# A Picture is Worth a Thousand Songs: Exploring Visual Aspects of Music

Alexander Schindler
Safety & Security Department
AIT Austrian Institute of Technology GmbH
Vienna, Austria
alexander.schindler@ait.ac.at

## ABSTRACT

The abstract nature of music makes it intrinsically hard to describe. To alleviate this problem we use well known songs or artists as a reference to describe new music. Music information research has mimicked this behavior by introducing search systems that rely on prototypical songs. Based on similarity models deducing from signal processing or collaborative filtering an according list of songs with similar properties is retrieved. Yet, music is often searched for a specific intention such as music for workout, to focus or for a comfortable dinner with friends. Modeling music similarities based on such criteria is in many cases problematic or visionary. Despite these open research challenges a more user focused question should be raised: Which interface is adequate for describing such intentions? Traditionally queries are either based on text input or seed songs. Both are in many cases inadequate or require extensive interaction or knowledge from the user.

Despite the multi-sensory capabilities of humans, we primarily focus on vision. Many intentions for music searches can easily be pictorially envisioned. This paper suggests to pursue a query music by image approach. Yet, extensive research in all disciplinary fields of music research, such as music psychology, musicology and information technologies, is required to identify correlations between the acoustic and the visual domain. This paper elaborates on opportunities and obstacles and proposes ways to approach the stated problems.

## 1. INTRODUCTION

In the second part of the last century the visual representation has become a vital part of music. Album covers grew out of their basic role of packaging to become visual mnemonics to the music enclosed [1]. Stylistic elements emerged into prototypical visual descriptions of genre specific music properties. Initially intended to aide or sway customers in their decision of buying a record, these artworks became an influential part of modern pop culture. The "look of music" became an important factor in people's appreciation of music and the rise of music videos in the early 80s provided further momentum to this development. In that period the structure of the prevalent music business changed completely by shifting the focus from selling whole records to hit singles [2]. Their accentuation and the new accompanied intensification of celebrity around pop music played an important role in the development of the *pinup culture*. This emphasis on visual aspects of music transcended from topics directly connected with music production into aspects of our daily life. The relationship to fashion tightened, yielding to different styles that discriminated the various genres. An early study on music video consumption among young adolescents revealed that one of the primary intentions was social learning such as how to dress, act and relate to others [3]. How much influence visual music had on us over half a century is hard to evaluate, but we grew accustomed to a visual vocabulary that is specific for a music style in a way that the genre of a music video can often be predicted despite the absence of sound (see Figure 1).

*Figure 1:* Examples of music genres that are easy to identify in images - a) Dance, b) Rock, c) Heavy Metal, d) Rap

The influence and outreach of music videos is on a constant rise as recently reported. In 2011 we conducted a survey among decision makers and stakeholders in the music industry which showed that YouTube is considered to be the leading online music service [4]. A different survey by Nielsen [5] involving 26,644 online consumers in 53 international markets revealed that '*watching music videos on computer*' was the most mentioned consuming activity, practiced by 57% of these consumers in the 3 month preceding the survey. 27% mentioned to also have '*watched music videos on mobile devices*' in the same period. Stowell and Dixon [6] identified YouTube as one of the most important technologies used in secondary school music classes. The authors suggest to offer MIR analysis for videos to enable open-ended exploration. Cunningham et al. [7] identified music videos as a passive trigger to active music search and state that visual aspects of music videos can strongly influence the amount of attention paid to a song.

Harnessing the visual layer of music videos provides a wide range of possible scenarios. At first hand it 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. On another hand it may provide new innovative approaches to search for or organize music. One intention is to use visual stimuli to query for music. This would enhance the traditional search spaces that are either based on audio content, textual or social data or a hybrid mixture of those. To use the visual domain to search for music it has to be linked to the acoustic domain. This requires profound understanding of audio-visual correlations. An apparent approach to identify such interrelationships is to analyze music videos. In a previous study we have shown that artist identification [8] can be significantly improved by visual features. This paper discusses an integrated view of music as a holistic combination of audio and video and outlays opportunities.

# 2. OBJECTIVES

This section outlines objectives and obstacles of an integrated audio-visual approach to music. Challenges for a selected set of music research disciplines are discussed.

## 2.1 Musicology

Research on audio-visual correlations related to music is currently gaining attention in some sub-disciplines of musicology. Yet, profound insights are required to derive generalized models from.

**Music Psychology:** including *Cognitive musicology* and *Cognitive neuroscience of music*. While the effects of music on film or movies are well known and studied [9] the effects of visual stimuli on the interpretation of music are yet beginning to attract attention [10]. Recently it has been discovered that visual cortex processes auditory information too [11]. Results from music psychological experiments may influence future approaches to *Music Therapy*.

**Popular Music Studies:** Visual music is a part of popular music. Audio-visual analysis could be applied to identify stylistic trends or influences (e.g. mass analysis of album art images).

**Sociomusicology:** This discipline traditionally uses a wide range of research methods to observe behavior and sociomusical interactions. Image processing could provide further insights by analyzing non-accoustical music references such as clothes or images in public spaces.

## 2.2 Digital Libraries

Applying audio-visual retrieval models to digital libraries provide new possibilities to organize and search their collections.

**Holistic view of Cultural Heritage Collections:** Most of the digital libraries provide separated access to their collections. As suggested in [12] audio-visual content can be used to enhance the learning experience for music in general. Analyzing artwork [13] could automatically provide extended meta-data that can be used to interlink images with recordings of classical music by means of similar properties. State-of-the-art image processing technology already provides means to extract features such as epoch, instrumentation, expression, etc.

**Extended Search Terms:** MIR technologies aim to assist digital libraries by providing new approaches for searching and exploring their collections. Yet most of these solutions are limited by the shortcomings of the two major approaches that are either based on audio content analysis or collaborative filtering. Except for user tag mining, categorizing music by terms that are not detectable in the audio signal or through social relationships of the listeners, is currently not feasible (e.g. Violence, Christmas, Party, Sex, or by instruments such as timpani, trumpet, etc.).

## 2.3 Music Information Retrieval

The need for multi-modal strategies for music information retrieval has strongly been argued by Liem et al. [14]. The MIRES roadmap for music information research [15] further identifies music videos as a potential rich source of additional information.

### *Sound and Visual Correlations.*

Music video directors and editors use a wide range of visual effects to create and stir emotions, to draw attention, to manipulate our perception or interpretation of the corresponding song.

**General Audio-Visual Correlations:** Color for example is a popular mean of expressing emotions. Movie directors pick warm colors for romantic scenes while using cold ones for fear and anxiety. Songwriters use similar effects to express emotions. Major chords are perceived happy, minor chords sad, dissonance creates an impression of dis-comfort, etc. Although, MIR can provide empirical evaluations of visual stimuli used in music videos or other music related visual media, it should be considered to incorporate domain knowledge from music psychology to build more reliable models.

**Lyrics and Visual Cues:** Narrative music videos provide visual interpretations of a song's lyrics. Previous cross-modal music categorization approaches including lyrics reported that a genre specific vocabulary can be identified (e.g. police, gun, jah, etc.) [16, 17]. Further multimedia-based evaluations should concentrate on the correlation of such keywords with visual cues.

### *Structural Analysis.*

Structural analysis is a fundamental predecessor of many analytical tasks. The goal is to automatically detect compositional elements in the audio content (e.g. verse, chorus, bridge, etc.).

**Music Segmentation and Summarization:** Music segmentation and summarization tries to identify coherent blocks within a composition such as verse or chorus. Typical approaches are based on self-similarities of sequentially extracted audio features. If no timbral or rhythmical differences can be detected, segmentation fails. The structure of a music video is often aligned to the compositional or lyrical properties of a track. Scenery or color changes depict the transition to another segment. Such transitions can be detected [18] and analyzed. Repetitions of the same images through the video clip for example may suggest repetitions of sections such as the chorus.

**Tempo and Rhythmic Analysis:** Current beat tracking approaches [19] rely on strong contrasting musical events to correctly estimate the track's tempo. Shot editing of music video can serve as valuable information in cases where audio based approaches fail. Further detecting and classifying dance motions can provide additional data for rhythmic analysis.

*Classification.*

Classification experiments play a significant role on Music Information Retrieval. The following tasks are prevalent in the Music Information Retrieval Evaluation eXchange (MIREX) [20]:

**Genre Recognition:** Automatic genre recognition has been extensively studied in the MIR domain [21, 22, 23]. A range of visual cues to describe abstract musical expressions (e.g. sunsets for romantic songs, gnashing teeth to express aggression, etc.) are frequently used in media. As a consequence the genre of music videos can often be recognized without hearing the corresponding song. Recent cross-modal approaches reported improved accuracy in detecting similar visual concepts (e.g. sky, people, plant life, clouds, sunset, etc.) [24]. It has to be evaluated which cues are discriminative for each genre and which state-of-the-art technology from the MIR domain can be applied to reliably extract this information.

**Artist Identification:** Artist recognition is an important task for music indexing, browsing and retrieval. Using audio content based descriptors only is problematic due to the wide range of different styles on either a single album or more severe throughout the complete works of an artist. Current image processing technologies (e.g. object detection, face recognition) could be used to either identify the performing artist within a video and recognize the corresponding name.

**Instrument Detection:** Instrument detection is an ambitious task currently suffering from poor performance of source separation approaches. Many music videos show the performing artists playing their instruments. This information can be used to augment audio based classification approaches.

**Emotion Recognition:** Visual arts have developed a wide range of techniques to express emotions or moods. Such techniques have to be identified, evaluated and modeled to harness their information for effective emotion recognition. It has to be evaluated to which extent the visual techniques of expressing emotions and those of music correlate. Such correlations may be used to combine the acoustic with the visual domain.

*Similarity.*

Definitions of similarity of music videos are yet missing. It is even undecided if similarity should be based on acoustic or visual properties or on a combination of both. Descriptions of music videos intuitively refer to visual cues. An extensive discussion including results from user evaluations should precede attempts to measure similarity between music videos.

## 2.4 General Objectives

This section envisions application scenarios that could be accomplished through the suggested approach.

**New Innovative Music Search:** Visual clues identified and extracted from music videos, such as correlations between color and music, may promote the development of new search methods for music or sounds. Traditionally textual or audio content descriptions of seed songs are used to query for music. Although not applicable to all kinds of music, the intention for a search is often better described through a picture. Based on insights from the suggested research, a music query using a picture of glass of wine and burning candles might return a collection of romantic music (see Figure 2).

*Figure 2:* Example images as input for music search algorithms.

**Music Visualization:** Identified audio-visual correlations can be used to improve music visualization or automatic lighting systems. LED-based color-changing lights are currently gaining popularity. An add-on to such systems could change the color according to musical transitions and mood.

**Music for Movies or Advertisements:** An investigation about the usage of music in films and advertising [25] suggests that a search function matching video features to music features would have potential applications in film making, advertising or similar domains. Understanding audio-visual correlations, especially in their cultural context, facilitates recommender systems suggesting music for discrete movie sequences.

**Adaptive Music Recommendation:** Similar to recommending music for videos, cameras can be used to observe the ambiance of an area. Based on the visual input from the camera, high level features such as number and motion of people, clothing, color of the room, weather conditions, sunrise, sunset, etc. can be used to select appropriate music.

# 3. CONCLUSIONS AND FUTURE WORK

This paper addresses the facilitation of visual music as a future grand challenge for various music related research disciplines. Objectives are discussed for Musicology, Digital libraries and Music Information Retrieval. The main obstacle to the proposed applications is the lack of sufficient cross-modal studies. State-of-the-art multi-modal retrieval approaches need to be analyzed towards their applicability to music classification and retrieval. As future work we plan large scale analysis of audio-visual correlations, by applying content based affect recognition methods from the image retrieval domain to album art images and music videos and comparing the estimated response to emotions extracted from audio content.

# 4. REFERENCES

# 5. REFERENCES


[1] S. Jones and M. Sorger, "Covering music: A brief history and analysis of album cover design," *Journal of Popular Music Studies*, vol. 11, no. 1, pp. 68–102, 1999.

[2] S. Frith, A. Goodwin, and L. Grossberg, *Sound and vision: the music video reader*. Routledge, 2005.

[3] S.-W. Sun and J. Lull, "The adolescent audience for music videos and why they watch," *Journal of Communication*, vol. 36, no. 1, pp. 115–125, 1986.

[4] T. Lidy and P. van der Linden, "Report on 3rd chorus+ think-tank: Think-tank on the future of music search, access and consumption," tech. rep., MIDEM 2011. Tech report, CHORUS+ EU Coord Action on Audiovisual Search, Cannes, France, 2011.

[5] Nielsen, "Digital music consumption and digital music access," 2011.

[6] D. Stowell and S. Dixon, "Mir in school? lessons from ethnographic observation of secondary school music classes," in

*Proc 12th Int Society for Music Information Retrieval Conf*, (Miami (Florida), USA), pp. 347–352, October 24-28 2011.

[7] S. J. Cunningham, D. Bainbridge, and D. Mckay, "Finding new music: A diary study of everyday encounters with novel songs," in *Proc 8th Int Society for Music Information Retrieval Conf*, (Vienna, Austria), 2007.

[8] A. Schindler and A. Rauber, "A music video information retrieval approach to artist identification," in *10th Symposium on Computer Music Multidisciplinary research (CMMR 2013)*, 2013.

[9] A. J. Cohen, "How music influences the interpretation of film and video: Approaches from experimental psychology," *Selected reports in ethnomusicology: perspectives in systematic musicology*, vol. 12, no. 1, pp. 15–36, 2005.

[10] M. M. Marin, B. Gingras, and J. Bhattacharya, "Crossmodal transfer of arousal, but not pleasantness, from the musical to the visual domain.," *Emotion*, vol. 12, no. 3, p. 618, 2012.

[11] P. Vetter, F. W. Smith, and L. Muckli, "Decoding sound and imagery content in early visual cortex," *Current Biology*, vol. 24, no. 11, pp. 1256 – 1262, 2014.

[12] C. Ribeiro, G. David, and C. Calistru, "Multimedia in cultural heritage collections: a model and applications," in *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*, pp. 186–195, Springer, 2007.

[13] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the international conference on Multimedia*, pp. 83–92, ACM, 2010.

[14] C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "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*, pp. 1–6, ACM, 2011.

[15] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordà, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer, *Roadmap for Music Information ReSearch*. 2013.

[16] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal music mood classification using audio and lyrics," in *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, pp. 688–693, Dec 2008.

[17] R. Mayer, R. Neumayer, and A. Rauber, "Combination of Audio and Lyrics Features for Genre Classification in Digital Audio Collections," in *Proceedings of the ACM Multimedia 2008*, pp. 159–168, ACM New York, NY, USA, October 27-31 2008.

[18] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Video shot detection and condensed representation. a review," *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 28–37, 2006.

[19] M. F. McKinney, D. Moelants, M. E. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *Jour. New Music Res.*, vol. 36, no. 1, pp. 1–16, 2007.

[20] J. S. Downie, K. West, A. Ehmann, E. Vincent, *et al.*, "The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview," in *6th Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 320–323, 2005.

[21] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[22] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Adaptive Multimedia Retrieval*, 2012.

[23] A. Schindler and A. Rauber, "Capturing the temporal domain in echonest features for improved classification effectiveness," in *Adaptive Multimedia Retrieval*, 2012.

[24] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative," in *Proceedings of the international conference on Multimedia information retrieval*, pp. 527–536, ACM, 2010.

[25] C. Inskip, A. Macfarlane, and P. Rafferty, "Music, movies and meaning: Communication in film-makers' search for pre-existing music, and the implications for music information retrieval," in *Proc 9th Int Conf on Music Information Retrieval*, (Philadelphia, USA), pp. 477–482, September 14-18 2008.