An Audio-Visual Approach to Music Genre Classification through Affective Color Features

Alexander Schindler[1], Andreas Rauber[2]

Motivation

The *"look of music"* became an important factor in people's appreciation of music. The rise of music video added an additional semantic layer to a music track. We grew accustomed to a visual vocabulary that is specific for a music style in a way that the genre of a music video can often be predicted by sight only (see Figure 1). By augmenting music information retrieval technologies with solutions emerging from the video retrieval domain open research challenges could be addressed that are currently problematic to solve through audio content analysis only (e.g., classifying Christmas songs).

Rise in Music Video Importance

In 2011 we conducted a survey among music industry's decision makers and stakeholders which showed that YouTube is considered to be the leading online music service. Other studies revealed that *'watching music videos on computer'* was the most mentioned consuming activity of a survey among 26,644 online consumers in 53 international markets. Music videos are also identified as passive triggers to active music search. Their visual aspects can strongly influence the amount of attention paid to a song.

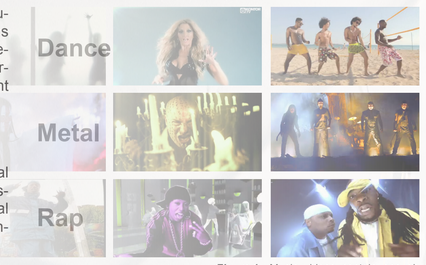

Figure 1: Music videos contain enough information to estimate their genres by sight.

The Music Video Dataset

The Music Video Dataset (MVD) is a collection of carefully selected music videos. It consists of two subsets that can be combined to one bigger data-set. Each set was assembled to serve different tasks. A set of quality and stratification criteria has been strictly applied to the selection process.

Due to copyright restrictions it is not possible to redistribute music videos or audio files. All extracted features are made available for download at:

http://www.ifs.tuwien.ac.at/mir/mvd/

MV-VIS

Music **V**ideo Data-set for **VIS**ual content analysis. Intended for classifying music videos by visual properties only. Non overlapping genres. Tracks within a class share similar audible characteristics.

MV-MM

Music **V**ideo Data-set for **M**ulti **M**odal content analysis and classification. Intended for multi-modal classification and retrieval tasks. Overlapping genres with high inter and intra class variance.

MV-MIX

Combination of the data-sets MVD-VIS and MVD-MM. The distinct genres of the sub-sets have been selected to facilitate a non-overlapping a union of the two sets providing a bigger set.

Table 1: The Music Video Dataset - Detailed Overview of structure including class descriptions.

Genre	Videos	Artists	BPM
MVD-VIS			
Bollywood	100	32	112 (21)
Country	100	70	129 (30)
Dance	100	84	128 (1)
Latin	100	72	131 (5)
Metal	100	76	131 (31)
Opera	100	NA	104 (26)
Rap	100	81	104 (25)
Reggae	100	75	114 (36)
MVD-MM			
80ies	100	72	119 (16)
Dubstep	100	78	135 (14)
Folk	100	66	116 (28)
Hard Rock	100	69	129 (27)
Indie	100	64	135 (26)
Pop Rock	100	65	125 (25)
Reggaeton	100	69	126 (42)
RnB	100	67	114 (28)
MVD-MIX			
16 Genres	1600	1040	

Feature Extraction

Audio features are extracted from the separated audio channel of the music videos. Visual features are extracted from each frame of a video and aggregated during post-processing by calculating the statistical measures. As a pre-processing step black bars at the borders of video frames, also called Letter-boxing or Pillar-boxing, are removed.

VIDEO

Global Color Statistics

Mean Saturation and Brightness based on the Improved Hue, Luminance and Saturation (IHLS) color space. Hue in IHLS is an angular value. Circular statistics has to be applied to assess mean Hue and its angular deviation. Saturation weighted aggregations are more robust towards weakly saturated colors.

Colorfulness

Used to computationally describe aesthetics in photographies. Calculate the dissimilarity of an image to an ideal color distribution.

Color Names

Color distributions of the reduced Web-safe elementary-color palette: Magenta, Red, Yellow, Green, Cyan, Blue, Black and White. Contrast, brightness and color enhancement using Contrast Limited Adaptive Histogram Equalization (CLAHE). Ordered Dithering was used for color quantization, since it reduces the effect of contouring but stays more consistent with the original colors.

Figure 2: Sequential image enhancement steps.

Emotional Values

Pleasure-Arousal-Dominance model based on empirical investigated emotional reactions. The emotional values are linearly related to saturation and brightness in the IHLS color space.

Itten Contrasts

A set of art-theory concepts defined for combining colors to induce emotions based on an proportional opponent color model. Contrast calculation is aligned to Wang's feature extraction. Following contrasts were used: *Contrast of Light and Dark, Contrast of Warm and Cold, Contrast of Saturation, Contrast of Hue.*

Lightness Fluctuation Patterns

Calculated analogous to the music feature Rhythm Patterns. For each frame a 24 bin histogram of the lightness channel is calculated. Fast Fourier Transform (FFT) is applied to the histogram space of all video frames. This results in a time-invariant representation capturing reoccurring patterns in the video. Based on the observation that light effects, motions and shots are usually beat synchronized in music videos, LFPs can be assumed to express rhythmic structures of music videos.

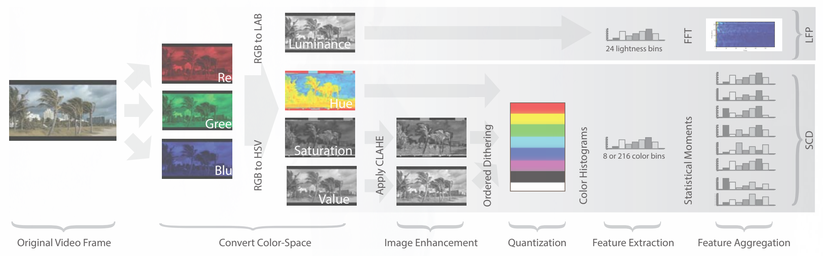

Figure 3: Feature extraction Workflow of Lightness Fluctuation Patterns

Wang Emotional Factors

Three factors based on emotional word correlations that are relevant for image retrieval based on emotion semantics.

Feature One: lightness description of a segmented image ranging from *very dark* to *very bright*, combined with the hue labels *cold* and *warm*.

Feature Two: description of warm or cool regions with respect to different saturations as well as a description of contrast.

Feature Three: combines lightness contrast with an estimation of sharpness. A no-reference perceptual blur measure was used.

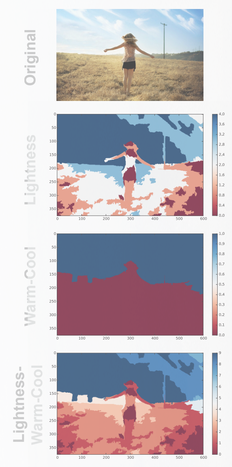

Figure 4: Visualization of contrast features.

AUDIO

MFCC

Mel Frequency Cepstral Coefficients (MFCC) are well known audio features derived from speech recognition.

Chroma

Chroma features project the spectrum onto 12 bins representing the semitones of the musical octave.

Psycho-accoustic Music Descriptors

Proposed by Rauber, Lidy et al. are based on a psycho-acoustically modified Sonogram representation that reflects human loudness sensation. **Statistical Spectrum Descriptors (SSD)** subsequently compute statistical moments for the 24 critical bands of hearing. **Rhythm Patterns (RP)** describe fluctuations in modulation frequency which provide a rough interpretation of the rhythmic energy of a song. **Rhythm Histograms (RH)** aggregate the modulation amplitude values of the individual critical bands computed in a RP, providing a lower-dimensional descriptor for general rhythmic characteristics. **Temporal Variants (TSSD, TRH)** describe variations over time through statistical moments calculated from consecutive segments of a track.

Table 2: Overview of all features used in the evaluation. The column '#' indicates the dimensionality of the corresponding feature set.

	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. [17]
	Lightness Fluctuation Patterns	80	Rhythmic fluctuations in video lightness

Evaluation

The empirical evaluation is based on the Music Video Dataset. Classification experiments using Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF) and Naive Bayes (NB) classifiers, were conducted. In a second analysis Chi-square feature selection was applied to evaluate the performance of the visual and audio-visual feature-spaces.

Figure 5: Distribution of named colors in the MV-VIS dataset

Results

Audio

Table 3 summarizes the results of the comparative classification experiments. The top segment of the table provides audio only results which serve as baseline for evaluating the visual and audio-visual approaches.

Table 3: Classification results for audio, visual and audio-visual features showing accuracies for Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF) and Naive Bayes (NB) classifiers. Bold-faced values highlight improvements of audio-visual approaches over audio features.

		MVD-VIS				MVD-MM				MVD-MIX			
		SVM	KNN	RF	NB	SVM	KNN	RF	NB	SVM	KNN	RF	NB
Audio	TSSD-RP-TRH	93.79	80.85	77.13	71.46	74.76	55.00	55.84	52.20	75.91	54.16	49.80	48.32
	TSSD	86.81	72.58	70.72	62.61	69.97	53.33	56.16	53.65	66.19	47.40	45.33	44.22
	RP	87.26	69.81	71.29	64.04	60.35	42.38	43.85	41.63	63.19	43.06	42.53	41.39
	SSD	85.78	73.18	72.80	58.81	68.74	50.28	54.43	48.41	65.11	44.64	46.18	38.92
	TRH	71.04	55.83	55.16	53.86	49.50	38.28	37.66	39.66	46.61	33.02	30.54	35.70
	MFCC	62.28	48.58	49.04	46.95	42.14	29.16	32.50	34.17	37.02	26.60	25.57	27.11
	Chroma	36.34	28.09	34.41	23.03	25.26	20.11	23.16	19.41	19.64	14.68	16.52	12.08
	Visual Features	50.13	34.04	38.60	39.38	31.69	21.16	22.86	23.38	32.22	17.89	19.36	21.16
Audio-Visual	TSSD-RP-TRH	**94.86**	**81.38**	76.51	**71.65**	**75.69**	**55.78**	54.36	51.36	**76.53**	**55.76**	49.15	**49.08**
	TSSD	**88.45**	71.65	68.80	**64.75**	**70.55**	52.60	54.34	52.25	**69.46**	46.15	43.12	**45.16**
	RP	**89.80**	**71.99**	69.90	**65.78**	**62.79**	**43.93**	43.74	41.61	**66.59**	**44.47**	40.61	**41.68**
	SSD	85.25	62.05	66.22	57.80	65.34	42.28	48.53	44.24	**65.21**	36.13	39.64	**38.76**
	TRH	**77.84**	**55.98**	**57.21**	**59.71**	**58.50**	32.79	35.60	**41.40**	**56.31**	31.39	**31.18**	40.09
	MFCC	**63.71**	41.53	45.78	46.28	**42.88**	24.38	27.40	27.35	**43.11**	22.33	22.89	**25.62**
	Chroma	**55.70**	**39.28**	**42.78**	**43.13**	**35.29**	**24.16**	**26.04**	**25.51**	**35.43**	**20.10**	**21.91**	**24.14**

Visual

Using visual features only an accuracy of 50.13% could be reached for Support Vector Machines (SVM) for the MVD-VIS set. Accuracies for other sets or classifiers range from 17.89% to 39.38%. Because all classes equal in size these results are above a baseline of 12.5% or 6.25% respectively. Yet, the performance of the visual features alone is not representative.

Audio-Visual

The audio-visual results show interesting effects. Generally, there is insignificant or no improvement of the performance over the top performing audio features. The results show that combining the visual features with chroma and rhythm descriptors has a positive effect on the accuracy while it is negative with spectral and timbral features.

Feature Selection

Applying ranked Chi-square attribute selection on the visual features shows, that affective features as well as the frequencies of black and white pixels have highest values. Further, more information is provided by variance and min/max aggregated values than by mean values.

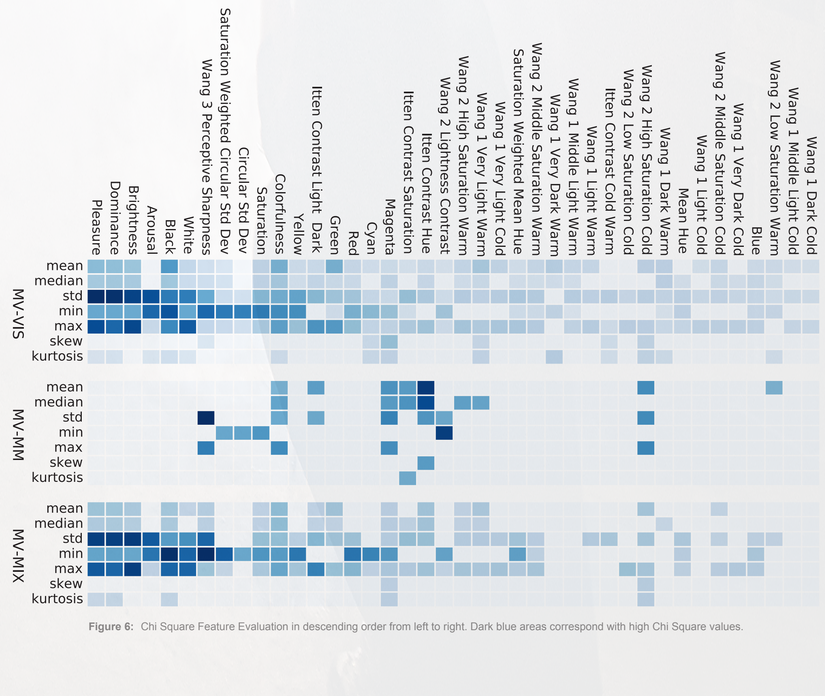

Figure 6: Chi Square Feature Evaluation in descending order from left to right. Dark blue areas correspond with high Chi Square values.

TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

[1)]Department of Software Technology and Interactive Systems
Vienna University of Technology
rauber@ifs.tuwien.ac.at

[2)]Next Generation Content Management Systems
AIT Austrian Institute of Technology
alexander.schindler@ait.ac.at

37th European Conference on Information Retrieval (ECIR'15), Vienna, Austria, March 29 - April 02 2015.