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

Alexander Schindler[1,2] and Andreas Rauber[1]

[1] Department of Software Technology and Interactive Systems
Vienna University of Technology
rauber@ifs.tuwien.ac.at
[2] Information management, Digital Safety and Security Department
AIT Austrian Institute of Technology
alexander.schindler@ait.ac.at

Abstract. This paper presents a study on classifying music by affective visual information extracted from music videos. The proposed audio-visual approach analyzes genre specific utilization of color. A comprehensive set of color specific image processing features used for affect and emotion recognition derived from psychological experiments or art-theory is evaluated in the visual and multi-modal domain against contemporary audio content descriptors. The evaluation of the presented color features is based on comparative classification experiments on the newly introduced 'Music Video Dataset'. Results show that a combination of the modalities can improve non-timbral and rhythmic features but show insignificant effects on high performing audio features.

1 Introduction

Over the past decades music videos distinctively influenced our pop-culture and became a significant part of it. Since their inception in the early 1980-ies music videos emerged from a promotional support medium into an art form of their own. The effort invested to produce a video creates enough information such that many music genres can be predicted by the moving pictures only. This potential of information provided was demonstrated in previous work on music video based artist identification [13], where a precision improvement of 27% could be observed over conventional audio features. Harnessing this potential presents a new way to approach existing Music Information Retrieval (MIR) problems such as an audio-visual approach to music video segmentation [4]. Approaches to affective content analysis of music videos are provided by [19] and [20]. In order to use the visual domain for music retrieval tasks, it has to be linked to the acoustic domain. Since substantial research on audio-visual correlations in music videos is yet scarce or not available, we base our approach on the simplified assumption that both layers intend to express the same emotions. In this paper we evaluate if this information - and more specifically the color information - is sufficient to discriminate music genres. Using color in content-based image retrieval has been extensively studied [9, 10, 12] and is yet described as problematic since it is highly influenced by lighting conditions during image acquisition. In music videos different illumination settings and colors are usually desired artistic effects. In the following section we introduce seven feature sets that derive from psychological experiments, art-theory or try to model human perception. Section 3 lays out the evaluation and introduces the Music Video Dataset to foster further research. After discussing the results in Section 4 conclusions and outlooks to future work are provided in Section 5.

	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

Table 1: Overview of all features. The column '#' indicates the dimensionality of the corresponding feature set.

2 Method

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 mean, median, standard deviation, min, max skewness, kurtosis. As a pre-processing step black bars at the borders of video frames, also called *Letterboxing* or *Pillarboxing*, are removed.

2.1 Audio Features

Psycho-accoustic Music Descriptors as proposed by [7] 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. *Temporal Variants (TSSD, TRH)* describe variations over time through statistical moments calculated from consecutive segments of a track. For the extraction, we employed the Matlab-based implementation, version 0.6411.

Mel Frequency Cepstral Coefficients (MFCC) are well known audio features derived from speech recognition. **Chroma** features project the spectrum onto 12 bins representing the semitones of the musical octave. We utilized MARSYAS [14] version 0.4.5.

2.2 Visual Features

Global Color Statistics calculate *Mean Saturation* and *Mean Bightness* based on the Improved Hue, Luminance and Saturation (IHLS) color space [18] which has the advantages of low saturation values of achromatic pixels and independence of saturation from the brightness function. Hue in IHLS is an angular value. Circular statistics has to be applied [5] to assess *angular mean Hue* and *angular deviation of Hue*. *Saturation weighted mean Hue and deviation of Hue* are more robust towards weakly saturated colors.

Global Emotion values refer to a Pleasure-Arousal-Dominance model based on investigated emotional reactions presented in [15]. The introduced relationship between saturation (S) and brightness (B) is calculated from the corresponding IHLS channels:

$$Pleasure = 0.69 * B + 0.22 * S \quad (1)$$

$$Arousal = -0.31 * B + 0.60 * S \quad (2)$$

$$Dominance = 0.76 * B + 0.32 * S \quad (3)$$

Colorfulness is one of the features used in [2] to computationally describe aesthetics in photographies. The proposed method is based on a partitioned RGB palette using Earth Mover's Distance (EMD) [11] to calculate the dissimilarity of a supplied image to an *ideal* color distribution of a *colorful* image.

Wang Emotional Factors Wang et al. [17] identified three factors based on emotional word correlations that are relevant for image retrieval based on emotion semantics. Three feature sets are calculated using fuzzy membership functions to assign values of the perceptual psychology motivated L*C*H* color space to discrete semantic words. *Feature One* includes lightness description of a segmented image ranging from *very dark* to *very bright*. These are combined with the calculated hue labels *cold* and *warm*. *Feature Two* provides a 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 sharpness estimation. A no-reference perceptual blur measure [1] was used. The sharpness is further calculated by $1 - blurIndex$. The contrast description overlaps with the *Itten contrasts* and is omitted.

Itten's Contrasts are a set of art-theory concepts defined by Johannes Itten [6] for combining colors to induce emotions based on an proportional opponent color model. The contrast calculation is aligned to the method presented in [8] which uses Wang's feature extraction [17] as a predecessor. Instead of a waterfall segmentation we used a Quick Shift [16] approach due to better performance at reasonable processing time. We calculated the following contrasts: *Contrast of Light and Dark*, *Contrast of Saturation*, *Contrast of Hue* and *Contrast of Warm and Cold*.

Color Names describe color distributions of the reduced Web-safe Elementary-color palette consisting of the 8 elementary colors Magenta, Red, Yellow, Green, Cyan, Blue, Black and White. To map a frame of a video to this palette it is converted to Hue Value Saturation (HSV) color-space. *Contrast, brightness and color enhancement* is applied through application of Contrast Limited Adaptive Histogram Equalization (CLAHE) [21]. *Color Quantization* to reduce the number of distinct colors of the frame to the desired palette is obtained by applying *error diffusion* which computes the mean square error between the original pixel value and its closest match which is then propagated locally to its surrounding pixels. *Ordered Dithering* was used since it reduces the effect of contouring but stays more consistent with the original colors. A 32x32 Bayer pattern matrix was used as threshold map. *Feature Calculation* is concluded by calculating the statistical moments mean, median, variance, min, max, skew and kurtosis of the reduced palette.

Lightness Fluctuation Patterns are calculated analogous to the music feature Rhythm Patterns (RP) [7] from the perceptually uniform LAB color space. 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 of the 24 lightness levels capturing reoccurring patterns in the video. Only amplitude modulations in the range from 0 to 10 Hz are used for the final feature set, since rhythm cannot be perceived from higher modulation frequencies. 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.

3 Evaluation - The Music Video Dataset

The empirical evaluation is based on the Music Video Dataset (MVD). We use empirical classification experiments and Chi-square feature selection to analyze the performance of the visual and audio-visual feature-spaces. The MVD is a collection of carefully selected music videos. It consists of different subsets that can be combined to bigger data-sets. The following sub-sets of the MVD are used to evaluate the features presented in Section 2:

MVD-VIS: The *M*usic *V*ideo *D*ataset for *VIS*ual content analysis and classification is intended for classifying music videos by their visual properties only. Special emphasis has been set on minimizing the intra- and maximising the inter-class variance in the acoustic domain of the dataset. Non overlapping sub-genres were chosen and tracks within a certain class share very similar musical characteristics. Music genre classification based on conventional audio features provides accuracy above-average (see Table 2) compared to current benchmarks of the Music Information Retrieval domain [3].

MVD-MM: The *M*usic *V*ideo *D*ataset for *M*ulti*M*odal content analysis and classification is intended for multi-modal classification and retrieval tasks. The overlapping classes have high inter and intra class variance. Genre classification based on audio features provides average results and serves as starting point for multi-modal approaches.

MVD-MIX: The MVD-MIX data-set is a combination of the data-sets MVD-VIS and MVD-MM. The distinct genres of the sub-sets have been selected in a way, that a union of the two sets provides a non-overlapping bigger set. Consequently the inter-class variance increases while the intra-class variance remains the same as for the individual sets. While the sub-sets are intended for developing content descriptors, the MVD-MIX should be used for audio-visual evaluations.

The dataset creation was preceded by the selection of the non-overlapping genres respectively to enable the combination of the two subsets into the bigger *MVD-MIX* dataset. Each genre consists of 100 selected videos. Resulting in dataset sizes of 800 music videos for *MVD-VIS* and *MVD-MM* each as well as 1600 for the MV-MIX dataset. Music videos were selected primarily by their audible properties. A set of selection criteria has been applied such as quality criteria of at least 90 kBits/s audio encoding and video resolution ranging from QVGA to VGA. Only official music videos were selected, no live performance, abstract or animated videos. Artist stratification is provided by selecting only two tracks per artist.

Data Provision: Due to copyright restrictions it is not possible to redistribute music videos or audio files. Yet, all videos have been retrieved from Google's Youtube platform and a list of corresponding Youtube video-ids is provided. It should be stated that the availability of these videos cannot be guaranteed and that some may vanish over time. To ensure comparability of results and reproducibility of the experiments, all features of this publication including a range of standard visual and acoustic features are being provided and customized features will be extracted and provided on request. All extracted features are made available for download at: `http://www.ifs.tuwien.ac.at/mir/mvd/`.

4 Results

Table 2 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. Using visual features only an accuracy of 50.13%

Table 2: 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**

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. 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. 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.

5 Conclusions and Future Work

We presented a comparative evaluation of audio-visual music classification that focused on the color information of music videos. A set of diverging approaches based on psychological or perceptive models has been applied to extract different kinds of semantic information. We further introduced a descriptor that captures rhythmical changes in illumination. The performance of the color features is generally noticed as weak, while some interesting effects on chromatic and rhythmic features in the audio-visual domain are observed.

Future work on music videos will extend the semantic space to include texture, local features and object detection. Results are expected to provide information about how appropriate these methods are to solve MIR problems and how they can be used to connect the audio with the visual domain to facilitate new scenarios such as query-by-image.

References

1. F. Crete, et al. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Electronic Imaging 2007*, pages 64920I–64920I, 2007.
2. R. Datta, et al. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.
3. Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions on*, 13(2):303–319, 2011.

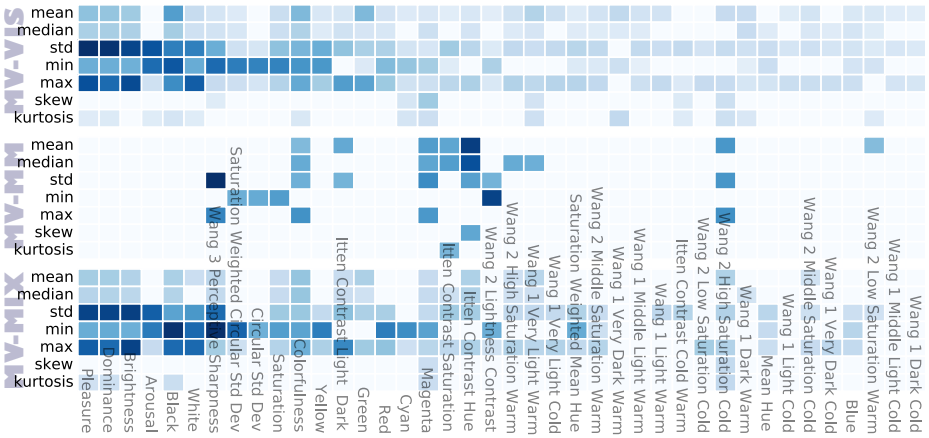

Fig. 1: Chi Square Feature Evaluation in descending order from left to right. Dark blue areas correspond with high χ^2 values.

4. O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentations of music videos. *IEEE Trans on Circuits and Sys. for Video Tech.*, 2007.
5. A. Hanbury. Circular statistics applied to colour images. In *8th Computer Vision Winter Workshop*, volume 91, pages 53–71. Citeseer, 2003.
6. J. Itten and E. Van Haagen. *The art of color: the subjective experience and objective rationale of color*. Van Nostrand Reinhold New York, NY, USA, 1973.
7. T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR*, 2005.
8. J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proc. Int. Conf. on Multimedia*, pages 83–92, 2010.
9. B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans on Circuits and Sys. for Video Tech.*, 11(6):703–715, 2001.
10. K. N. Plataniotis and A. N. Venetsanopoulos. *Color image proc. and applications*, 2000.
11. Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
12. R. Schettini, G. Ciocca, S. Zuffi, et al. A survey of methods for colour image indexing and retrieval in image databases. *Color Imaging Science: Exploiting Digital Media*, 2001.
13. A. Schindler and A. Rauber. A music video information retrieval approach to artist identification. In *10th Symp. on Computer Music Multidisciplinary research*, 2013.
14. G. Tzanetakis, P. Cook. Marsyas: A framework for audio analysis. *Organised sound*, 2000.
15. P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.
16. A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008*, pages 705–718. Springer, 2008.
17. W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *IEEE International Conference on Systems, Man and Cybernetics*. 2006.
18. H. Wildenauer, P. Blauensteiner, A. Hanbury, and M. Kampel. Motion detection using an improved colour model. In *Advances in visual computing*. Springer, 2006.
19. A. Yazdani, K. Kappeler, and T. Ebrahimi. Affective content analysis of music video clips. In *Music information retrieval with user-centered and multimodal strategies*, 2011.
20. S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *Multimedia, IEEE Transactions on*, 12(6):510–522, 2010.
21. K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press Professional, Inc., 1994.