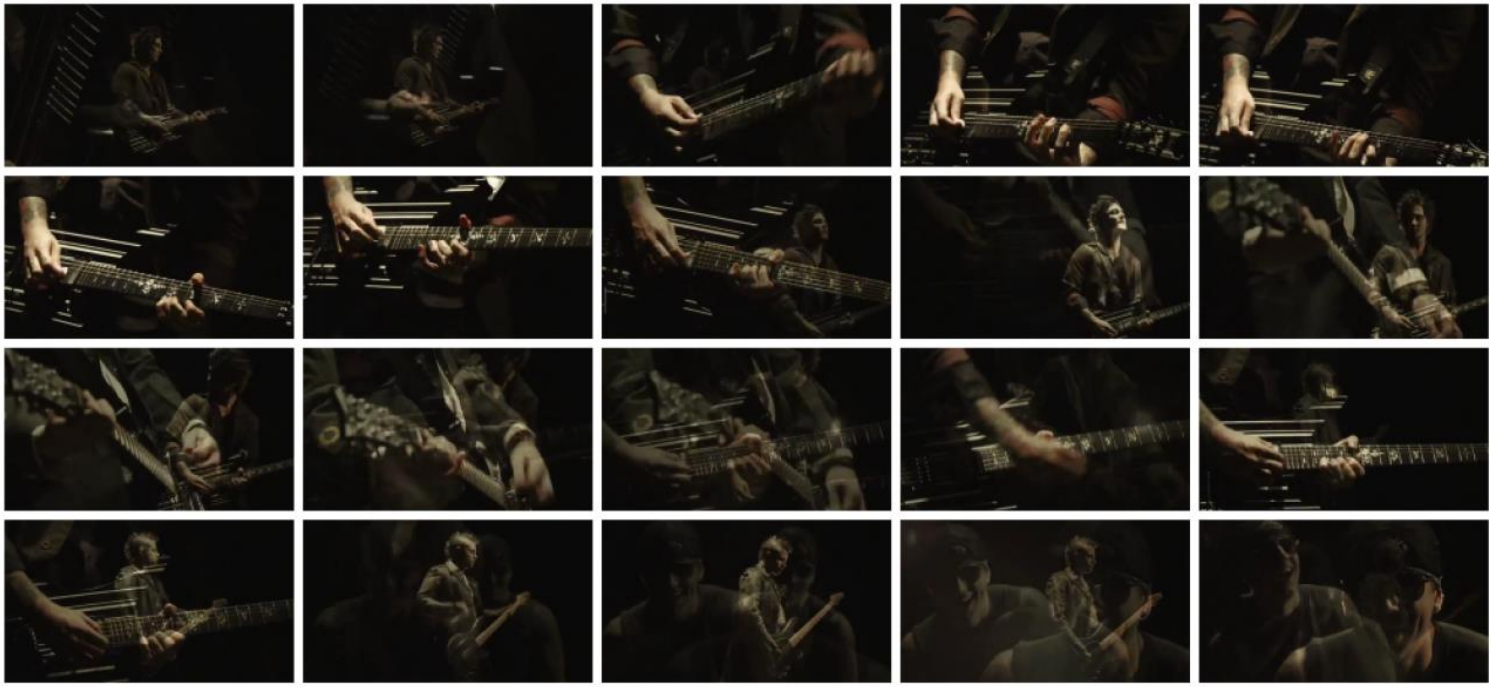


FIGURE 7.3 – Example of *Blending* video frames. Within 0.7 seconds (20 frames at 29 frames per seconds) 7 different scenes are blended.

Conclusions

- Feature development failed on missing definition of
  - Transition types
  - Coherent edited scenes vs. Fast paced transitions
- Position paper on Shot-boundary Detection in Music Videos
  - Documentation of findings and conclusions
  - Overview of different transition types
  - Initial taxonomy suggested
  - Discussion of problematic / undefined types
  - Suggestions for future work
- Provided Annotation Tool

Promising Approach

Music-related Visual Object Detection

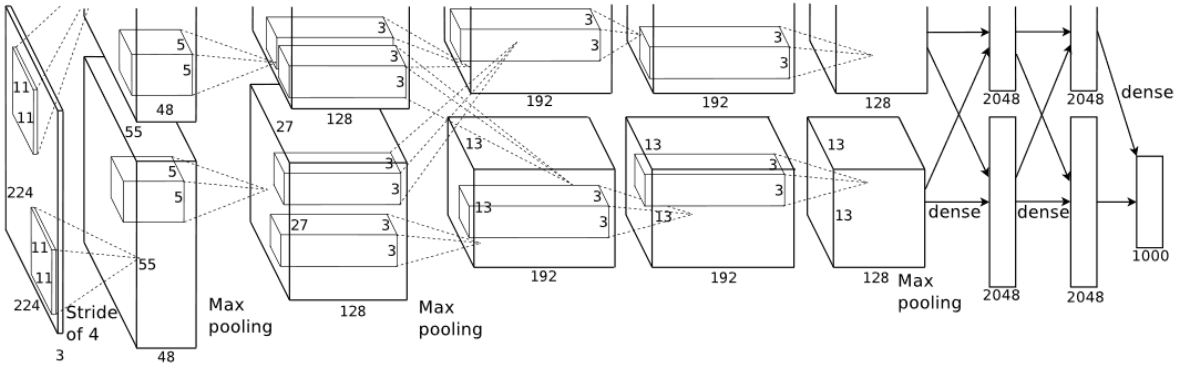
Alexander Schindler and **Andreas Rauber**. Harnessing Music related Visual Stereotypes for Music Information Retrieval. ACM Transactions on Intelligent Systems and Technology (TIST) 8.2 (2016):

Problemstatement

- Visual stereotypes
  - Play important role in social interaction
  - Trigger categorization process / expectations about behaviour, attitude, etc.
  - Drama theory and film make profound usage of visual stereotypes
- Extramusical Connotations (Meyer)
  - Visual associations, culturally shared
  - Artist development departments of Music
    - Visual concepts to categorize new Artists – genre, style, etc.
    - Make easily recognizable
    - Advertise in visual media – print, TV, etc.
- Music related Visual Stereotypes / Visual Language
  - Music Videos took part in the development
- Intention
  - Analyze visual objects in music videos
  - Identify music-related visual stereotypes

Approach

- Apply Object Detection to each Frame of Music Videos
  - AlexNet
    - Initially: every 10th frame (2014)
    - Titan X: every frame
  - Evaluate different approaches to aggregate Softmax Activations
    - Time-invariance
    - Best performance in Music Genre Classification



Feature 1 - ImageNet

- AlexNet pre-trained on Imagenet Dataset
  - 1000 object categroies
  - 40% animals, 60% objects of daily life, including music instruments
- Intention
  - Detect objects placed in music videos

Feature 2 – MIT Places Dataset

- AlexNet pre-trained on Places Dataset
  - 205 scene categroies
  - Street, bar, beedroom, etc.
- Intention
  - Detect music videos sceneries

Example ImageNet Detection Results

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

Most Salient Visual Objects in Music Videos

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

Evaluation 1: Visual Concept Detection

- Predict Genre up to 74.4% from visual objects only!
- Exceed performance of MFCC Genre classification
- Places also provide relevant information

			MVD-VIS			MVD-MM			MVD-MIX		
		Dim	SVM	KNN	NB	SVM	KNN	NB	SVM	KNN	NB
(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
(c) High-level Visual Concepts											
ImageNet	$v_{in}1$ 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
Places	$v_{pl}1$ MEAN	205	57.13	37.15	43.05	42.90	25.94	30.55	38.24	19.08	25.32
	$v_{pl}2$ MAX	205	58.36	42.28	45.35	38.91	25.51	31.44	36.63	21.74	27.33
	$v_{pl}3$ STD	205	60.74	40.70	39.39	43.95	27.58	28.99	39.33	20.90	23.03
	$v_{pl}4$ $v_{pl}1$+$v_{pl}2$	510	59.46	43.11	43.85	41.25	27.08	31.58	38.26	22.28	26.72
	$v_{pl}5$ $v_{pl}1$+$v_{pl}3$	510	60.49	39.74	40.99	43.40	26.45	30.33	39.72	20.50	24.88

Evaluation 2: Combinations of Visual Features

- No improvement for clearly acoustically defined genres
  - ImageNet and Places
  - Correlated Information
  - Overlapping Concepts (e.g. Villa, train)
- Improvement adding Color Information
  - Especially with higher number of classes / weaker genre definitions

			MVD-VIS			MVD-MM			MVD-MIX		
		Dim	SVM	KNN	NB	SVM	KNN	NB	SVM	KNN	NB
(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
(d) Visual Combinations											
$vc1$	$v_{in}5$+$v_{co}8$	2360	72.86	45.81	53.75	55.59	31.84	38.08	51.94	27.48	33.85
$vc2$	$v_{in}3$+$v_{pl}3$	1205	72.70	44.38	47.24	54.11	32.54	31.86	50.51	27.46	27.66
$vc3$	$v_{in}5$+$v_{pl}3$	2205	73.80	48.54	52.73	55.21	33.74	36.75	52.21	28.41	33.03
$vc4$	$v_{in}5$+$v_{pl}5$	2510	73.95	48.35	53.14	55.26	33.74	36.71	52.48	28.59	33.24
$vc5$	$vc4$+$v_{co}8$	2870	74.25	47.93	54.43	56.05	32.71	37.61	54.18	28.28	33.79

ImageNet + Places

ImageNet + Places + Color

Evaluation 3: Audio-Visual Combinations

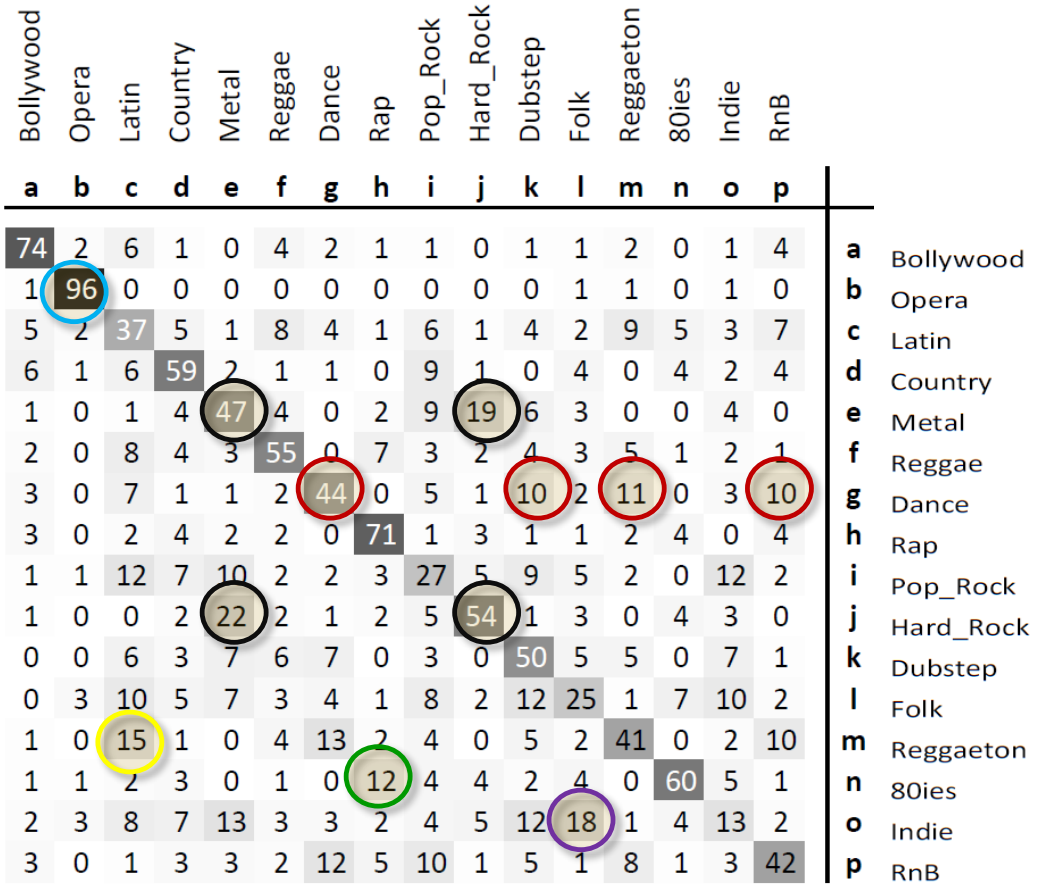
- **Improvements**
  - MVD-VIS => 2.94%
  - MVD-MM => 6.84%
  - MVD-MIX => 10.82%
- **Largest Improvement:**
  - TSSD (MVD-MIX) => 16.43%

- Better definition of weak genres
- Better sability with higher number of genres

			MVD-VIS			MVD-MM			MVD-MIX		
		Dim	SVM	KNN	NB	SVM	KNN	NB	SVM	KNN	NB
(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
(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	95.65	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

Confusion Matrix – MVD-Mix ImageNet

- Opera
  - visually good discriminable
- Metal / Hard_Rock
  - Visually Similar
  - Performance focused / Instruments
  - Fire / Pyrotechnic effects
- Dance / Dubstep / Reggaeton /RnB
  - Female Dancers / Artists
  - Club / Beach scenery
  - Cars
- Reggaeton / Latin
  - Reggaeton is Latin
  - Some videos share similar scenery / Villa
- Indie / Folk
  - Related genres
  - Visually simlar
- 80ies / Rap
  - Possible production effect
  - Old videos 80s / 90s



Evaluation 4: MVD-Themes Classification

- **Non-Audible Themes /** Cross-genre Classification
- **Eval 1:** Discriminating different Themes (TH)
  - Problematic for audio
  - Outperformed by Visual Concept Detection / (+36% accuracy, 37.5% audio-visual)
- **Eval 2:** Discriminating Themes from other Genres (Multi-Tagging)
  - Highly problematic for audio, especially with many genres
  - Outperformed by Visual Concept Detection / (+34.5% accuracy, 45.5% audio-visual)

+45.5%

	Audio-Only				Visual-Only				Audio-Visual			
Theme	**VIS**	**MM**	**MIX**	**TH**	**VIS**	**MM**	**MIX**	**TH**	**VIS**	**MM**	**MIX**	**TH**
Christmas				52.9				**88.9**				**90.4**
K-Pop				86.0				**91.7**				**90.0**
Protest Song				47.5				**75.5**				**77.5**
Broken Heart				54.9				**70.2**				**68.6**

Exercise: Tag that Song

- **Genre**: Punk
- **Style**: Pop Punk
- **Mood**: Happy
- **Theme**: **?**

Conclusions and Future Work

Conclusions

- Hypothesis: Existence of music related visual information

- **Which visual features are able to capture task related information?**
  - **Faces:** Improve Artist Identification Precision relatively by 27% over the audio-only baseline
  - **Color:** Predict Genre up to 50.1% accuracy. Slight improvement with audio-visual combination
  - **Objects:** Predict Genre up to 74.4% accuracy. Huge audio-visual improvements.

  - **Conclusions:**
    - Music related visual media contains music related visual information
    - This information can already be extracted using low-evel features
    - High-level visual objects directly relate to music concepts (e.g. instruments, cowboy-hat, etc.)
    - Visual Objects are outperforming low/mid-level features (e.g. MFCC, TRH)

Conclusions

- **Is it possible to verify concepts within the production process?**
  - MVD-VIS dataset: high acoustic coherence (93.7%)
    - ➔ Visual concept detection (74.4%). Confusion-Matrix provides further indications.
    - ➔ Highly discriminative visual patterns per genre
    - ➔ deliberate visual accentuation verified
  - Contemporary Country Videos, movement to warm orange tones
    - ➔ Average values of red and yellow are significantly highest

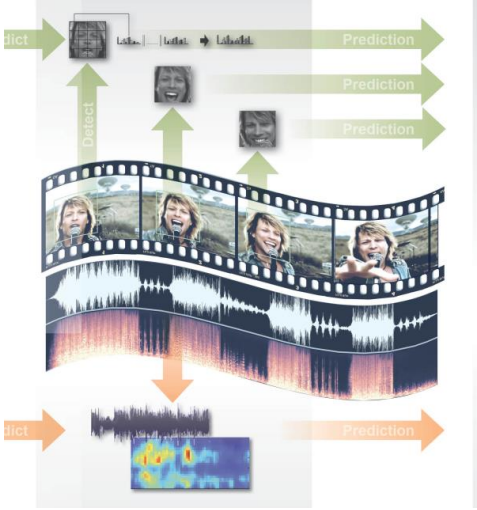
- **Verification of Music related Visual Stereotypes**
  - Common stereotype of „darkness“ of Heavy Metal
    - ➔ significantly more black pixels in MVD-MIX Dataset
  - Country Music and Cowboys
    - ➔ Cowboy-hat most salient visual object / Farming vehicles (pickup truck, tractor, etc.)
  - Over-sexualization of contemporary Dance music
    - ➔ Swim- and underwear, body parts, pole dancing, most salient visual objects
  - Heavy Metal and Fire
    - ➔ Matchstick (no „Fire“ concept in ImageNet)

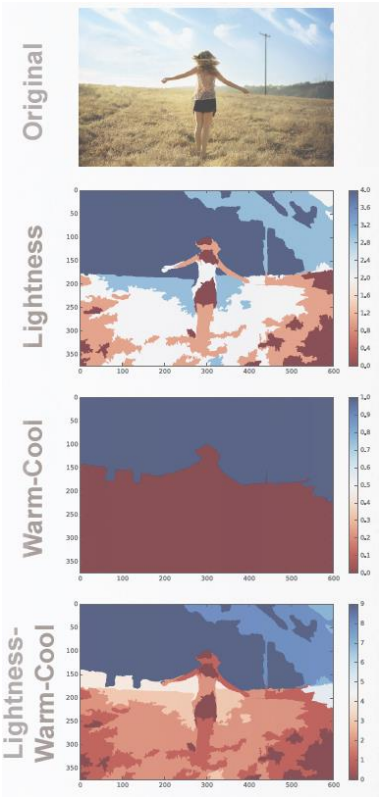
Awards

- DCASE 2016: Winning contribution (w. Thomas Lidy)
  - Domestic Audio Tagging task
- MIREX 2016: Winning contributions (w. Thomas Lidy)
  - Classical Composer Identification
  - Genre Classification (Latin)
  - Mood Classification
  - KPOP genre classification
- FMT 2016: Best Paper Runner-up
  - Alexander Schindler, Thomas Lidy, and Andreas Rauber. Comparing shallow versus deep neural network architectures for automatic music genre classification.
- CBMI 2019: Best Student Paper
  - Alexander Schindler and Peter Knees. Multi-Task Music Representation Learning from Multi-Label Embeddings.

Research Exploitation

- Direct/Indirect exploitation in 6 Research Projects
  - Europeana Sounds, FLORIDA, VICTORIA, 3 Innovationsschecks, 2 Innovations-Lehrgänge
  - 2 proposals as consortia lead pending
  - 3 currently in planning for open calls
- Direct/Indirect contributions to academic lectures
  - Information Retrieval, Advanced Information Retrieval, Intelligent Audio and Music Analysis
- Contributions to Community
  - 2 Tutorials (ISMIR 2018, ML-Prague 2017)
  - 1 Lecture at Summer School
  - Github Repositories with Tutorials, Example code





Thank You!

Alexander Schindler

http://www.ifs.tuwien.ac.at/~schindler

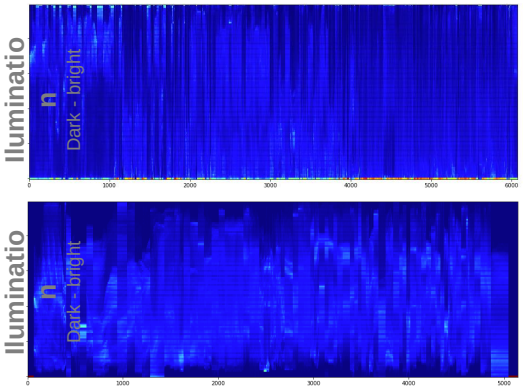
stage	0.3162
electric guitar	0.1169
bassoon	0.0649
accordion	0.0611
drumstick	0.0386
microphone	0.0313
marimba	0.0276

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



APPENDIX

Awards

- DCASE 2016: Winning contribution (w. Thomas Lidy)
  - Domestic Audio Tagging task
- MIREX 2016: Winning contributions (w. Thomas Lidy)
  - Classical Composer Identification
  - Genre Classification (Latin)
  - Mood Classification
  - KPOP genre classification
- FMT 2016: Best Paper Runner-up
  - Alexander Schindler, Thomas Lidy, and Andreas Rauber. Comparing shallow versus deep neural network architectures for automatic music genre classification.
- CBMI 2019: Best Student Paper
  - Alexander Schindler and Peter Knees. Multi-Task Music Representation Learning from Multi-Label Embeddings.

Research Exploitation

- Direct/Indirect exploitation in 6 Research Projects
  - Europeana Sounds, FLORIDA, VICTORIA, 3 Innovationsschecks, 2 Innovations-Lehrgänge
  - 2 proposals as consortia lead pending
  - 3 currently in planning for open calls
- Direct/Indirect contributions to academic lectures
  - Information Retrieval, Advanced Information Retrieval, Intelligent Audio and Music Analysis
- Contributions to Community
  - 2 Tutorials (ISMIR 2018, ML-Prague 2017)
  - 1 Lecture at Summer School
  - Github Repositories with Tutorials, Example code

Preamble

- AIT High Performance Image Processing Group (4 years)
- Acquired substantial Image Processing Know How
- Project participation
  - 2 Projects (Document Image Quality Inspection)
- Publications
  - 9 co-author
  - 1 first author
- Changed to Data Science / Information Retrieval Group in 2014

Timeline Albumarts

- ISMIR 2011
  - MSD
- ISMIR 2012
  - Audio Samples
  - Extracted Feature-sets
  - Ground-truth assignments (genre, styles)
- AMR 2012
  - Evaluation of Echonest Features
- 2012
  - MSD Album Arts Dataset (MAAD)
- DCASE 2016/2017
  - DNN for Audio (Winning Contribution)
- MIREX 2016
  - DNN for Music (Winning Contribution)
- FMT 2016
  - First large-scale audio benchmarks for the MSD (Best Paper nominee)
- CBMI 2019
  - Ground-truth assignments, music representation learning (Best Student Paper Award)

Timeline Music Videos

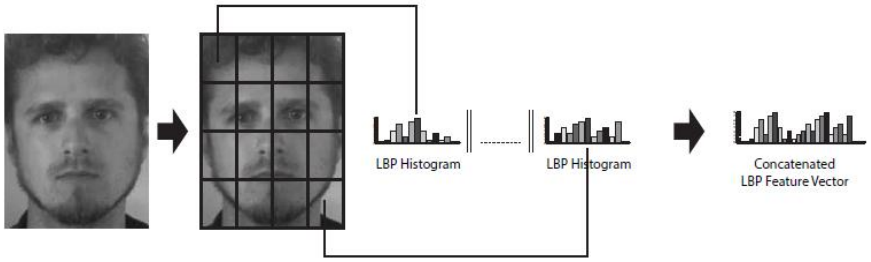
- CMMR 2013
  - Artist Identification
- DLfM 2014
  - Opportunities of MVIR
- 2014
  - Full Music Video Dataset
- ECIR 2015
  - Colors in Music Videos
- TIST 2016
  - Music related visual concepts
- MMM 2019
  - Shot-Boundary detection

Obstacles of Music Distribution

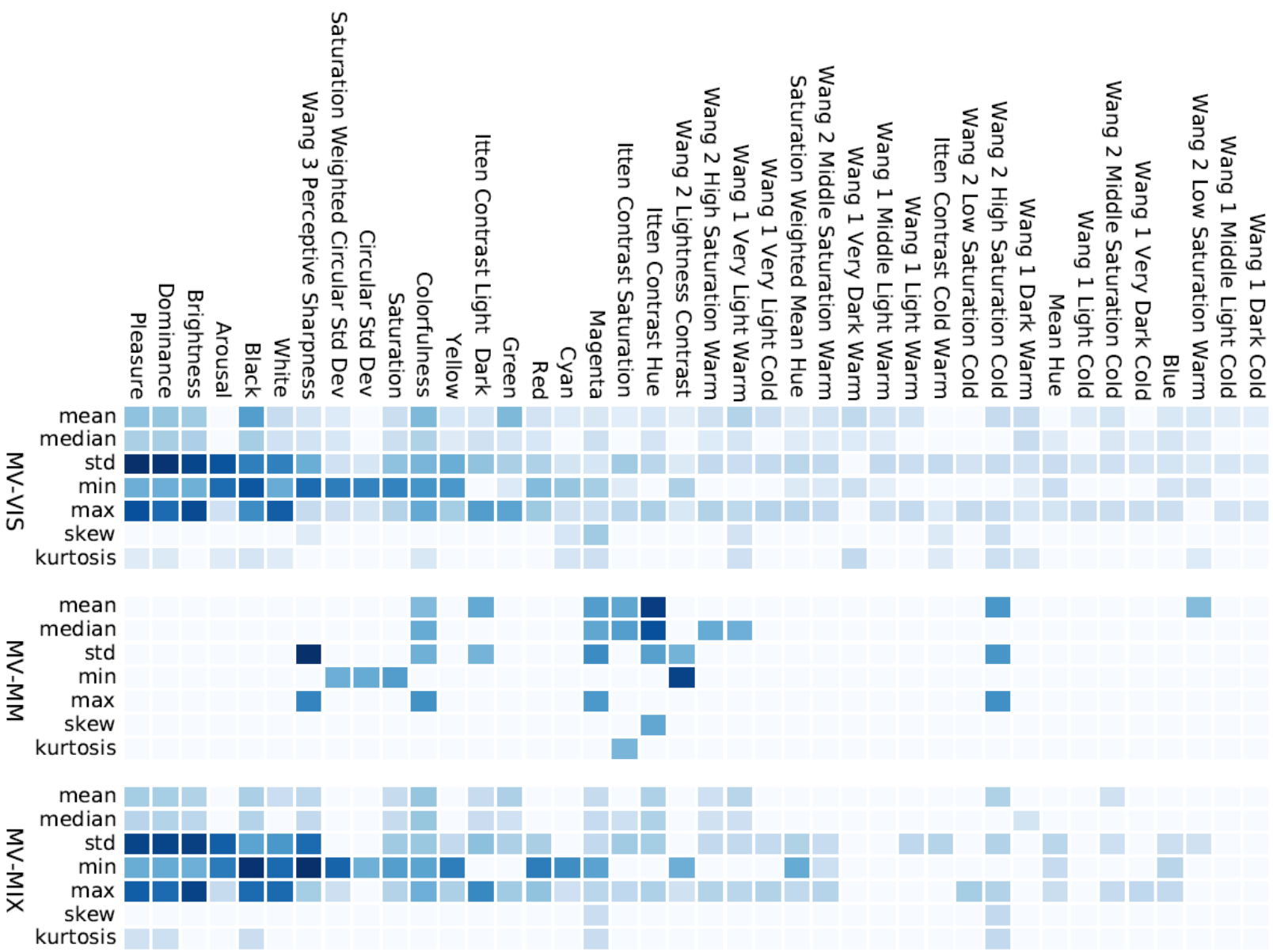
- Music is an acoustic phenomenon
- To experience => listen to it
- No problem in smart phone era
  - Spotify
  - Youtube
- Problem before (two decades ago)
  - CD
  - LP
  - Cassette

Face Recognition

- Local Binary Patterns (LBP)
  - Texture descriptor.
    - Combination of structural and statistical texture analysis
  - Robust against different facial expressions, illumination changes and aging subjects.
  - Thresholding 3x3 neghborhood by center pixel ➔ result = 8bit numeric label
  - Histogram of 256 labels / 8 Regions to retain spatial information
  - Nearest Neighbors based recognition w. Chi Square dissimilarity.



Low-Level Feature Importance



Types of Music-related Visual Media

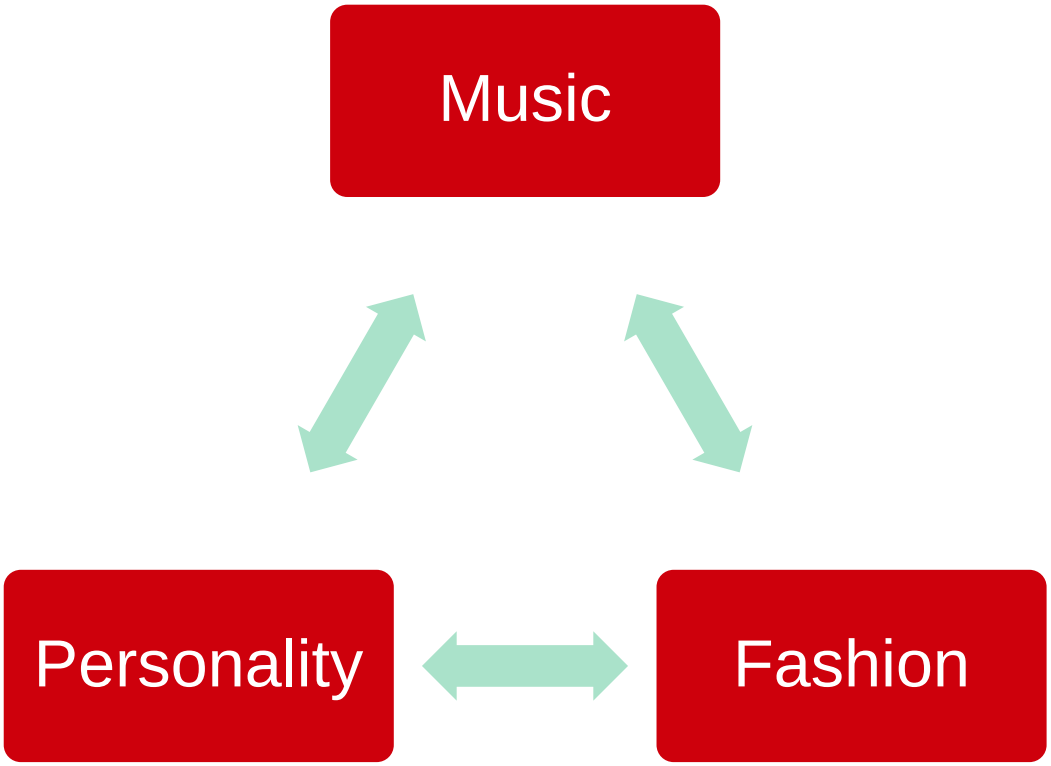
- Music Advertising
- Expressing music reference
  - Style of event
  - Personal taste

Fashion

Party Flyers

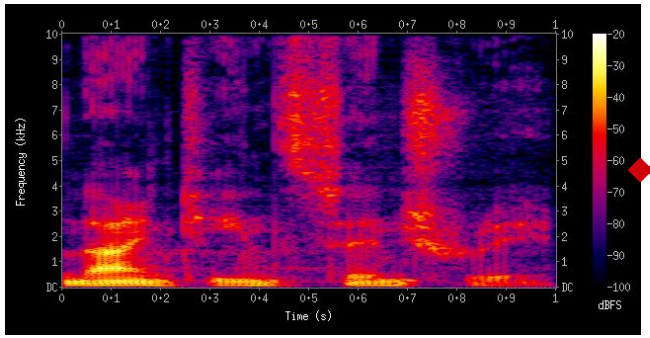
Album Covers

Music-Semantics Complexity

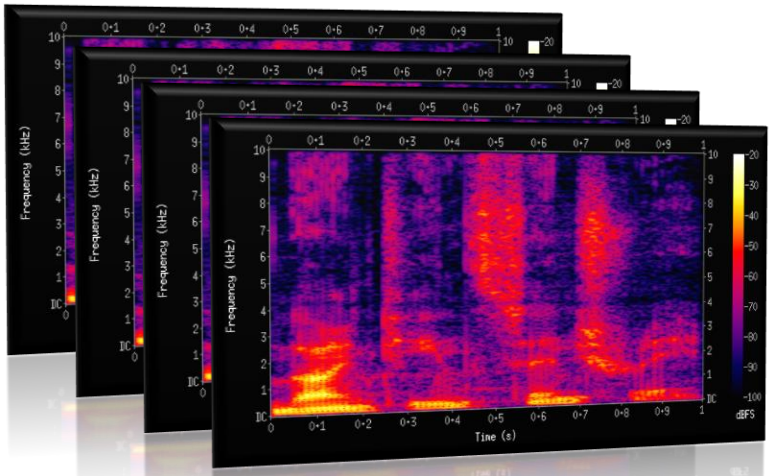


Music-Semantics Complexity

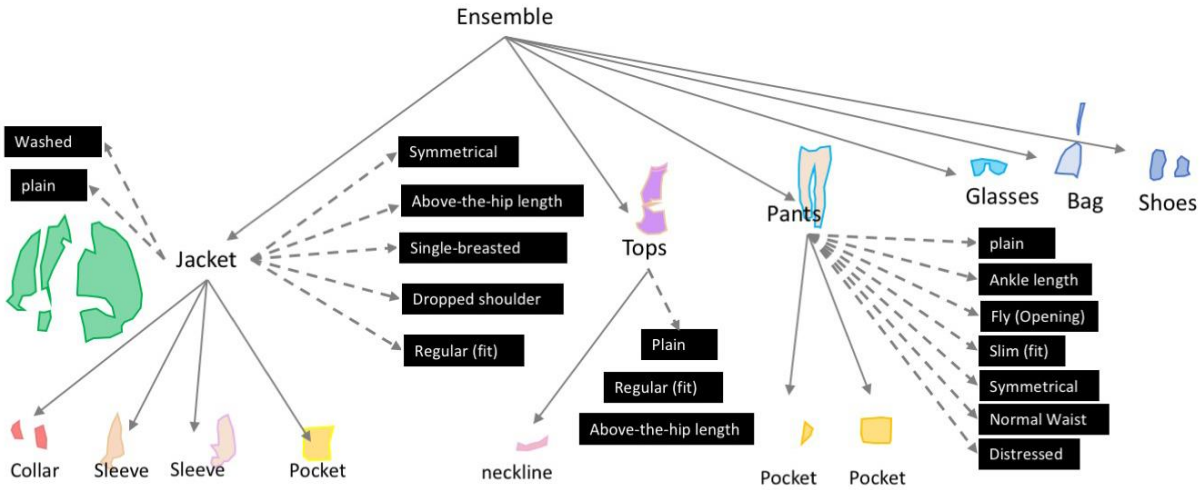
Music-Semantics Complexity

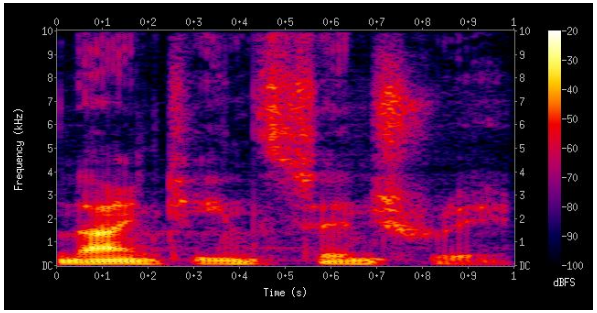


Punk, Autumn, Sunset,
Sad, Birthday, Liked by
my firends, for work,
with firends, heavy, …



Future Work


Ensemble
Washed
plain
Jacket
Symmetrical
Above-the-hip length
Single-breasted
Dropped shoulder
Regular (fit)
Collar
Sleeve
Sleeve
Pocket
Tops
Plain
Regular (fit)
Above-the-hip length
neckline
Pants
Pocket
Pocket
Glasses
Bag
Shoes
plain
Ankle length
Fly (Opening)
Slim (fit)
Symmetrical
Normal Waist
Distressed


Frequency (kHz)
Time (s)
dBFS

*"**Albums should be** as **bold** and **dashing** as we can make them; they should **stand out** in dealers' windows **screaming for attention**, yet always **reflecting** the **spirit of the music** inside. Color should be violent and strong. Copy should be pared to a minimum, and each album should reflect the quality of the Columbia name"*

— Pat Dolan, CBS, 1940

Example

- **Genre**: Punk
- **Style**: Pop Punk
- **Mood**: Happy
- **Theme**: **?**

$a1$	Chroma	19.64	14.68	12.08
$a2$	MFCC	37.02	26.60	27.11
$a3$	SSD	65.11	44.64	38.92
$a4$	RP	63.19	43.06	41.39
$a5$	TRH	46.61	33.02	35.70
$a6$	TSSD	66.19	47.40	44.22
$a7$	$a4$+$a6$	74.64	53.06	**48.54**
$a8$	$a4$+$a3$+$a5$	72.73	49.88	43.65
$a9$	$a4$+$a3$	74.38	51.60	43.52
$a10$	$a4$+$a5$+$a6$	**75.91**	**54.16**	48.32

	Audio-Only			
Theme	**VIS**	**MM**	**MIX**	**TH**
Christmas	67.6	36.7	29.5	52.9

- 1600 Tracks
- 16 Genres
- 46 Christmas Songs

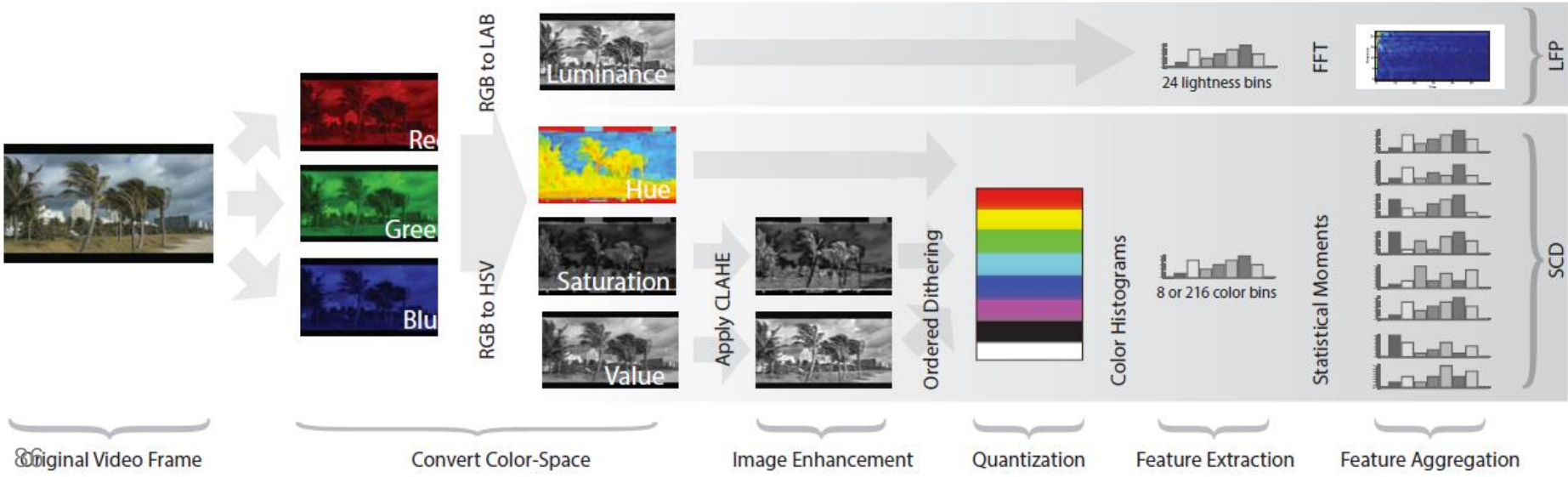
Artist Identification

Results

Artist Name	Audio			Video			Audio-Visual Ensemble		
	Prec	Recall	f1	Prec	Recall	f1	Prec	Recall	f1
Aerosmith	0,33	0,52	0,39	0,14	0,33	0,20	0,36	0,57	0,44
Avril Lavigne	0,50	0,45	0,47	0,62	0,25	0,36	0,64	0,45	0,53
Beyonce	0,33	0,26	0,29	0,28	0,42	0,33	0,55	0,32	0,40
Bon Jovi	0,28	0,36	0,32	0,20	0,04	0,07	0,24	0,27	0,25
Britney Spea	0,32	0,33	0,33	0,16	0,17	0,16	0,34	0,42	0,38
Christina Agu	0,48	0,71	0,57	0,18	0,43	0,26	0,33	0,50	0,40
Foo Fighters	0,41	0,47	0,44	0,00	0,00	0,00	0,62	0,53	0,57
Jennifer Lop	0,22	0,24	0,22	0,33	0,14	0,20	0,27	0,19	0,22
Madonna	0,27	0,28	0,24	0,50	0,12	0,19	0,30	0,24	0,27
Maroon 5	0,20	0,10	0,13	0,12	0,80	0,20	0,35	0,70	0,47
Nickelback	0,55	0,38	0,44	1,00	0,18	0,30	0,58	0,44	0,50
Rihanna	0,29	0,19	0,23	0,40	0,10	0,15	0,75	0,14	0,24
Shakira	0,44	0,40	0,41	0,25	0,21	0,23	0,28	0,65	0,39
Taylor Swift	0,60	0,32	0,41	0,50	0,06	0,10	1,00	0,16	0,27
	0,37	0,36	0,35	0,33	0,23	0,20	0,47	0,40	0,38

Acoustic & Visual Features

	Short Name	#	Description
Audio	Statistical Spectrum Descriptors (SSD)	168	Statistical description of a psycho-accoustic transformed audio spectrum
	Rhythm Patterns (RP)	1024	Description of spectral fluctuations
	Rhythm Histograms (RH)	60	Aggregated Rhythm Patterns
	Temporal SSD and RH		Temporal variants of RH (TRH #420), SSD (TSSD #1176)
	MFCC	12	Mel Frequency Cepstral Coefficients
	Chroma	12	12 distinct semitones of the musical octave
Visual	Global Color Statistics	6	mean saturation and brightness, mean angular hue, angular deviation, with/without saturation weighting
	Colorfulness	1	colorfulness measure based on Earth Movers Distance
	Color Names	8	Magenta, Red,Yellow,Green,Cyan Blue, Black, White
	Pleasure, Arousal, Dominance	3	approx. emotional values based on brightness and saturation
	Itten Contrasts	4	Contrast of Light and Dark, Contrast of Saturation, Contrast of Hue and Contrast of Warm and Cold
	Wang Emotional Factors	18	Features for the 3 affective factors by Wang et al. [281]
	Lightness Fluctuation Patterns	80	Rhythmic fluctuations in video lightness

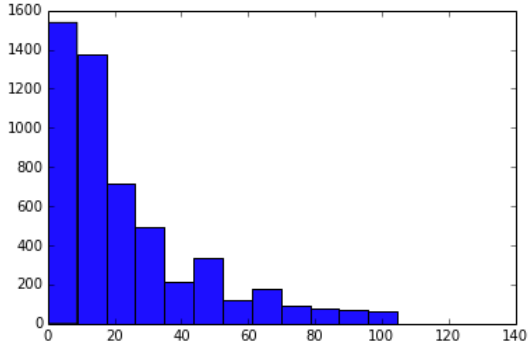


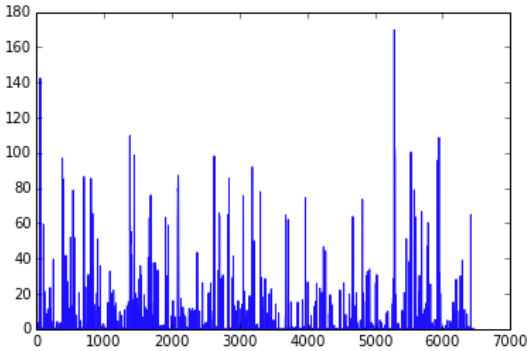
Low-Level features Results

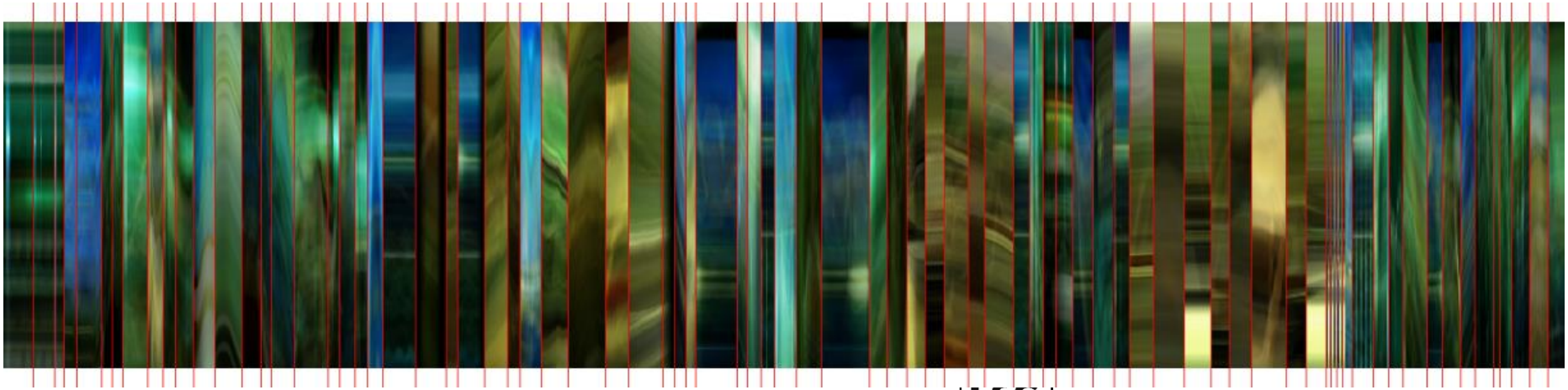
(b) Low-level Color and Affect related Image Features											
$v_{co}1$	LFP	60	33.21	23.59	25.45	20.38	16.74	16.46	16.93	11.71	13.36
$v_{co}2$	CF	7	34.89	25.49	31.50	21.84	17.06	20.41	18.53	11.92	16.49
$v_{co}3$	IC	28	36.80	27.55	27.51	24.83	**19.43**	19.68	21.44	13.54	12.66
$v_{co}4$	GEV	21	39.45	**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**

Shot Detection with Clustering

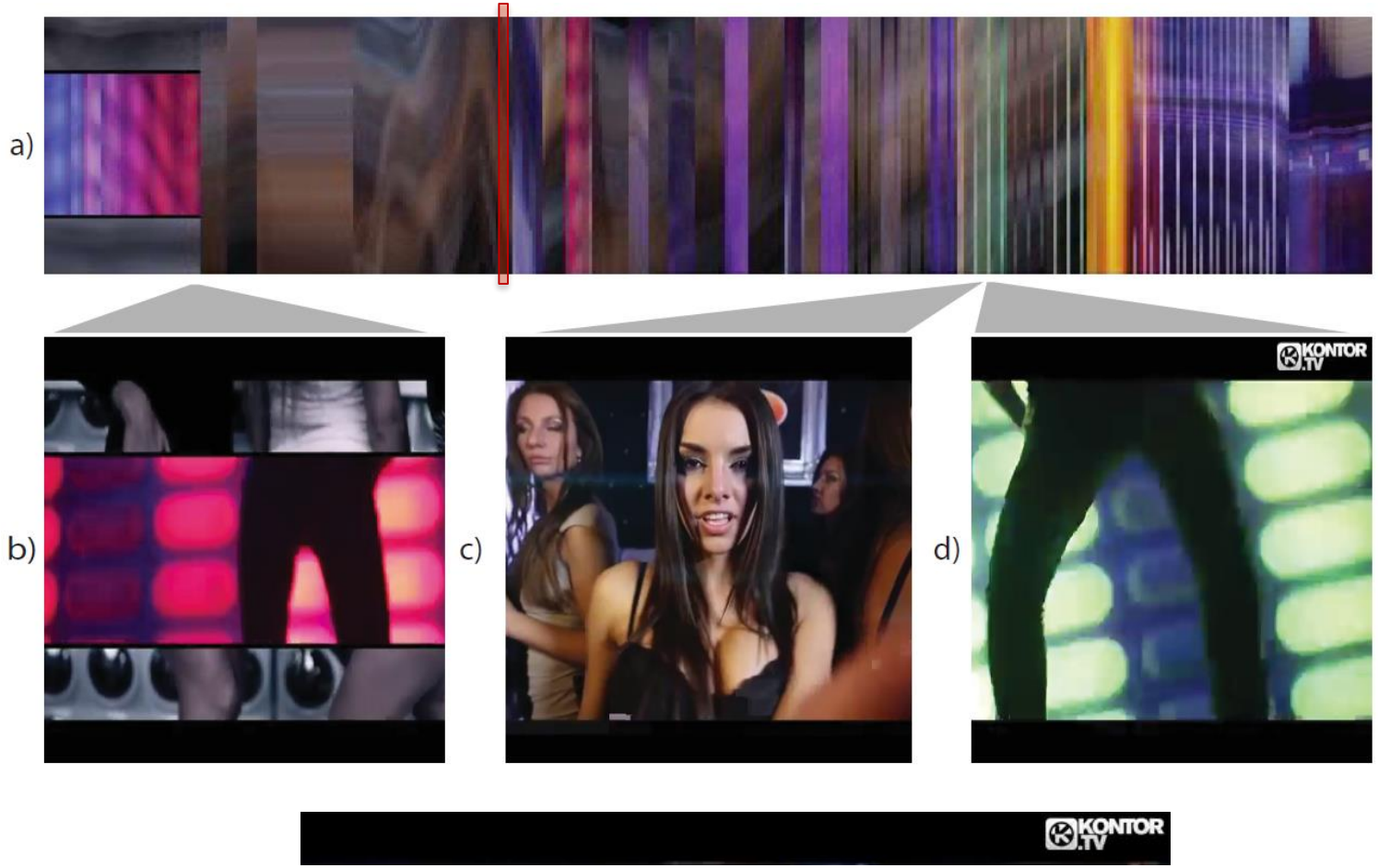
- Chasanis, Vasileios T., Aristidis C. Likas, and Nikolaos P. Galatsanos. "Scene detection in videos using shot clustering and sequence alignment." *IEEE transactions on multimedia* 11.1 (2008): 89-100.







Frame Swapping



Skip Frames

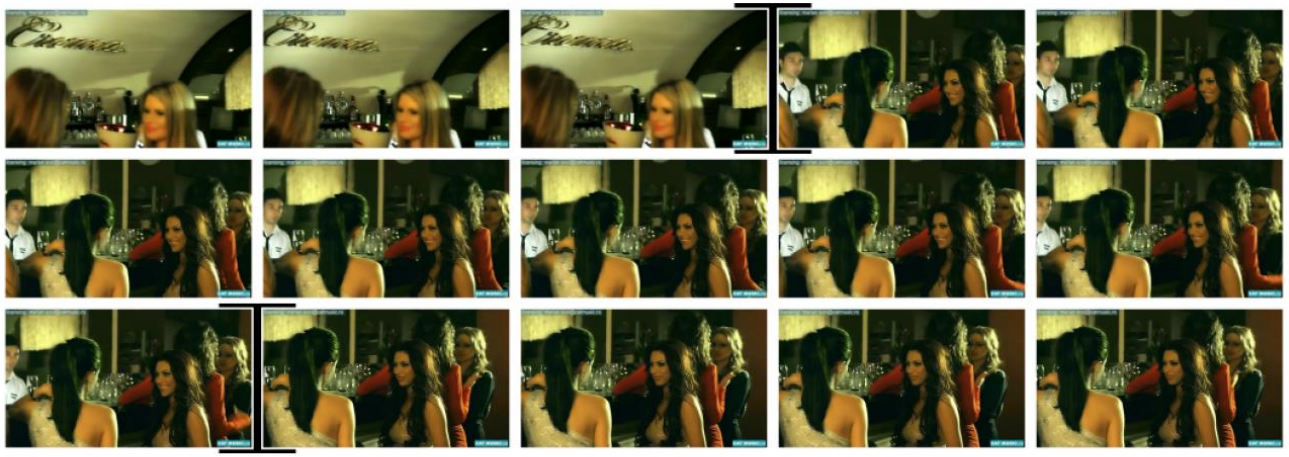


FIGURE 7.1 – Example of **Skip Frames**. Replaying these frames with the 25 frames per second of the original video reveals, that this scene was recorded in a single camera movement, starting from the left corner of the bar and panning to the right until the focus is on two women. The depicted frames show that a large segment was skipped after frame number 3 and a small segment after frame number 11.

(Virtual) One-Shot Music Videos

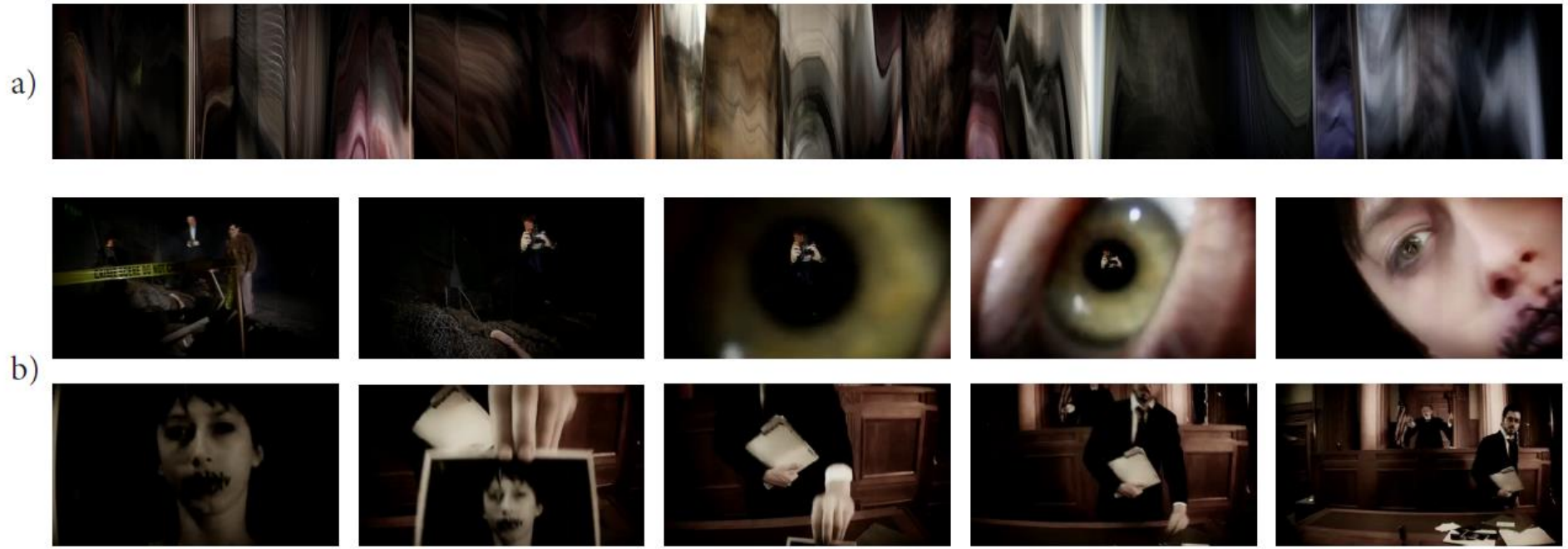
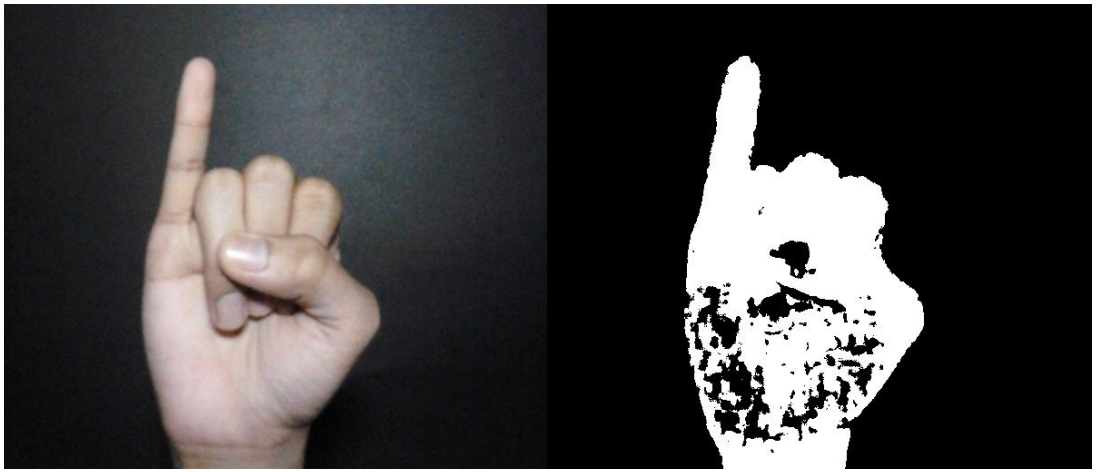


FIGURE 7.8 – Example 3: One-shot Music Video. a) Mean-color-bar depicting that there are no sharp cuts in the video. b) example video frames of the starting sequence of the music video. These frames demonstrate how zooming out is used to transition between scenes.

(Too) Complex Features

Skin Ratio Detection



By Color Range

- H = [0,50]
- S = [0.20,0.68]
- V = [0.35,1.0]

- Works on natural color distributions
- Difficult to parametrize globally
- Not invariant to Illumination changes

Other Approaches

- Grey-World Theory
- Retinex
- Skin Probability Maps

Chakraborty, Biplab Ketan, M. K. Bhuyan, and Sunil Kumar. "A Weighted Skin Probability Map for skin color segmentation." *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2016.

Artificially Distorted Color Distributions

- Post-processing
  - Filter (Color, Blur, etc.)
- White-Balancing
  - Single perturbation
    - Linear mapping
  - Multiple perturbations difficult
    - Non-linear mapping

Color Quantization Complexity

a) Original image with **Letterboxing**

b) Pre-processing: remove letterboxing

c) Color Quantization: Naive nearest neighbor match

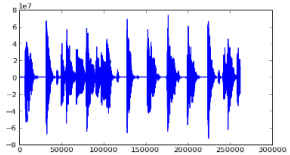
d) Ordered Dithering (OD)

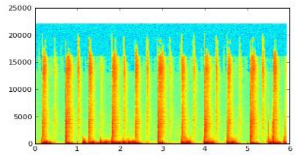
e) OD + **enhanced brightness (CLAHE)**

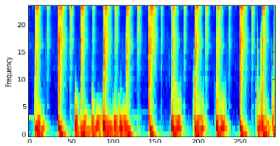
f) OD + **enhanced brightness, saturation**

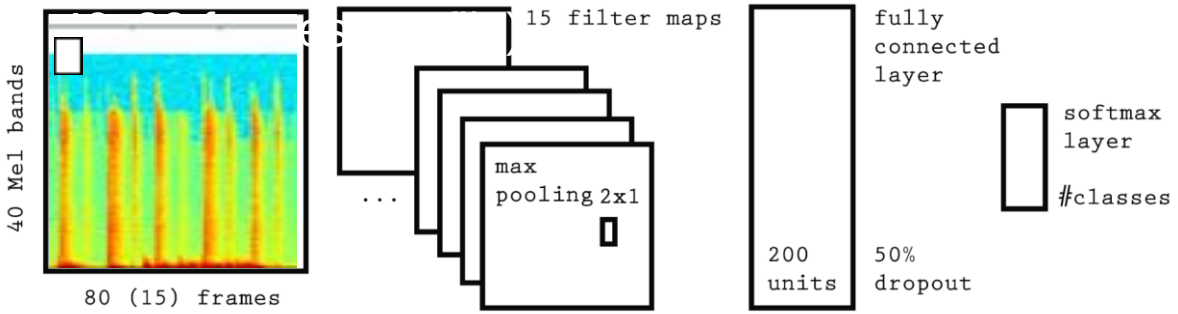
Deep Learning for Audio Retrieval

Pre-Processing: Waveform → Spectrogram → 40 Mel bands → Log scale

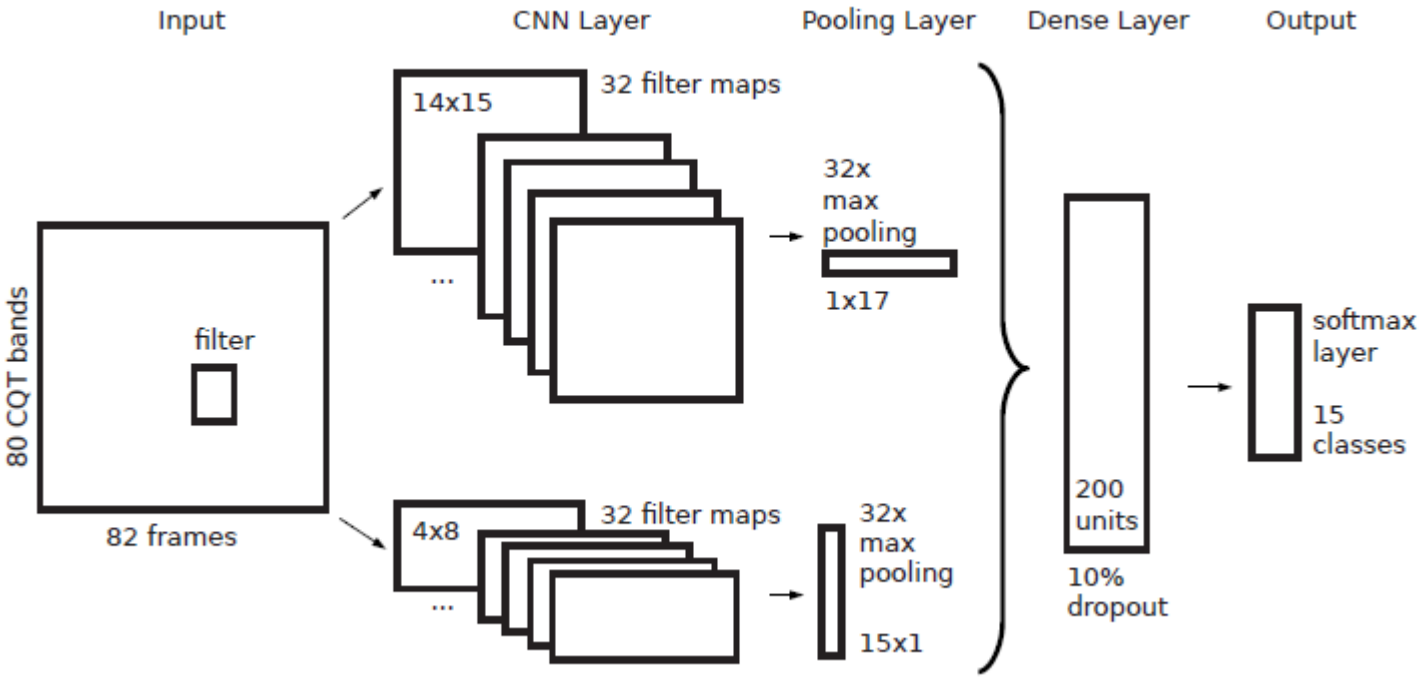






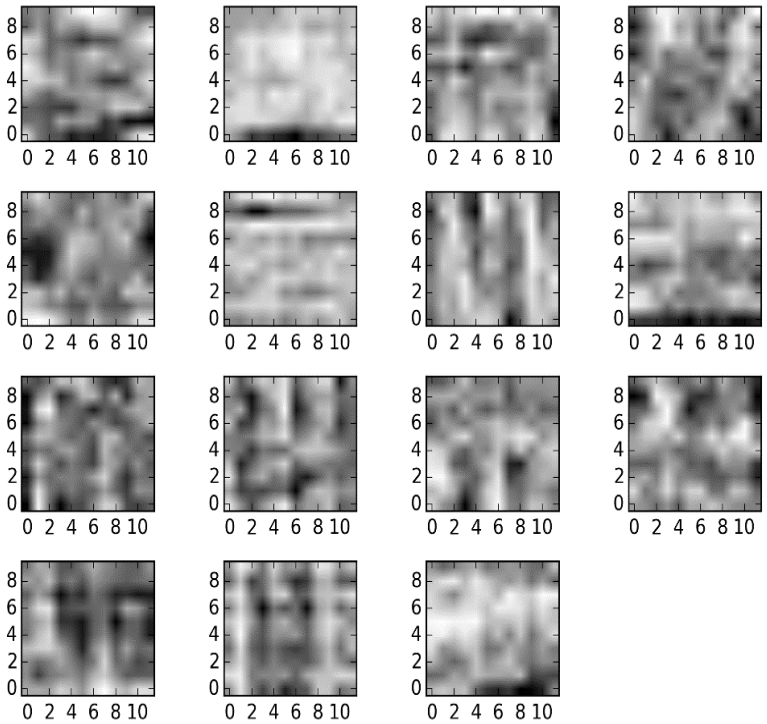


Convolutional Neural Network

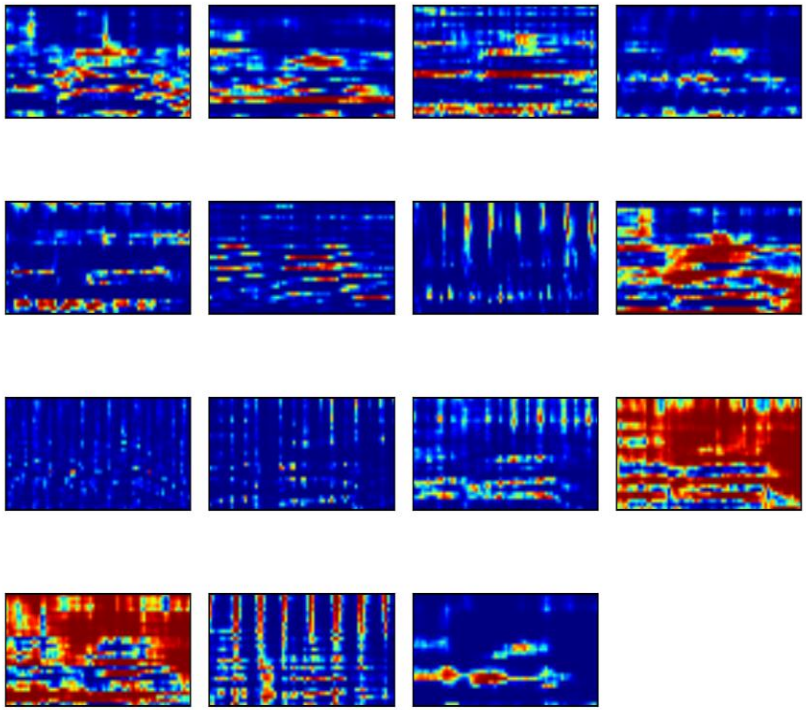


Audio Processing

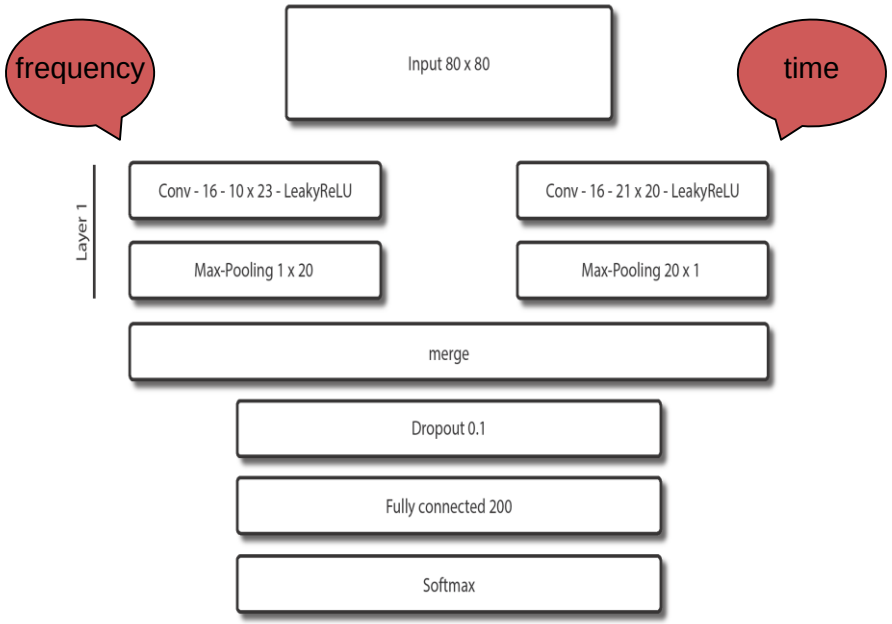
Learned Filter Weights

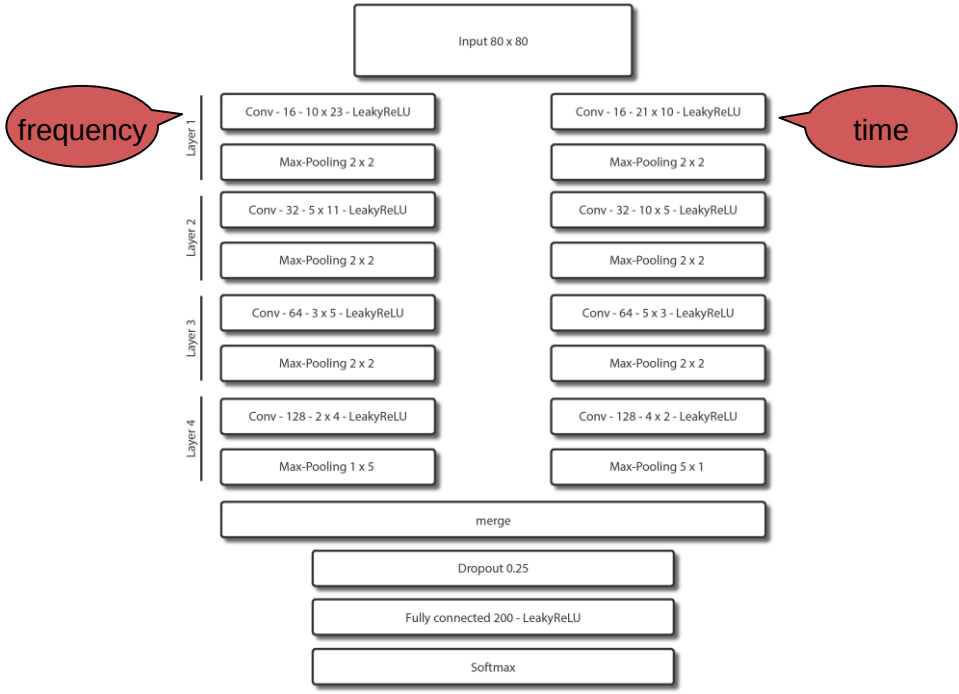


Convolved Spectrograms



Deep vs. Shallow





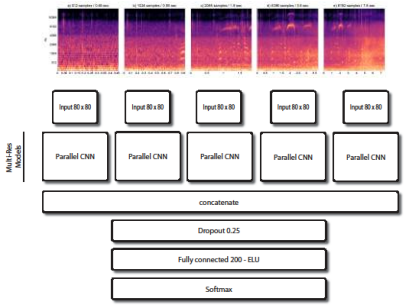
	100 epochs		200 epochs	
	Shallow	**Deep**	**Shallow**	**Deep**
GTZAN	78.1	**78.6**	**80.8**	80.6
ISMIRgenre	**85.5**	84.1	84.9	**85.1**
Latin	92.4	**94.4**	93.5	**95.1**
MSD	63.9	**67.2**	/	/

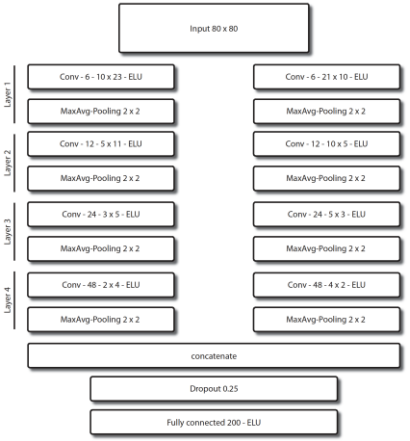
Success

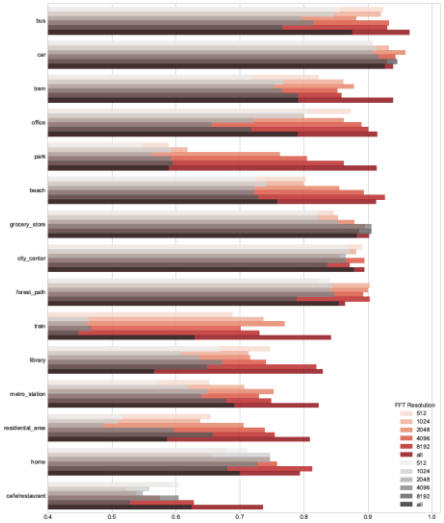
- DCASE 2016
  - Task winner: Domestich Audio Tagging
- MIREX 2016
  - Classical Composer Identification
  - Latin Genre Classification
  - Music Mood Classification
  - KPOP Genre (Annotated by Korean Annotators) Classification
  - KPOP Genre (Annotated by American Annotators) Classification
  - KPOP Mood (Annotated by Korean Annotators) Classification
  - KPOP Mood (Annotated by American Annotators) Classification

Multi-resolution Convolutional Neural Networks

- Netzwerk Architecture







fft win size	instance raw	grouped raw	instance augmented	grouped augmented
512	64.14 (2.84)	70.32 (2.96)	69.06 (4.33)	76.63 (4.44)
1024	66.32 (2.58)	71.27 (3.06)	71.70 (5.46)	77.06 (5.46)
2048	66.83 (1.52)	70.23 (1.99)	76.24 (2.53)	80.46 (3.30)
4096	69.50 (2.83)	71.92 (3.23)	79.20 (3.03)	81.66 (3.29)
8192	69.66 (2.58)	71.47 (2.95)	82.26 (2.40)	83.73 (2.63)
grouped single		73.12		83.19
multi-res	72.23 (4.15)	74.30 (4.81)	85.22 (2.11)	87.29 (2.02)
multi-res do	69.39 (2.77)	72.05 (3.26)	82.51 (2.37)	86.04 (3.03)

Audio Analysis for Security Related Porjects

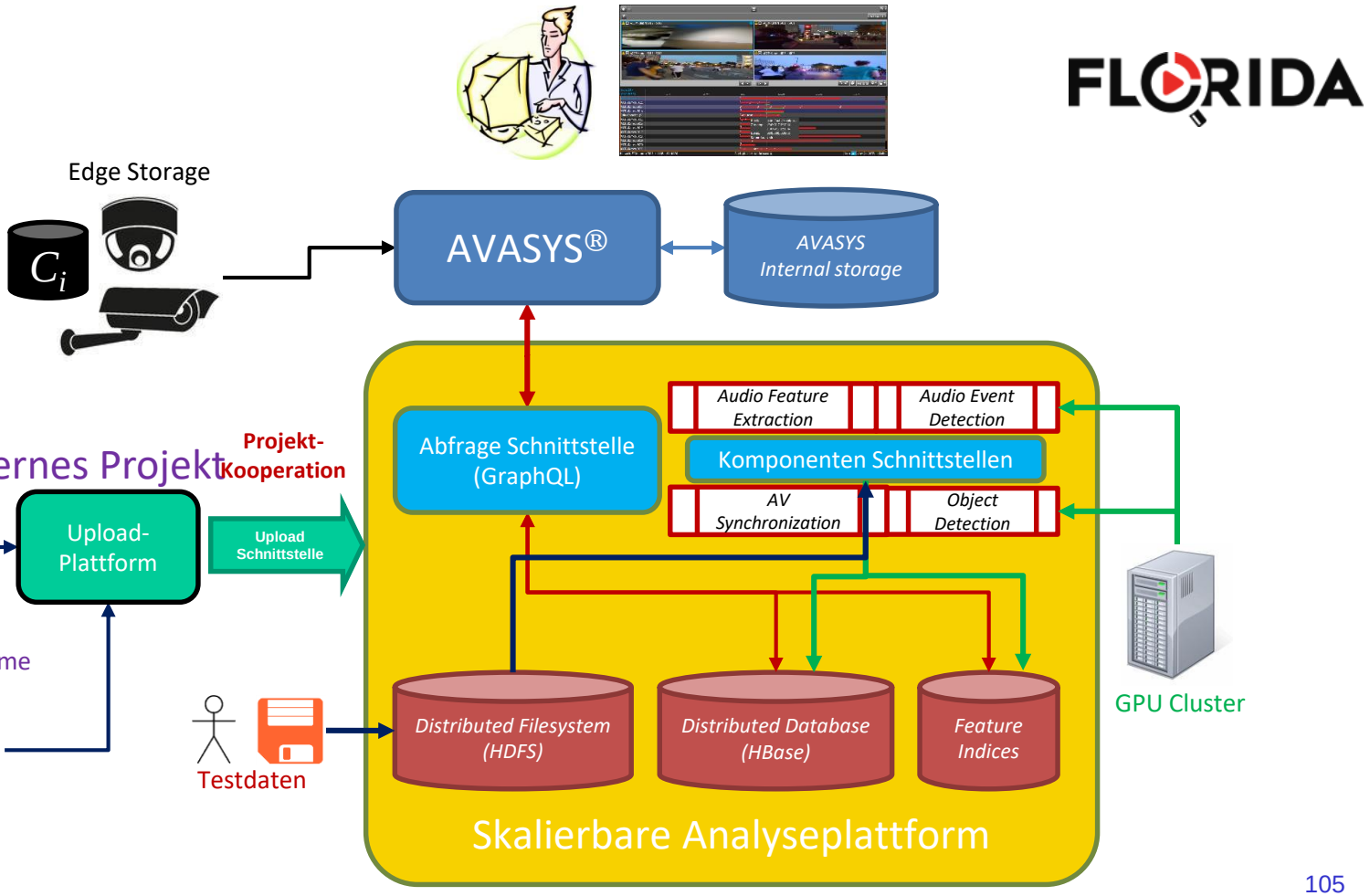



Audio Event Detection

Audio Synchronization

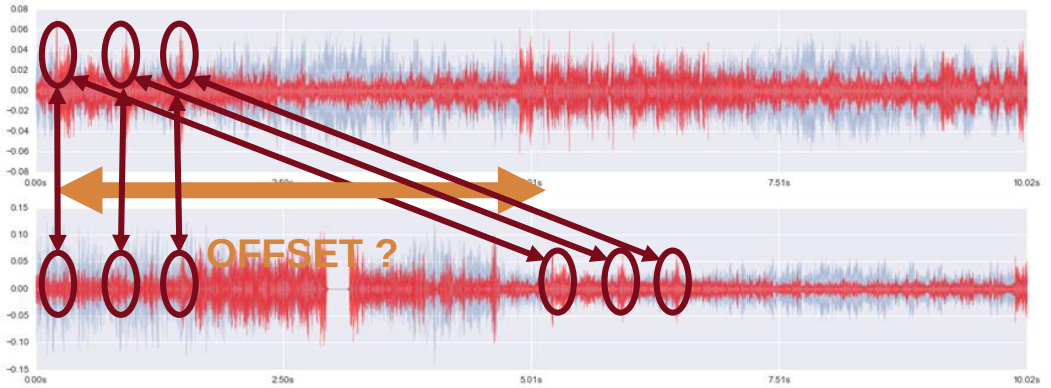
Audio Similarity Retrieval

System Architektur



Audio-based Video-Synchronisation

- Time-Information in video-metadata often unreliable (e.g. personal camera)
- Salient, loud accoustic events create a specific pattern
  - Z.B. honking, screams, gunshots, etc.
  - These are independant from: angle of view, camera position
- Synchronisation by overlapping audio patterns



Audio-Event Detection und Classification

- Query the collection by: „all videos with audible gunshots“
- Digital Audio Analysis
  - Detecting Audio Events
  - Segmentation of audio-track (start – stop of event)
- Audio Feature-Extraction
  - Timbre, loudness, length, spectral envelope, etc.
- Machine Learning
  - Training models for automatic classification

Searching Videos with similar audio content

- Example:
  - Suspicious person with back to camera
  - Passing car honks twice
  - Searching for videos with similar audio pattern (two consecutive honks)
- Audio feature extraction
- Searching for segments with similar audio feature values

Ergebnis-Liste

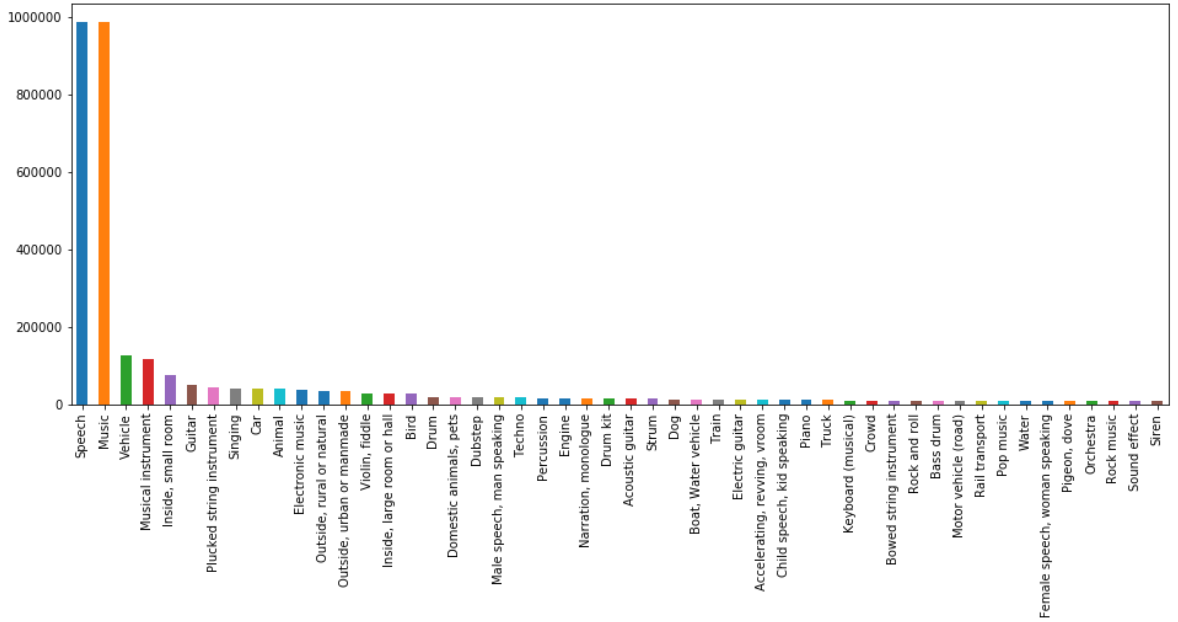
90% ähnlich

85% ähnlich

70% ähnlich

AudioSet

- 2.084.320 YouTube videos
- 527 labels
- Human-labeled
- Verified annotations
- Sounds heard within Segment

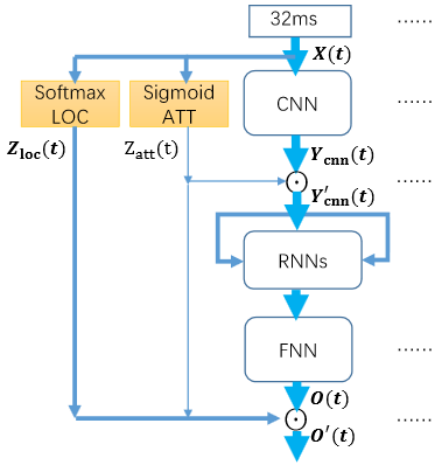


Audio Event Detection

Recurrent-Convolutional Neural Networks

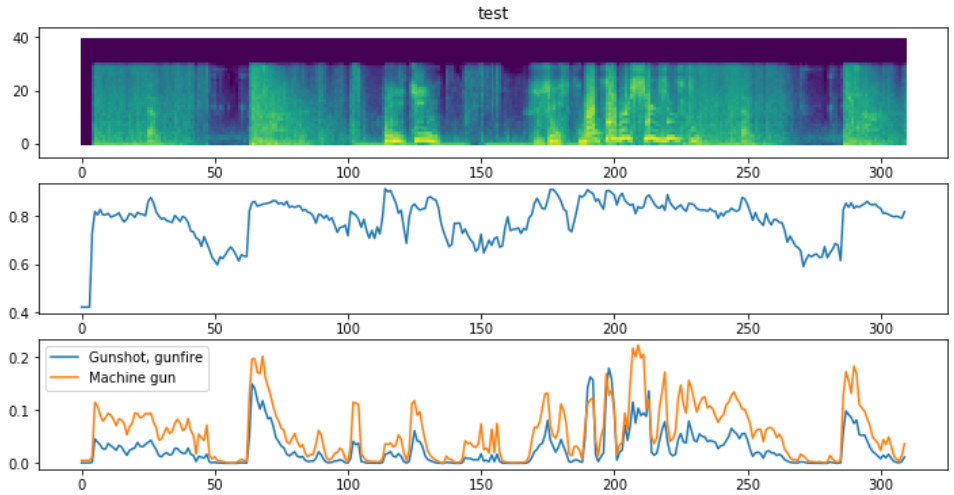
1. Attention Layer
   - Lernt relevante von kontextuell irrelevanter Informatio unterscheiden
   - Feedback-Schleife zu RCNN

2. Localization Layer
   - Zeitlich verteilte Fully Connected Layers
   - Bestimmen das Label für jedes Segment / Zeitfenster



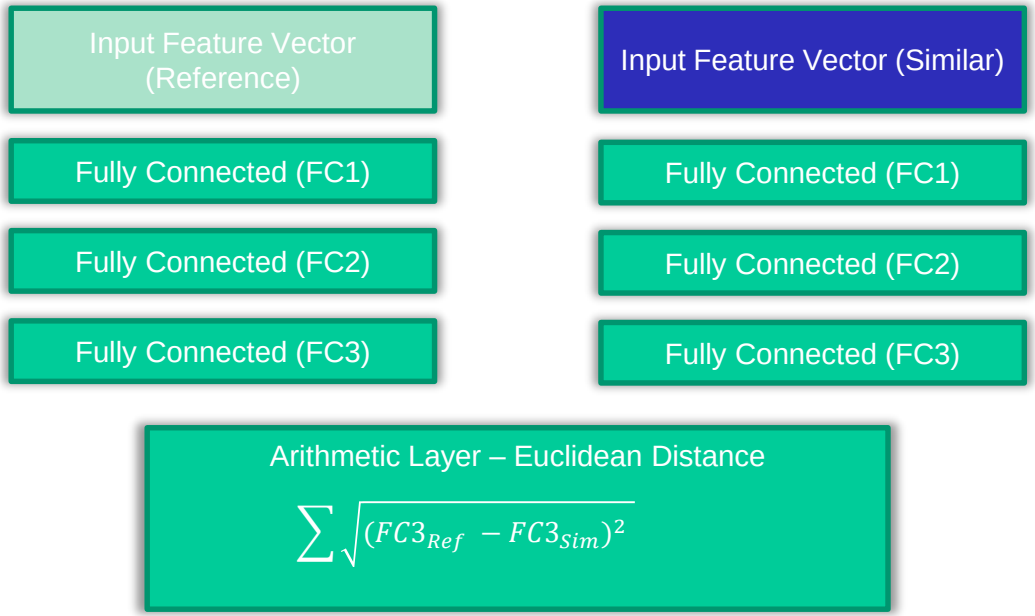
Ergebnisse

- Beispiel: AudioSet (Zufälliges Beispiel)
  - Attention Layer
    - Ignoriert „stille“ Sequenzen
  - Schüsse sehr gut lokalisiert



Siamese Networks

- Semantische Ähnlichkeit zwischen zwei Input Vektoren
- Shared Deep Neural Network layers
- Training Prozess:
  - Aufeinanderfolgende Sequenz von ähnlichen und verschiedenen Vektor-Paaren
  - Contrastive-Loss
  - Distanz zwischen
    - Ähnlichen minimieren
    - Nicht ähnlichen maximiern



FLORIDA Demo

Thank You!

Alexander Schindler
Department of Information Systems and Engineering
Vienna University of Technology
http://www.ifs.tuwien.ac.at/~schindler