Harnessing Music related Visual Stereotypes for Music Information Retrieval

Alexander Schindler, Austrian Institute of Technology
Andreas Rauber, Technische Universitt Wien

Over decades music labels have shaped easily identifiable genres to improve recognition value and subsequently market sales of new music acts. Referring to print magazines and later to music television as important distribution channels, the visual representation thus played and still plays a significant role in music marketing. Visual stereotypes developed over decades which enable us to quickly identify referenced music only by sight without listening. Despite of the richness of music related visual information provided by music videos, album covers as well as T-shirts, advertisements and magazines, research towards harnessing this information to advance existing or approach new problems of music retrieval or recommendation is scarce or missing. In this paper we present our research on visual music computing which aims to extract stereotypical music related visual information from music videos. To provide comprehensive and reproducible results we present the Music Video Dataset, a thoroughly assembled suite of datasets with dedicated evaluation tasks that are aligned to current Music Information Retrieval tasks. Based on this dataset we provide evaluations of conventional low-level image processing and affect-related features to provide an overview of the expressiveness of fundamental visual properties such as color, illumination and contrasts. Further we introduce a high-level approach based on visual concept detection to facilitate visual stereotypes. This approach decomposes the semantic content of music video frames into concrete concepts such as vehicles, tools, etc., defined in a wide visual vocabulary. Concepts are detected using convolutional neural networks and their frequency distributions as semantic descriptions for a music video. Evaluations showed that these descriptions show good performance in predicting the music genre of a video and even outperform audio-content descriptors on cross-genre thematic tags. Further, highly significant performance improvements were observed by augmenting audio-based approaches through the introduced visual approach.

CCS Concepts: •**Information systems** → **Multimedia and multimodal retrieval; Music retrieval;** *Multimedia information systems;* •**Computing methodologies** → **Visual content-based indexing and retrieval;** *Object recognition;* Supervised learning by classification;

General Terms: Multimedia, Video Retrieval, Audio Retrieval

Additional Key Words and Phrases: Music Videos, Visual Concept Detection, Video Analysis

ACM Reference Format:
Alexander Schindler and Andreas Rauber, 2016. Harnessing Music related Visual Stereotypes for Music Information Retrieval. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 1 (July 2016), 20 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Malcolm McLaren, the manager of the punk rock band "The Sex Pistols", stated in 1977 "Christ, if people bought the records for the music, this thing would have died a death long ago". In his provoking way he indicated that there are far more dimensions to music purchase behavior than the artistic quality of an act. Punk rock was a result of social-political differences of the late 1970s and dressing in leather jackets, spike

Author's addresses: A. Schindler, Digital Safety and Security Department, Austrian Institute of Technology, Vienna, Austria A. Rauber, Institute of Software Technology and Interactive Systems, Vienna, Austria.
 1539-9087/2016/07-ART1 $15.00
DOI: 0000001.0000001

bands and Mohawk hairstyle was a visual commitment to the views and rebellious attitudes expressed by the genre. Such visual stereotypes are adopted by and learned over generations from mass media. As a consequence stereotypes such as cowboy-shoes or hats are often visualized when referring to Country music [Shevy 2008]. Such visual associations are constantly rehashed to recognizably shape the image of a music act [Negus 2011]. Using these visual associations one could sketch the soundtrack of the past decades without playing a single note. Flowery clothes, twinkling mirror-balls, dreadlocks, shoulder pads, skulls and crosses, Cowboy hats, Baseball caps and lots of jewelry are visual synonyms for music styles ranging from the early sixties to contemporary Hip-Hop. Visual stereotypes play an essential role in our social interaction with unfamiliar others [Haake and Gulz 2008]. They trigger a categorization process in which we quickly form expectations on a person's likely behavior, attitudes, opinions, personality, manners, etc. and thus shape our personal attitude towards that person. Drama theory and film make profound usage of such concepts [Finkelstein 2007]. The vicious antagonist is often recognizable at a glimpse of their first appearance by capitalizing on stereotypes such as ragged clothes, scars and over-exaggerated armament.

Accordingly, in his influential book "Emotion and Meaning in Music" [Meyer 1956] Meyer addresses "extramusical" connotations, explicitly referring to classes of visual associations which are cognitive schema that are culturally shared by groups of individuals. Half a century later it had been shown that these associations became culturally independent as a consequence of global mass media exposure [Shevy 2008]. Artist development departments utilize visual concepts to categorize a new act by an appropriate genre and to create its visual image that is used for promotion in mass media [Negus 2011]. Consequently, easily identifiable genres are a desired goal of music labels. Music videos not only make use of stereotypical visual themes, but also take part in their development and propagation. This lead to the circumstance that many music properties such as genre, instrumentation, tempo and even rhythm can be estimated by the visual representation of muted music videos. From an information retrieval perspective it can thus be hypothesized that the visual layer of music videos contains enough music related information to approach the Music Information Retrieval (MIR) problem space via solutions deriving from the visual computing domain. This presents a new approach to existing MIR problems such as classification (e.g. genre, mood, artist, etc.), structural analysis (e.g. music segmentation, tempo, rhythm, etc.) or similarity retrieval. Taking the widespread comprehension of music related visual stereotypes into account these concepts can be used to facilitate new innovative approaches to search for or organize music. We provide a comprehensive overview of the objectives of exploring the visual aspects of music in [Schindler 2014]. The potential of transferring methods from the visual computing domain to MIR tasks was demonstrated by a previous study [Schindler and Rauber 2013] on artist identification. This evaluation showed that state-of-the-art approaches can be significantly improved through visual features extracted from music videos. Nevertheless, extensive research on harnessing the music related information of the visual layer of music videos is scarce and will be discussed in detail in Section 2.

In this publication we present our research on visual music computing for which the aims were previously described in [Schindler 2014]. In Section 3 we provide a brief overview of music video types and production techniques to provide a common understanding of music video characteristics, properties and complexity. To provide comprehensive and plausible results we first thoroughly assembled a suite of datasets (see In Section 4) we introduce the Music Video Dataset as a collection of benchmark datasets with clearly defined evaluation tasks and scenarios. Based on this dataset we provide a series of evaluations in Section 8 on the sequentially introduced audio-content based music descriptors (see Section 5), low-level image processing features (see Section 7.1)

and the new introduced high-level semantic descriptors based on visual vocabularies (see Section 7). Finally, we present our conclusions and future work in Section 9.

2. RELATED WORK

Music Genre Recognition (MGR) is a well researched MIR [Downie 2003] task. As for many other audio-based MIR tasks the algorithmic design consists of two parts. First, audio-content descriptors are extracted from the audio signal which are commonly referred to as audio *features*. Usually this is accomplished through transformations of the signal's short-term Fourier transform (STFT) magnitude response and statistical aggregations. One of the most common features are the Mel Frequency Cepstral Coefficients (MFCC) [Logan et al. 2000]. A comprehensive overview of music features is provided by [Lartillot and Toiviainen 2007; Tzanetakis and Cook 2002; Lidy and Rauber 2005]. In a second step these features are used to train machine learning based models, using popular supervised classifiers including k-nearest neighbors (k-NN), Gaussian mixture models (GMM) or Support Vector Machines (SVM). Comprehensive surveys on music and genre classification are provided by [Scaringella et al. 2006; Fu et al. 2011]. It should be noted that recent publications argue that relying on classification accuracy alone can be a misleading criterion in tasks as MGR [Sturm 2013]. Arguments of the discussion refer to the ambiguities and subjectivity inherent to music genre [McKay and Fujinaga 2006] as well as cross-cultural differences in genre perception [Shevy 2008] and question the validity of state-of-the-art content-based audio descriptors because they do not reflect human music perception but rather train on abstract patterns derived from the audio spectrum [Sturm 2014]. In this publication we do not insist on recognizing the correct genre but rather use the same methodology to classify music videos by artificially assigned labels which refer to common acoustic properties (see Section 4). Through isolated experiments on classes that have clearly defined boundaries based on isolated stylistic elements the rational of this study is to identify and capture extramusical concepts related to music genres [Shevy 2008; Sturm 2013].

Multimodal Approaches to MIR approach tasks such as MGR by utilizing music information of different modalities such as song lyrics [Mayer et al. 2008; Hu and Downie 2010], web pages [Schedl et al. 2006] and social media/tagging [Lamere 2008]. Visual related music information extraction has been reported to utilize album art images for MGR [Mayer 2011], artist identification [Libeks and Turnbull 2011] and similarity retrieval [Brochu et al. 2003]. In [Lıbeks and Turnbull 2010] image features extracted from promotional photos of artists are used to estimate the musical genre of the image. A multimodal approach to mood classification using audio and image features is reported in [Dunker et al. 2008]. Perceptual relationships between art in the auditory and visual domains were analyzed in [Mattek and Casey 2011].

Music Video Processing Multi media analysis of music videos is part of reported automatic music video creation systems. [Foote et al. 2002] presented an approach based on audio and video segmentation. Based on calculated suitability scores segments of the source video are selected and combined to a new music video. [Hua et al. 2004] and [Yoon et al. 2009] build upon [Foote et al. 2002] to automatically build music video-like videos from personal home videos. [Cai et al. 2007] extracts salient words or phrases from the lyrics and uses them to search for corresponding images on the web. Also contrast features similar to those used in this publication were used. An approach to automatically determine regions of interest in music videos is reported in [Kim and Kim 2007]. An approach to automatic music video summarization is presented in [Shao et al. 2006]. An audio-visual approach to segmentations of music videos was proposed in [Gillet et al. 2007], including an evaluation of audio-visual correlations with

an intended application in audio retrieval from video. Approaches to affective content analysis of music videos are provided by [Zhang et al. 2010] and [Yazdani et al. 2011]. A comprehensive evaluation of color and affect related features extracted from music videos for genre classification is presented in [Schindler and Rauber 2015]. A study on artist identification in music videos using face recognition is presented in [Schindler and Rauber 2013]. An approach using convolutional neural networks (CNNs), in order to learn mid-level representations from combinations of low-level MFCCs and RGB color-values in order to build higher level audio-visual representations, is presented in [Acar et al. 2014]. Reported accuracies are low and comparable to low-level visual features evaluated in [Schindler and Rauber 2013]. [Sasaki et al. 2015] describes an audio-visual approach for a recommender system that uses a video as input query. To index music via the video domain mood values calculated from the audio content were matched to mood information derived from the color information. The visual features used by the authors [Valdez and Mehrabian 1994] were also evaluated in this publication. Unfortunately we were not able to draw the same conclusions about the effectiveness of these features for music classification. [Acar et al. 2014] apply convolutional neural networks to MFCCs and color values to learn higher level representations for classifying music videos. The approach is evaluated on the DEAP dataset [Koelstra et al. 2012], a music video dataset with valence and arousal valued ground truth data.

Transfer Learning refers to concept of transferring knowledge from a related task that has already been learned within a given context to a problem of a different context or even a different research domain [Pan and Yang 2010]. A well known example is the transfer of the well evaluated automatic speech recognition method based on MFCC to the MIR domain [Logan et al. 2000]. A comprehensive summary of further examples is provided by [Weninger et al. 2012]. In this study we transfer knowledge from the visual computing domain to MIR by applying trained visual concept detection to music classification tasks.

3. MUSIC VIDEO - AN INTRODUCTION

Formerly simply known as *promotional videos* music video production came to pass in the early 80s, although they were initially criticized to provoke a diminishing of the interpretative liberty of the individual music listener by imposing the narrative visual impression upon him or her [Frith et al. 2005]. Contemporary video production typically uses a wide range of film making techniques such as screen-play, directors, producers or director of photography [Negus 2011]. This section provides a brief overview of music video types, concepts and properties, exceeding those relevant for the experiments performed in the evaluation to provide a basis for potential future work.

3.1. Types of Music Videos

Music videos are short films intended to promote a song and its performing artist. Music video production is usually based on one of three major concepts [Frith et al. 2005]. *Illustration* visually displays or adds further explanation to the meaning of the song as usually told by the lyrics. This is accomplished by either using a narrative plot or visual clues that are related to lyrics or harmonics. *Amplification* further emphasizes the meaning of song or distinct relevant messages of the lyrics and reinforces them through constant repetitions and dominating visual accentuation. *Contradiction* ignores the underlying meaning of lyrics and melodics. The visual layer either counterpoints the meaning of the song or becomes completely abstract. Generally, three main types of music videos can be considered [Frith et al. 2005]. **Narrative videos** use illustration to portray the story told by the lyrics, but may also contradict the underlying song by telling a completely different story. Usually the narrator is the

performing artist who often acts as the protagonist of the story. **Concept videos** do not narrate the meaning of the lyrics but use illustration through visual concepts and metaphors. The plot is mostly obscure and surreal, trying to attract and entertain the audience by constantly keeping their attention on the screen. **Performance videos** present the artist performing the song - typically in an environment corresponding to the song or genre (e.g. with friends, live or staged performance, studio recording, etc.). Further music video types include *animation*, where images are created by computer, hand drawn or molded out of plastic material, *lyric video* where the focus of the video is on the artistic display of the lyrics, or *mixtures*.

3.2. Stylistic Properties of Music Videos

A wide range of cinematographic styles and effects can be found in music videos. While film making has developed standardized rules how to aesthetically apply these concepts, music videos do not follow these rules and often give these elements different functions and meanings [Vernallis 2004]. The following list summarizes the most important stylistic features of music videos:

Camera Work and Movement: Abiding the continuing progression of the musical track movement is one of the most common properties of music videos. Most of the shots of a video make use of one or more of the following techniques: *tracking shot* - the camera moves towards or away from the subject, *pan* - the camera rotates to the left or right to reveal more of the environment, *tilt* - similar to panning but camera rotates vertically, *zoom in* or *out* - magnify or reduce the subject by changing the focal length. *Tempo* of fore- and background movement is often used to express moods.

Lighting and Color: Lighting and color can influence our interpretation of an individual scene. *Color* can be used to attract attention or to express moods (e.g. warm colors are often used to create romantic or pleasurable ambiance). *Lighting* in visual production has developed many techniques to create all kinds of effects and moods (e.g. directional light, back and spot lighting, shadows, etc.).

Visual and stylistic coherence: A common aimed at property of music videos. Shots of a video provide a coherent impression that should correspond to the underlying track. Usually such effects are achieved through the application of global filters (e.g. color, blur, etc.) during post-production.

Close Ups: Usually a full-portrait shot composed from below the shoulder line. There is often a pressing demand of music labels to promote and iconize the contracted artist through a proliferated presence on screen. Close ups are further applied to underscore a musical hook or the peak of a phrase or, more recently, to advertise brand names or products added to the video to earn additional revenue.

Editing and Shots: Contrary to film making, editing does not intend to remain unnoticed but is perceived as an art-form. Traditional rules to guide the viewer in time and space are ignored due to its short form and the demand to showcase the star. Shots are combined by favoring compositional elements such as color or shape over content. *Jump Cuts* are disjunctive edits frequently employed using drastic shifts in color, scale or content. *Low angle shots* reproduce the impression of looking up to an artist performing on stage while *High-angle shots* harmonize with key moments of a song. Mixing those in a series disorients the viewer who seeks additional guidance in the music. Shot length and boundaries commonly align to tempo and rhythm of the track.

4. THE MUSIC VIDEO DATASET

To facilitate comparable results and reproducible research on music related visual analysis of music videos we sequentially introduced the Music Video Dataset (MVD) [Schindler and Rauber 2013; 2015]. The MVD follows the Cranfield paradigm [Clever-

Table I. The Music Video Dataset - Detailed Overview of structure including class description, number of artists, average Beats per Minutes and standard deviation per genre.

MVD-VIS			**MVD-MM**		
Genre	**Videos**	**Artists**	**Genre**	**Videos**	**Artists**
Bollywood	100	32	80s	100	72
Country	100	70	Dubstep	100	78
Dance	100	84	Folk	100	66
Latin	100	72	Hard Rock	100	69
Metal	100	76	Indie	100	64
Opera	100	NA	Pop Rock	100	65
Rap	100	81	Reggaeton	100	69
Reggae	100	75	RnB	100	67
MVD-MIX			**MVD-Themes**		
MVD-VIS + MVD-MM 16 Genres	1600	1040	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

don 1967] and provides test collections and of multimedia documents and corresponding ground truth assignments within the context of well defined tasks. The main focus in on the development and evaluation of visual or audio-visual features that can be used to augment or substitude audio-only based approaches. The MVD consists of four major subsets that can be combined to two bigger task related collections. This section provides an overview of the sub-set characteristics and their corresponding tasks. For further information on distinct dataset properties and class/genre descriptions please refer to the MVD webpage[1].

MVD-VIS: The *M*usic *V*ideo *D*ataset for *VIS*ual content analysis (MVD-VIS) is intended for feature development and optimization. To facilitate this, 800 tracks of eight clearly defined and well differentiated sub-genres were aggregated (see Table I). Their tracks were selected concerning minimal variance in acoustic characteristics, thus sharing high similarity in instrumentation, timbre, tempo, rhythm and mood. Audio classification results provided in Table II reflect that state-of-the-art audio-content based approaches can accurately discriminate such well differentiated classes. Based on the premise that the tracks of a class sound highly similar, these results should serve as a baseline for the task of identifying patterns of similarity within the visual layer as well as developing means to extract this information.

MVD-MM: The structure of the *M*usic *V*ideo *D*ataset for *M*ulti*M*odal content analysis (MVD-MM) is aligned to the *MVD-VIS* but its classes are less well differentiated. The heterogeneous distributions of inter and intra class variance refer to problems of

[1]http://www.ifs.tuwien.ac.at/mir/mvd/

imprecision and subjectivity of music genre definitions [McKay and Fujinaga 2006] which are observed in current music classification datasets [Sturm 2013]. This variance was intentionally introduced to facilitate comparability with results reported on these datasets [Schindler and Rauber 2014]. The task is to evaluate and improve the performance of visual features in such environments and if state-of-the-art audio-only based approaches can be improved through audio-visual combinations.

MVD-ARTISTS: This dataset for audio-visual based artist identification using music videos is a set of 20 popular western music artists listed in Table I extending the dataset which has been introduced in [Schindler and Rauber 2013]. Popular artists were chosen to meet the requirement of collecting enough music videos for each musician. To demonstrate the previously mentioned problems of content based artist identification the selected artists belong predominately to the two genres Pop and Rock.

MVD-THEMES: The MVD-Themes set is a collection of thematically tagged classes that span across musical genres. The task is aligned to the MusiClef multi-modal music tagging task [Orio et al. 2012]. The strong contextual and non-audio connotations of the themes should be described through information extracted from the visual layer. To address cross-lingual and cross-cultural challenges [Lee et al. 2005] of multi-modal approaches analysing song lyrics [Mayer et al. 2008; Hu and Downie 2010] (most of these approaches were evaluated only for the English language) the MVD-THEMES set includes performances in various languages. The following Themes are provided:

— **Christmas:** Tracks that can be related to Christmas. *Genres* covered: Alternative-, Indie- and Hard-Rock, 60s-, 80s- and 90s-Pop, Dance, Rock 'n Roll, RnB, Soul, Big Band, Country, A capella. *Langauges:* English, Thai.

— **K-Pop:** Korean-Pop is strongly influenced by western music [Lee et al. 2013] and characterized through visual content (synchronized dance formations, colourful outfits). *Genres* covered: Pop, Dance, RnB, Rap. *Langauges:* Korean, English (chorus).

— **Broken Heart:** Songs about sorrowfully loosing someone beloved either through death or end of a relationship. *Genres* covered: Rock, Hard Rock, Metal, Pop, RnB, Country, Folk, 80s. *Languages:* English, Taiwanese.

— **Protest Songs:** Songs protesting against war, racism, police power and social injustice. *Genres* covered: Pop, Folk, Rap, Reggae, Rock, Punk Rock, Metal, Punjabi, Indie, 80s. *Languages*: English, French, German, Egyptian Arabic, Hindi.

MVD-MIX and MVD-COMPLETE: The *MVD-MIX* dataset is a combination of the datasets MVD-VIS and MVD-MM. The distinct genres of the subsets have been selected to facilitate a union of the two sets providing a non-overlapping bigger set. While MVD-VIS is intended for feature development and optimization and MVD-MM is for evaluation, the MVD-MIX set is for evaluating the performance of the features concerning their stability towards a higher number of classes. The *MVD-Complete* dataset is a combination of the MVD-MIX and the MVD-Artists datasets providing 2212 music videos for similarity search and recommendation. Since the classes of the two separate sets are not related, no class-labels are provided.

4.1. Dataset Creation and Data Provision

To align the dataset to contemporary music repositories a pre-selection was based on the Recording Industry Association of America's (RIAA) report on consumer expenditures for sound recordings [of the Census and States 2009] which separates profiles into the following genres: Rock, Pop, Rap/Hip Hop, R&B/Urban, Country, Religious, Classical, Jazz, Soundtracks, Oldies, New Age, Children's and Other. Each class consists of 100 videos which were primarily selected by their audible properties. The class labels were artificially assigned and only refer to commonly known music genres. They

do not infer to be accurate in musicological terms and are not result of a common agreement. After listening to the tracks the videos were inspected to ensure they conform the set of pre-defined quality criteria: **Quality filter:** A minimum of 90 kBits/s audio encoding; A video resolution ranging from QVGA (320x240) to VGA (640x480). **Content filter:** Only official music videos; No lyric-videos (Videos showing only lyrics); No non-representational (not showing artists); No live performance, abstract or animated videos; No videos with intro/outro longer than 30 seconds; No or minimal graphical overlays (e.g. channel logo, advertisment, etc.). **Stratification:** Only two tracks by the same artist (exceptions: Bollywood, Opera); Artists of the *MVD-Artists* dataset do not feature other artists of this set; The stratification rule for the Bollywood and Opera class was substituted by: Bollywood: only two tracks from the same movie; Opera: only two tracks of the same opera/performance. These criteria and the variance constraints of the subset's genres made the selection process complex and exhaustive. More than 6000 videos were examined and downloaded from Youtube in MPEG-4 format.

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. Thus, to ensure comparability of results, a range of standard visual and acoustic features are being provided. 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/.

5. AUDIO CONTENT DESCRIPTORS

The audio features used for the experiments are well evaluated music content descriptors widely used in the music information retrieval domain [Fu et al. 2011] and provide a good timbral, temporal, rhythmic and harmonic description of the music content.

Psycho-acoustic Music Descriptors as proposed by [Lidy and Rauber 2005] are based on a psycho-acoustically modified Sonogram representation that reflects human loudness sensation. *Statistical Spectrum Descriptors (SSD)* subsequently computes seven statistical measures for the 24 critical bands of hearing. Mean, median, variance, skewness, kurtosis, min- and max-values, for different segments of a song are aggregated by calculating the median of the descriptors of all segments. *Rhythm Patterns (RP)* describe rhythmical characteristics by applying a discrete Fourier transform to the psycho-acoustically transformed Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band. These fluctuations in modulation frequency 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):* Temporal Statistical Spectrum Descriptor (TSSD) and Temporal Rhythm Histograms (TRH) describe variations over through statistical moments calculated from consecutive segments of a track. For the extraction, we employed the Matlab-based implementation, version 0.6411[2]

Mel Frequency Cepstral Coefficients (MFCC) [Tzanetakis and Cook 2000] are the most commonly used features in and outside the MIR domain. They are derived from speech recognition and also apply log-scale transformations to anneal the feature response to the human auditory systems. MFCCs are good descriptors of timbre.

[2]http://www.ifs.tuwien.ac.at/mir/downloads.html

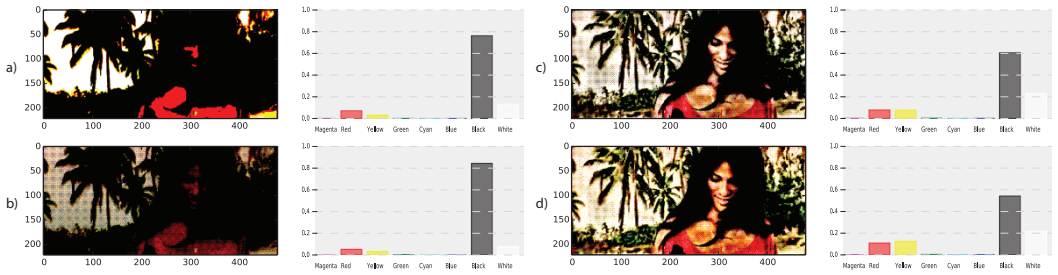


Fig. 1. Visualization of the image enahncement steps: a) simple nearest neighbor color quantization. b) Ordered Dithering (OD). c) OD with enhanced brightness. d) OD with enhanced brightness and saturation.

Chroma [Tzanetakis and Cook 2000] features project the entire spectrum onto 12 bins representing the 12 distinct semitones of the musical octave. Both MFCC and Chroma were added to the evaluation as an outline, because their performance on conventional datasets is well known. The features are extracted using the well known MARSAYS toolset [Tzanetakis and Cook 2000] version 0.4.5[3].

6. VISUAL CONTENT DESCRIPTORS

The selection for this part-evaluation presents consists of well-known low-level image processing features to provide a comprehensible estimation of their expressibility. Further, color based features derived from art-theory and empirical psychological studies, used for affective image retrieval, are applied to study possible effects of intentional color usage in music videos.

Global Color Statistics (GCS) are calculated based on the Improved Hue, Luminance and Saturation (IHLS) color space [Hanbury and Serra 2002] which has the advantages of low saturation values of achromatic pixels and saturation is independent of the brightness function. *Mean Saturation* and *Mean Brightness* values are calculated. Hue in IHLS is an angular value requiring circular statistics [Hanbury 2003] to assess *angular mean Hue* and *angular deviation of Hue*. *Saturation weighted mean Hue and deviation of Hue* introduces the relationship between hue and saturation by weighting the unit hue vectors by their corresponding saturation.

Global Emotion Values (GEV) refer to a Pleasure-Arousal-Dominance emotion model based on investigated emotional reactions presented in [Valdez and Mehrabian 1994]. The authors introduce a linear relationship between saturation and brightness as a model for the emotional variables. The values were calculated from the luminance and saturation channel of the previously described independent IHLS color space.

Colorfulness (CF) is one of the features used in [Datta et al. 2006] to computationally describe aesthetics in photographies. The proposed method is based on a partitioned RGB palette using Earth Mover's Distance (EMD) [Rubner et al. 2000] to calculate the dissimilarity of a supplied image to an *ideal* color distribution of a *colorful* image.

Wang Emotional Factors (WEF) Wang et al. [Wei-ning et al. 2006] identified three factors based on emotional word correlations that are relevant for emotion semantics based image retrieval and defined three corresponding feature sets. Fuzzy membership functions are used to assign values of the perceptual psychology motivated L*C*H* color space to discrete semantic words. *Feature One* includes lightness description of image segments ranging from *very dark* to *very bright* (Figure 2b). These are combined with the classification of hue into *cold* and *warm* colors, resulting in 10 dimensional

[3]http://sourceforge.net/projects/marsyas/

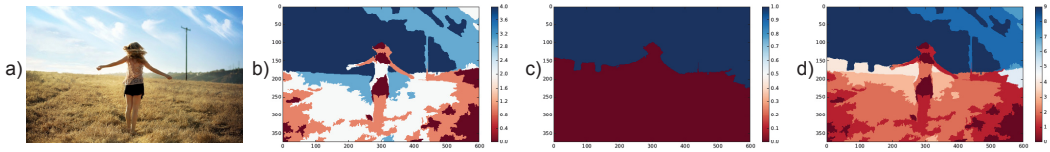


Fig. 2. Wang Emotional Factors: a) Original Image. b) Lightness. c) Warm-Cool. d) Lightness-Warm-Cool

histogram. *Feature Two* provides a description of warm or cool regions with respect to different saturations as well as a description of contrast (Figure 2c). *Feature Three* combines lightness contrast with an sharpness estimation (Figure 2d). A no-reference perceptual blur measure [Crete et al. 2007] was used to calculate the sharpness. The contrast description of the third factor overlaps with the *Itten contrasts* and is omitted.

Itten's Contrasts (IC) are a set of art-theory concepts defined by Johannes Itten [Itten and Van Haagen 1973] for combining colors to induce emotions. The contrasts are based on an proportional opponent color model with 180 distinct colors that are mixtures from 12 pure colors. The contrast calculation is aligned to the method presented in [Machajdik and Hanbury 2010] which is based on Wang's feature extraction [Wei-ning et al. 2006]. Instead of the described waterfall segmentation we used the Quick Shift approach by [Vedaldi and Soatto 2008], 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*. For the missing descriptions of aggregation methods used to calculate the distinct values from the fuzzy membership functions of [Wei-ning et al. 2006], we used the average value.

Color Names (CN) 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 video frame 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) [Zuiderveld 1994] to the value channel V using a region size of 22x22 pixels and a clip limit of 1.0. Saturation and color enhancement was applied similarly to the corresponding channels with slightly adapted values. Figure 1 shows the effects before and after enhancement. *Color Quantization* to reduce the number of distinct colors to a desired palette is obtained by applying *error diffusion* or *dithering* 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. Figure 1d) shows a quantized image using ordered dithering compared to a naive nearest-neighbor-match approach in Figure 1a). *Feature Calculation* is concluded by calculating the statistical moments mean, median, variance, min, max, skew and kurtosis of the reduced palette.

Lightness Fluctuation Patterns (LFP) are calculated analogous to the music feature Rhythm Patterns (RP) [Lidy and Rauber 2005]. In a first step each frame of a music video is converted to the perceptually uniform LAB color space. This corresponds to the psychoacoustic transformations applied to the audio data as a pre-processing step of RP feature calculation. Then 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.

7. VISUAL OBJECT VOCABULARY

In this section we present a high-level approach to facilitate visual stereotypes in music videos based on visual concept detection. It is based on the assumption that music video directors make extensive use of genre related items, apparels or sceneries to outline the video's reference to that specific music genre. A good example is the cowboy hat as a reference to country music. Recently the attention of the computer vision research community has been attracted by the success of deep convolutional neural networks (CNN) [Krizhevsky et al. 2012]. These new approaches made remarkable improvements in visual concept detection, now facilitating comprehensive image understanding. Using these visual computing approaches it is possible to decompose the semantic content of music video into concrete encapsulated concepts such as guitars, vehicles, landscapes, etc. and to measure their frequency distribution over video frames.

7.1. Visual Feature extraction

New machine learning frameworks provide means for sharing trained and evaluated models enabling the concept of transfer learning [Pan and Yang 2010] by transferring the knowledge of one task to other tasks with little or no prior knowledge such as labeled data. The presented approach utilizes the Caffe deep learning framework [Jia et al. 2014] developed by the Berkeley Vision and Learning Center[4]. This framework provides a series of convenient advantages. Besides its fast computational capabilities, its openness and community engagement, one of its remarkable contributions is a sharing platform for trained models called *Model Zoo*. Researchers of all communities are encouraged to upload their models so they can be re-used and applied to different domains. This concept is also referred to as *Transfer Learning* [Pan and Yang 2010], where learning in a new task is improved through the transfer of knowledge from a related task that has already been learned. In reference to this publication this means harnessing the learned models of the visual computing domain to learn semantic descriptions of music videos to approach high level music concepts such as music genre. The outlined approach is based on a pre-trained deep convolutional neural network which won the ILSVRC-2012 image classification task [Krizhevsky et al. 2012] achieving a top-5 test error rate of 15.3%. The model is online available[5] and consists of eight learned layers, five convolutional and three fully-connected.

ImageNet [Russakovsky et al. 2015] is an image database organized according to the WordNet [Miller 1995] hierarchy in which each node of the hierarchy is depicted by hundreds and thousands of images. The dataset currently consists of more than 10,000,000 labeled images depicting more than 10,000 object categories - also called 'synonym sets' or 'synsets'. The images were collected from the web and labeled by human labelers using Amazon's Mechanical Turk crowd-sourcing tool. Figure 3 illustrates example synsets with corresponding images. Each synset is a collection of dozens to thousands of images in different resolutions and variants. As illustrated in Figure 3 the synset *Cowboy hat* includes images of hats as well as people wearing hats. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky et al. 2015] is an annual benchmarking campaign to steer the competition in visual concept recognition and retrieval development. The image classification task, as one of three tasks, focuses on the ability of algorithms to identify objects present

[4]http://bvlc.eecs.berkeley.edu/
[5]https://github.com/BVLC/caffe/wiki/Model-Zoo

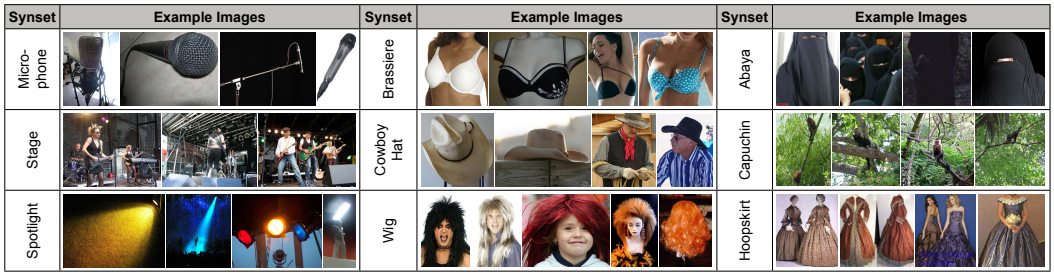


Fig. 3. Example ImageNet Synsets which are also included in the ILSVRC2012 competition categories.

in images of a subset of 1000 synsets. These 1000 synset are used as visual object vocabulary to semantically analyze music videos. A full description of the ILSVRC-2012 categories is provided online[6] and consists to approximately 40% of animal categories. The remaining synsets are objects of daily life such as clothing, food, means of transportation, electronic devices, landscapes as well as music instruments.

Places205 [Zhou et al. 2014] is a CNN trained on 205 scene categories of Places Database (used in NIPS'14) with 2.5 million images. The architecture is the same as Caffe reference network.

To extract the visual features the CNN model is applied to every frame of a video to retrieve the predicted probability values for the visual concepts. The sum of all their values equals 1, resulting in a uniform feature vector. To compare music videos of different length these vectors need to be aggregated into representational vector for the corresponding video. The identification of appropriate aggregations was part of the evaluation and will be discussed in Section 8.4.

8. EVALUATION

The need for multi-modal strategies for MIR has been expressed by [Liem et al. 2011; Weninger et al. 2012; Urbano et al. 2013] and the MIRES roadmap for music information research [Serra et al. 2013] identified music videos as a potential rich source of additional information. Yet, no relevant work has been reported on evaluating the performance of visual features extracted from music videos towards describing music related properties. In this section we intend to close this gap by describing a series of performance evaluations of the previously introduced audio and visual features performed on the MVD. These evaluations are centered around the assumption that information provided by the visual layer of music videos is music related and can be harnessed to describe music content.

8.1. General Experimental Setup

We choose MGR as a proxy problem to perform the subset related evaluation tasks. Feature and system performance is evaluated in terms of classification accuracy.

Audio Feature Extraction was based on audio files from the separated audio channel of the music videos. FFMPEG[7] (version 2.1) was used to extract and store the audio streams in mp3-format using a samplerate of 44.100 Hz and a bit rate of 128 kbit/s.

Visual Feature Extraction was based on frame wise processing of each video. General image pre-processing included the removal of *Letterboxing* or *Pillarboxing* which describes the process of artificially applying black bars at the borders of video frames

[6]http://image-net.org/challenges/LSVRC/2012/browse-synsets
[7]http://www.ffmpeg.org/

to transfer them into either a wide-screen or standard television aspect ratio (see Figure 1a). Removing this often applied stylistic effect is a normalization step to enhance comparability of videos with and without bars (see Figure 1b).

Experimental results of all sections were obtained using the Weka machine learning toolkit [Hall et al. 2009] (version 3.7.5) based on the following set-up: Stratified 10-fold cross-validation was used to evaluate the mean accuracy of ten repeated runs of the separate feature sets on each of the sub-datasets using the following three classifiers: *Support Vector Machine (SVM):* linear SVM and complexity parameter c=1; *K-Nearest Neighbors (KNN):* with k=1, using Euclidean Distance (L2); *Naive Bayes (NB):* simple probabilistic classifier based on applying Bayes' theorem

8.2. Performance Evaluation of the Audio Content Based Features

The Music Video Dataset was compiled to foster the development and evaluation of visual features that are able to capture the subtle relationship between the visual layer of music videos and the underlying music itself. In this sense the audio classification results presented in the upper part of Table II (a) serve two purposes. First, they provide a general conspectus of the MVD and prove that the intended requirements towards the MVD's subsets were met. Second, they serve as a baseline for all consecutive evaluations to compare the performance of the visual features against the audio-only classification results. We provide results for MFCC due to their popularity in music research domains and their advanced quantity of evaluations [Siedenburg et al. 2016]. We further included Chroma features because they provide an abstract kind of harmonic description and are frequently used in different tasks such as audio fingerprinting [Miotto and Orio 2008] or synchronization of audio with music scores [Ewert et al. 2009]. The psychoacoustic features were included due to their reported advantage in classification tasks [Lidy and Rauber 2005]. The aggregation of the *MVD-VIS* subset aimed at clearly defined and well differentiated classes. Classification accuracies provided in Table II (a) show that these requirements have been met which can be observed by the high accuracies for the combined feature-sets *RP-TRH-TSSD* (a10). The high accuracies of the distinct feature-sets *RP* (a4) and *SSD* (a3) conclude that the selected classes are well differentiated by spectral and rhythmical characteristics - at least in terms that are captured by the corresponding feature sets. The weak definition of the *MVD-MM* classes is observable as well. Analysis of the confusion matrices showed extensive mutual confusions of the classes *Hard Rock*, *Pop Rock* and *Indie*. These confusions stretch out to *Metal* and *Country* for the combined *MVD-MIX* dataset. The presented results are comparable to the evaluation provided by [Schindler and Rauber 2014] where the same feature sets had been evaluated on the four de facto MIR music classification benchmark sets. Although (a10) provides best results there is no significant improvement ($p < 0.05$) over (a9) which has half the dimensions of (a10), thus we will use (a9) for further evaluations.

8.3. Performance Evaluation of the low-level color based affective features

This evaluation should serve as a comprehensive bottom-up evaluation to investigate the performance of low-level image features in describing the music related semantic content of music videos. The evaluation is based on the seven color and illumination based image processing feature sets described in Section 6. An initial evaluation of this low-level visual features had been provided in [Schindler and Rauber 2015], but only the performance of the combined visual feature space was evaluated. This evaluation extensively discusses the performance of the distinct color related image features for which the results are presented in Table II (b). The results indicate that Color Names (CN), Wang Emotional Factors (WEF) and Global Color Statistics (GCS) perform better

Table II. Classification results for audio, visual and audio-visual features showing accuracies for Support Vector Machines (SVM), K-Nearest Neighbors (KNN) and Naive Bayes (NB) classifiers. Bold-faced values highlight improvements of audio-visual approaches over audio features. Bold values in the *Audio* and *Visual Concepts* sections depict top-results for the corresponding classifier. Bold values in the *Audio-Visual* section depict improvement over the corresponding audio-only results. Underlined values are significant on a 0.05 level.

			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
(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**
(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**
$v_{pl}1$	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
(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	55.28	**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
(e) Audio-Visual Combinations											
$av1$	$a10$+$v_{in}5$	5036	**96.73**	**81.13**	65.00	**81.60**	**55.73**	49.31	**86.73**	**59.01**	47.48
$av2$	$a9$+$v_{in}5$	3608	**95.63**	77.05	64.16	**77.83**	49.54	**46.58**	**79.44**	51.31	**43.71**
$av3$	$a9$+$v_{pl}5$	2118	**94.50**	**79.95**	**68.08**	**72.96**	**53.29**	**45.99**	**77.40**	**53.73**	**45.51**
$av4$	$av2$+$v_{pl}5$	4118	**95.76**	75.76	61.00	77.55	**50.31**	44.59	**80.16**	**52.43**	41.79
$av4$	$a6$+$v_{in}5$	3176	**94.65**	68.61	63.64	**78.49**	53.01	50.41	**82.62**	**48.94**	**48.53**
$av5$	$a4$+$v_{in}5$	3440	**91.24**	68.80	**63.40**	**71.95**	**43.78**	**44.86**	**74.14**	**45.53**	**42.69**
$av6$	$a3$+$v_{in}5$	2168	**89.85**	62.11	57.89	**70.13**	43.16	42.93	**70.30**	37.98	38.88

in discriminating the different classes but are not reliable to describe music genres. However it was able to confirm the common stereotype of "darkness" related to *Heavy Metal* music [Farley 2009]. Videos of the class *Metal* contain significantly more black pixels (independent t-test, $p < 0.05$) than other classes of the *MVD-MIX* subset and the second smallest brightness values. The lowest values on *brightness* and *colorfulness* for *Opera* are a strong indicator for the inferior lightning conditions at such venues and thus have no relation to the music itself.

8.4. Performance Evaluation of the Visual Vocabulary

This evaluation focuses on the visual concept detection based approach presented in Section 7. The rational behind this evolution is to estimate if the high-level semantic concepts extracted from the frames of the music videos contain relevant information to discriminate the classes of the MVD. Part of the evaluation is to assess appropriate methods to aggregate the feature responses of the music video frames into a single feature vector. The resulting vector is the representative descriptor for the corresponding music video. This single vector representation is a requirement of utilized machine learning algorithms and further abstracts from variations in length. Seven statistical moments (minimum, maximum, mean, median, standard deviation, variance, kurtosis, skewness) were analyzed by testing all possible combinations. The best-performing results using *Visual Concepts* are listed in Table II (c). Results show high classification accuracies for visual vocabularies based on the *ImageNet* model. The most relevant aggregations for this model are *MAX* ($v_{in}3$) and *STD* ($v_{in}2$). This can explained by the type of feature response which is the probability of a certain concept of the vocabulary to be present in a frame. In that sense the *MAX* aggregated features represent which concepts had high probability values and thus 'reliably' appeared in at least one frame of the video. This information appeared to be most discriminative with an accuracy of 73.15% on the *MVD-VIS* dataset ($v_{in}3$). The best performing combination ($v_{in}5$) reached 74.36% which is not significantly better than ($v_{in}3$) ($p < 0.05$). The visual vocabulary results for the *Imagenet* model improved the previously reported accuracy of 50.13% ($v_{co}8$) using low-level visual features [Schindler and Rauber 2015] by 24.23%. This high accuracy supports the initial hypothesis (see Section 1) that music videos make use of easy identifiable visual concepts. The low improvement of the combination with color features ($vc1$) and ($vc5$) indicates that this information is provided by high semantic concepts (e.g. apparel, buildings, music instruments, vehicles, etc.). We will discuss this further in the analysis of visual stereotypes in Section 8.7. Comparing the results of the visual vocabulary based approach with state-of-the-art audio features shows that the described visual approach outperforms the music features *Chroma* and the de-facto standard feature *MFCC*.

8.5. Performance Evaluation of Audio-Visual Combinations

The rational behind a multi-modal approach is to utilize information of different modalities to improve the performance for a dedicated task. In the previous sections the two modalities *audio* and *video* have been evaluated separately. This evaluation attempts to answer the question, if their combinations can improve classification accuracy, using the *audio* results (see Table II (a)) as baseline. Again different combinations of audio features and visual vocabulary aggregations were evaluated in classification experiments. To reduce the number of required experiments, weak performing aggregations identified in the previous task have been skipped. Only *MEAN*, *MAX* and *STD* aggregations have been used. Feature-sets have been combined using an early fusion approach. The *Audio-Visual* section of Table II shows selected best-performing combinations for the *SVM* classifier. The additional information can be harnessed well by the *SVM* classifier where all results showed noticeable and some remarkable im-

Table III. Classification results for cross-genre music themes evaluation using Support Vector machines and 10-fold cross-validation. Values represent accuracies for the corresponding theme. Columns titled with one of the *MVD-{VIS,MM,MIX}* subsets' names represent results where the theme had been added as additional class to the corresponding set. Columns marked with *TH* represent results where only the four classes of the MVD-THEMES dataset have been used in the experiments.

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

provements over the baseline. Despite the high degree of optimization of the *MVD-VIS* subset, an improvement of 2.94% ($av1$) was accomplished over the best performing audio combination ($a10$). An even higher improvement was observed for the less differentiated set *MVD-MM* (+6.84%) and the bigger *MVD-MIX* dataset (+10.82%). The visual information ($v_{in}5$) showed outstanding improvements for the *TSSD* ($av4$) audio feature set. Improvements of 7.84% on the optimized which is still significantly lower than the best performing combination ($av1$) but not as the combination with similar vector dimensions ($av2$). The highest improvement of 16.43% for ($av4$) was observed for the bigger *MVD-MIX* dataset. All mentioned improvements are significant ($p < 0.05$).

8.6. Analysis of non-audible Music Themes

This evaluation attempts to answer the question, if the presented multi-modal approach can improve the classification performance of cross-genre music themes such as *Christmas*? The task corresponding to this evaluation refers to music tagging and is aligned to the MusiClef multi-modal music tagging task [Orio et al. 2012]. The evaluation was performed in two ways. In a general experiment the accuracy for discriminating the classes of the *MVD-THEMES* dataset was evaluated. In a second experiment each theme was added separately to the datasets *MVD-VIS,MM,MIX* and the accuracy for discriminating this theme from the other classes was measured. This refers a common tagging scenario where audible and non-audible semantic tags are mixed-up. The audio results again serve as baseline. The represent the problem of audio-content based features towards describing non-audible concepts. Classification accuracies for all themes, especially combined the larger *MVD-MIX* dataset, are low. Except for *Broken Heart* almost all visual vocabulary based approaches already performed better than the audio-only results such as *Christmas* combined with *MVD-MIX* (+34,5%). More remarkable for the same combination is the observed improvement of 45.5% using the introduced audio-visual approach. Similar high improvements were observed for the themes *K-Pop* and *Protest Song*. Only the heavily spreading theme *Broken Heart* showed only improvements in the discrimination of different themes but almost none for the mixed-up experiments.

8.7. Analysis of Visual Stereotypes

The final evaluation is concerned with the analysis if the extracted features are able to capture genre or thematic related visual stereotypes. We calculated the term frequencies of the *ImageNet* Synsets for each class. For each concept of a class the largest minimal difference to the term frequencies of the other classes was calculated. This resulted in a list of the most salient visual concepts for each genre. Table IV provides an overview of a small selection of the genres. For the discussion of visual stereotypes we removed Synsets such as *Abaya* (see Figure 3) which are weakly defined and are often assigned to dark or blurred video frames. We start the discussion of visual stereo-

Table IV. Salient ILSVRC Synsets descendingly ranked by their minimal difference to other genres.

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

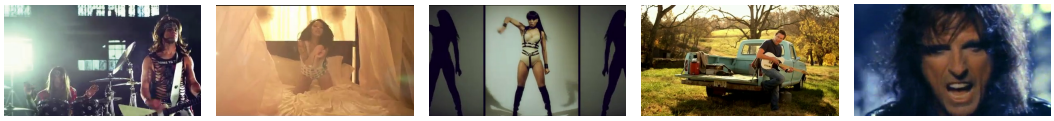
types with the genre *Country* for which the *cowboy hat* and utility vehicles such as the *pickup truck* or a *tractor trailer* are highly salient concepts. These observations correspond to an evaluation of perceived extramusical associations with country music [Shevy 2008] and to aesthetic description provided in [Frith et al. 2005]. It was also possible to confirm the mentioned *"movement towards the warm, orange tones that became the dominant 'look' of many contemporary country videos."* [Frith et al. 2005]. The average values for *red* and *yellow* are significantly highest and second highest for this genre. The second stereotype addresses the over-sexualization of contemporary popular *Dance* music [Hall et al. 2012]. Including the most salient concept *Brassiere* (see Figure 3) eight of the provided top ranking examples referred to revealing clothes, closeups of body parts or people dancing on *poles*. Figure 4 provides two examples of dance video frames. For Heavy *Metal* music we already confirmed the common stereotype of "darkness". The Synset *matchstick* refers to all kinds of fire which is a typical element related to *Metal* [Farley 2009]. Further common concepts such as *spotlight*, *electric guitar*, *drums* and *loudspeaker* are performance related elements. For *Opera* a high number of top-ranking classical instruments related concepts can be observed.

8.8. General Discussions

Figure 4 illustrates the top-seven synsets of five example music video frames ranked by their estimated probability. Although, weakly defined Synsets such as *Abaya* (see Figure 3) suffer from semantic significance related to music content,we would like to mention that such concepts were observed to be highly discriminative because they capture video related properties such as low saturation or blurred frames due to fast motion or panning. The salient Synsets for *Reggae* music exhibits many animal related concepts. This is an artifact of the used model. Reggae videos are often shot at seashores or in tropical landscapes. This is reflected by the high number of animals detected as salient concepts. Although these animals do not appear on the music videos, the images used in the training set show similar landscapes (see example *Capuchin* in Figure 3) in the background which the model trained on.

9. CONCLUSIONS AND FUTURE WORK

This paper introduced semantic analysis of music video content. The focus of the presented research was on harnessing the information provided by the visual layer to approach the Music Information Retrieval problem space. Different feature-sets were analyzed according different tasks. The results show that high-level concept detection approaches based on convolutional neural networks not only outperform traditional low-level image features, but also are superior to audio-content based descriptors in semantic music tagging tasks. The introduced audio-visual approaches improve to audio-



stage	0.3162	mosquito net	0.0932	maillot	0.2745	lumbermill	0.1925	wig	0.4399
electric guitar	0.1169	wardrobe	0.0857	bolo tie	0.0732	tow truck	0.1215	neck brace	0.0577
bassoon	0.0649	brassiere	0.0815	Windsor tie	0.0550	harvester	0.1152	chimpanzee	0.0418
accordion	0.0611	shower curtain	0.0471	letter opener	0.0486	thresher	0.0513	hair spray	0.0375
drumstick	0.0386	candle	0.0400	brassiere	0.0390	jeep	0.0484	orangutan	0.0366
microphone	0.0313	plastic bag	0.0204	bikini	0.0384	half track	0.0473	cloak	0.0267
marimba	0.0276	hoopskirt	0.0187	bassoon	0.0364	pickup truck	0.0460	Windsor tie	0.0236

Fig. 4. Example frames of music videos and their top-ranked synsets ordered by descending probability.

only baseline by up to 16.43% for genre classification and up to 45.5% for thematic music tagging tasks. The evaluation also showed, that object vocabularies are a good tool to capture the semantic and genre stereotypical information of music videos. Future work will focus on training specialized convolutional neural networks that include a wider range of music and genre related concepts.

References

Esra Acar, Frank Hopfgartner, and Sahin Albayrak. 2014. Understanding Affective Content of Music Videos through Learned Representations. In *MultiMedia Modeling*. Springer, 303–314.

Eric Brochu, Nando De Freitas, and Kejie Bao. 2003. The sound of an album cover: Probabilistic multimedia and IR. In *Workshop on Artificial Intelligence and Statistics*.

Rui Cai, Lei Zhang, Feng Jing, Wei Lai, and Wei-Ying Ma. 2007. Automated music video generation using web image resource. In *Acoustics, Speech and Signal Processing. ICASSP*.

Cyril Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*, Vol. 19. 173–194.

Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. 2007. The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Electronic Imaging*.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*. Springer, 288–301.

J Stephen Downie. 2003. Music information retrieval. *Annual review of information science and tech.* (2003).

Peter Dunker, Stefanie Nowak, André Begau, and Cornelia Lanz. 2008. Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach. In *Proc 1st ACM int conf on Multimedia information retrieval*. ACM, 97–104.

Sebastian Ewert, Meinard Müller, and Peter Grosche. 2009. High resolution audio synchronization using chroma onset features. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Helen Farley. 2009. Demons, devils and witches: the occult in heavy metal music. *Heavy metal music in Britain* (2009), 73–88.

Joanna Finkelstein. 2007. *Art of Self Invention: Image and Identity in Popular Visual Culture*. IB Tauris.

Jonathan Foote, Matthew Cooper, and Andreas Girgensohn. 2002. Creating music videos using automatic media analysis. In *Proceedings of the tenth ACM international conference on Multimedia*. ACM.

Simon Frith, Andrew Goodwin, and Lawrence Grossberg. 2005. *Sound and vision: the music video reader*. Routledge.

Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. 2011. A survey of audio-based music classification and annotation. *Multimedia, IEEE Transactions on* 13, 2 (2011), 303–319.

Olivier Gillet, Slim Essid, and Gaël Richard. 2007. On the correlation of automatic audio and visual segmentations of music videos. *Circuits and Systems for Video Technology, IEEE Transactions on* (2007).

Magnus Haake and Agneta Gulz. 2008. Visual stereotypes and virtual pedagogical agents. *Journal of Educational Technology & Society* 11, 4 (2008), 1–15.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11, 1 (2009).

P Cougar Hall, Joshua H West, and Shane Hill. 2012. Sexualization in lyrics of popular music from 1959 to 2009: Implications for sexuality educators. *Sexuality & Culture* (2012).

Allan Hanbury. 2003. Circular statistics applied to colour images. In *8th Computer Vision Winter Workshop*.

Allan Hanbury and Jean Serra. 2002. A 3D-polar coordinate colour representation suitable for image analysis. *submitted to Computer Vision and Image Understanding* (2002).

Xiao Hu and J Stephen Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 159–168.

Xian-Sheng Hua, Lie Lu, and Hong-Jiang Zhang. 2004. Automatic music video generation based on temporal pattern analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM.

Johannes Itten and Ernst Van Haagen. 1973. *The art of color: the subjective experience and objective rationale of color*. Van Nostrand Reinhold New York, NY, USA.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

Wonjun Kim and Changick Kim. 2007. Automatic region of interest determination in music videos. In *41th Asilomar Conf on Signals, Systems and Computers*. IEEE, 485–489.

Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on* 3, 1 (2012), 18–31.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

Paul Lamere. 2008. Social tagging and music information retrieval. *Journal of new music research* 37, 2 (2008), 101–114.

Olivier Lartillot and Petri Toiviainen. 2007. A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects*. 237–244.

Jin Ha Lee, Kahyun Choi, Xiao Hu, and J Stephen Downie. 2013. K-Pop Genres: A Cross-Cultural Exploration.. In *ISMIR*. 529–534.

Jin Ha Lee, J Stephen Downie, and Sally Jo Cunningham. 2005. Challenges in Cross-Cultural/Multilingual Music Information Seeking.. In *ISMIR*. 1–7.

Janis Lıbeks and Douglas Turnbull. 2010. Exploring Artist Image using Content-based Analysis of Promotional Photos. In *Proc Int Computer Music Conf*.

J. Libeks and D. Turnbull. 2011. You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *MultiMedia, IEEE* 18, 4 (April 2011), 30–37.

Thomas Lidy and Andreas Rauber. 2005. Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification.. In *ISMIR*.

Cynthia Liem, Meinard Müller, Douglas Eck, George Tzanetakis, and Alan Hanjalic. 2011. The need for music information retrieval with user-centered and multimodal strategies. In *Proc 1st int ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 1–6.

Beth Logan and others. 2000. Mel Frequency Cepstral Coefficients for Music Modeling.. In *ISMIR*.

Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the international conference on Multimedia*. ACM, 83–92.

Alison Mattek and Michael Casey. 2011. Cross-Modal Aesthetics from A Feature Extraction Perspective: A Pilot Study.. In *ISMIR*.

Rudolf Mayer. 2011. Analysing the Similarity of Album Art with Self-Organising Maps. In *Advances in Self-Organizing Maps*. LNCS, Vol. 6731. Springer.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. 2008. Rhyme and Style Features for Musical Genre Classification by Song Lyrics.

Cory McKay and Ichiro Fujinaga. 2006. Musical genre classification: Is it worth pursuing and how can it be improved?. In *ISMIR*. 101–106.

Leonard B. Meyer. 1956. Emotion and Meaning in Music. (1956).

George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

Riccardo Miotto and Nicola Orio. 2008. A Music Identification System Based on Chroma Indexing and Statistical Modeling. In *ISMIR*.

Keith Negus. 2011. *Producing pop: Culture and conflict in the popular music industry*. out of print.

Bureau of the Census and United States. 2009. *Statistical abstract of the United States*. US Government Printing Office.

Nicola Orio, Cynthia CS Liem, Geoffroy Peeters, and Markus Schedl. 2012. MusiClef: multimodal music tagging task. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on* 22, 10 (2010), 1345–1359.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40, 2 (2000), 99–121.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015).

Shoto Sasaki, Tatsunori Hirai, Hayato Ohya, and Shigeo Morishima. 2015. Affective Music Recommendation System Based on the Mood of Input Video. LNCS, Vol. 8936. Springer International Publishing.

Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. 2006. Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE* 23, 2 (2006), 133–141.

Markus Schedl, Tim Pohle, Peter Knees, and Gerhard Widmer. 2006. Assigning and Visualizing Music Genres by Web-based Co-Occurrence Analysis.. In *ISMIR*. Citeseer, 260–265.

Alexander Schindler. 2014. A Picture is Worth a Thousand Songs: Exploring Visual Aspects of Music. In *Proceedings of the 1st International Workshop on Digital Libraries for Musicology (DLfM '14)*.

Alexander Schindler and Andreas Rauber. 2013. A Music Video Information Retrieval Approach to Artist Identification. In *10th Symposium on Computer Music Multidisciplinary research (CMMR 2013)*.

Alexander Schindler and Andreas Rauber. 2014. Capturing the Temporal Domain in Echonest Features for Improved Classification Effectiveness. LNCS, Vol. 8382.

Alexander Schindler and Andreas Rauber. 2015. An Audio-Visual Approach to Music Genre Classification through Affective Color Features. In *Advances in Information Retrieval*. LNCS, Vol. 9022. 61–67.

Xavier Serra, Michela Magas, Emmanouil Benetos, Magdalena Chudy, S. Dixon, Arthur Flexer, Emilia Gómez, F. Gouyon, P. Herrera, S. Jordà, Oscar Paytuvi, G. Peeters, Jan Schlüter, H. Vinet, and G. Widmer. 2013. *Roadmap for Music Information ReSearch*.

Xi Shao, Changsheng Xu, Namunu C Maddage, Qi Tian, Mohan S Kankanhalli, and Jesse S Jin. 2006. Automatic summarization of music videos. *Trans Multimedia Comp., Communications, and Appl.* (2006).

Mark Shevy. 2008. Music genre as cognitive schema: extramusical associations with country and hip-hop music. *Psychology of music* 36, 4 (2008), 477–498.

Kai Siedenburg, Ichiro Fujinaga, and Stephen McAdams. 2016. A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology. *Journal of New Music Research* (2016).

Bob L Sturm. 2013. Classification accuracy is not enough. *Journal of Intelligent Information Systems* (2013).

Bob L Sturm. 2014. A simple method to determine if a music information retrieval system is a horse. *Multimedia, IEEE Transactions on* 16, 6 (2014), 1636–1644.

George Tzanetakis and Perry Cook. 2000. Marsyas: A framework for audio analysis. *Organised sound* (2000).

George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on* 10, 5 (2002), 293–302.

Julián Urbano, Markus Schedl, and Xavier Serra. 2013. Evaluation in music information retrieval. *Journal of Intelligent Information Systems* 41, 3 (2013), 345–369.

Patricia Valdez and Albert Mehrabian. 1994. Effects of color on emotions. *Journal of Experimental Psychology: General* 123, 4 (1994), 394.

Andrea Vedaldi and Stefano Soatto. 2008. Quick shift and kernel methods for mode seeking. In *Computer Vision–ECCV 2008*. Springer, 705–718.

Carol Vernallis. 2004. *Experiencing music video: Aesthetics and cultural context*. Columbia University Press.

Wang Wei-ning, Yu Ying-lin, and Jiang Sheng-ming. 2006. Image retrieval by emotional semantics: A study of emotional space and feature extraction. In *Int Conf on Systems, Man and Cybernetics*. IEEE.

Felix Weninger, Björn Schuller, Cynthia Liem, Frank Kurth, and Alan Hanjalic. 2012. Music information retrieval: An inspirational guide to transfer from related disciplines. *Dagstuhl Follow-Ups* 3 (2012).

Ashkan Yazdani, Krista Kappeler, and Touradj Ebrahimi. 2011. Affective content analysis of music video clips. In *Int ACM workshop on Music information retrieval with user-centered and multimodal strategies*.

Jong-Chul Yoon, In-Kwon Lee, and Siwoo Byun. 2009. Automated music video generation using multi-level feature-based segmentation. In *Handbook of Multimedia for Digital Entertainment and Arts*. Springer.

Shiliang Zhang, Qingming Huang, Shuqiang Jiang, Wen Gao, and Qi Tian. 2010. Affective visualization and retrieval for music video. *Multimedia, IEEE Transactions on* 12, 6 (2010).

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information Proc. Systems*.

Karel Zuiderveld. 1994. Contrast limited adaptive histogram equalization. In *Graphics gems IV*. 474–485.