From Water Music to 'Underwater Music': Multimedia Soundtrack Retrieval with Social Mass Media Resources*

Cynthia C. S. Liem

Multimedia Computing Group, Delft University of Technology, The Netherlands
c.c.s.liem@tudelft.nl

Abstract. In creative media, visual imagery is often combined with music soundtracks. In the resulting artefacts, the consumption of isolated music or imagery will not be the main goal, but rather the combined multimedia experience. Through frequent combination of music with non-musical information resources and the corresponding public exposure, certain types of music will get associated to certain types of non-musical contexts. As a consequence, when dealing with the problem of soundtrack retrieval for non-musical media, it would be appropriate to not only address corresponding music search engines in music-technical terms, but to also exploit typical surrounding contextual and connotative associations. In this work, we make use of this information, and present and validate a search engine framework based on collaborative and social Web resources on mass media and corresponding music usage. Making use of the SRBench dataset, we show that employing social folksonomic descriptions in search indices is effective for multimedia soundtrack retrieval.

1 Introduction

Music is not just isolated sound, but also a continuously recurring element in our daily lives. We choose to listen to it as accompaniment to daily activities, such as studying, commuting and working out, or to get into or out of certain moods. We use it as an atmosphere-creating element during significant community events. We may (sometimes involuntarily) be confronted with it in public spaces, and through the multimedia we consume. If certain types of music tend to co-occur with certain types of non-musical context, the non-musical context may become an integral part of the meaning associated to the music.

Traditionally, digital music indexing has had a positivist orientation, focusing on direct descriptions of isolated musical content. However, considering the notions above, it would be interesting to see if music can also *indirectly* be indexed and retrieved based on typical non-musical contextual usage. This would imply that query vocabularies for music would reach beyond limited music-theoretical

* The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-43997-6_18

and mood-oriented vocabularies, or acoustically similar songs in case of a query-by-example scenario.

In previous work [9, 10, 8], we developed the notion of a search engine framework targeting such more subjective, contextual and identity-establishing effects of music. In [9], we considered the use case of a soundtrack retrieval system for user-generated video, and sketched the outline of a demonstrator system in which music search was based on vocabulary-unrestricted free-text query expressions of the intended narrative of the final multimedia result. To connect music to non-musical contexts, inspired by literature from the humanities, we employed folksonomic social collaborative resources on music usage in mass media as our prime data source of knowledge. In parallel, in [10], we investigated what free-text stories were generally associated to production music fragments, and showed that produced descriptions could be associated back to 'stimulus' fragments in a stable way.

The ideas from [9] were not validated yet. In [8], we offered preliminary insights into the capability of folksonomic social collaborative resources to connect musical and non-musical tags associated with soundtrack songs. In the current work, we now will present and validate the general system framework more thoroughly. In this, we will look in a more systematic way at effects of folksonomic tag frequency filtering, as well as various retrieval setups based on different information resource combinations for indexing.

An unusual feature of our approach is that it inherently allows for subjectivity in the query expression and relevance assessment. While this allows for creative media repurposing, this makes evaluation a non-trivial matter. In order to still perform an evaluation which is as unbiased as possible, we employ the SRbench dataset [15], which was meant as a benchmark for soundtrack recommendation systems for image collections.

The remainder of this paper is structured as follows. At first, we will discuss related work considering (aspects of) our problem from various disciplinary angles. Subsequently, we will discuss the general outline of our proposed framework. Then, we will introduce the benchmark dataset against which we will evaluate our approach, followed by a more detailed discussion of the evaluation setup. Evaluation results are then reported and discussed, after which we will finish the paper with a Conclusion and Outlook to future work.

2 Related work

2.1 The humanities perspective: music, meaning and mass media

In this subsection, we will discuss several works from musicology and media studies in which musical meaning is considered in relation to (multi)media soundtrack usage. Our discussion is by no means a comprehensive review (that would be beyond the scope of this paper), but aims to point out representative thoughts.

Cohen [3] indicates that good synchronization between temporal visual development and a music soundtrack will lead to attributes associated in one domain

to be transferred to the other, reporting how associated emotions to a musical fragment get attributed to an animated bouncing ball, as soon as the ball and the music are in sync. Cook [4] also considers transfer of associated meaning between music and multimedia, stressing the importance of mass media expressions. In his book, he relates and contrasts musical structure and development to events in associated media, proposing three basic models for analyzing musical multimedia which will have different interpretation effects: *conformance*, *complementation* and *contest* between music and the associated media.

Considering the function of soundtracks in film, Lissa [11] initially proposed a typology describing various reasons why a soundtrack would be a good match to displayed imagery. These reasons do not all require temporal development. Tagg and Clerida [16] ran an extensive study in which many free-text associations were obtained in response to played mass media title tunes, leading to a revision of Lissa's typology. In the context of our soundtrack retrieval work, in [10] we ran a crowdsourcing study in which production music fragments were played to online audiences, and free-text associated cinematic scenes were acquired. Considering self-reported reasons for cinematic associations, another revision of the original typology from Lissa, Tagg and Clerida was made, identifying 12 categories of reasons for non-musical connections to the played music. For the associated cinematic scene descriptions, consistency was found in the event structure of the provided user narratives, e.g. displaying agreement in whether a scene involves a goal, and whether this goal would be achieved during the scene.

2.2 Contextual notions in music information retrieval

In the digital domain, the field of music information retrieval traditionally has taken a positivist viewpoint on music, with major research outcomes (of which several examples can be found in the review by Casey et al. [2]) focusing on direct description of musical objects in terms of signal features, which then can be used for classification, similarity assessment, retrieval and recommendation. However, the notion of 'context' has been emerging. In Schedl et al. [12], distinction is made between user context and music context, with 'music context' considering contextual information related to music items, artist or performers, such as an artist's background, while 'user context' considers contextual information surrounding consumption by a user, such as the user's mood or activity. Music content is considered as an audio-centric phenomenon, and semantic labels and video clips are considered as parts of music context.

A slightly different terminology and conceptual division are used in the work of Kaminskas and Ricci [5]. Here, the authors distinguish between environment-related context (e.g. location, time, ambience), user-related context (activity, demographical information and emotional state) and multimedia context (text and images). Here, text includes lyrics and reviews; again, this paradigm implies that music fundamentally is considered as an auditory object.

In our current work, we bypass audio representations of music. Instead, we aim to demonstrate that *non-auditory, extra-musical associations to music can offer a valid alternative way to describe, index and retrieve music items.*

2.3 Automated soundtrack suggestion approaches

Various works have been published aiming to automatically suggest soundtracks to videos or slideshows, many of them [6, 18, 7, 13] considering emotion to be the main relevance criterion for matching music to video or slideshow imagery. Emotion-based feature representations are then proposed to be extracted from the audio and visual signals.

In Stupar and Michel [14], cinema is chosen as prime example of how soundtracks are used with imagery, and an approach is proposed in which soundtrack matches are learned in relation to imagery by comparing soundtracks from cinematic films to soundtracks in a music database. Cai et al. [1] propose a system to automatically suggest soundtracks as accompaniment to online consumed web content, again performing matching based on emotional information as inferred from the music and text features.

It is striking how emotion is omnipresent in these automated approaches, while it is not as prevalent in the soundtrack match typology found in our studies [10]. Furthermore, direct signal-to-signal matching employing features obtained from the signal will work on high-quality training data, but encounters quality issues when considered in the context of lower-end content such as user-generated video.

A final disadvantage of direct signal-to-signal matching approaches is that an absolute optimum is implied for the match from music to media item. This contrasts with the notion in multimedia production that the same footage can be used, transformed and repurposed in many radically different (and sometimes contrasting) ways, depending on the accompanying soundtrack.

3 Proposed framework

Our proposed framework is particularly inspired by the earlier described notion in humanities literature that contextual extra-musical meaning associated to music has been established by the way in which music occurs as part of mass media. An interesting online social and collaborative resource holding such associations at scale is the Internet Movie Database (IMDb). In the IMDb, users enter various types of movie information, including plot summaries, keyword summaries, actor listings and reviews. The IMDb also has the possibility to add soundtrack listings to a movie. As a consequence, we can use this resource to associate film plot descriptions—describing the context in which soundtracks occur—to soundtrack listings. While the resource does not allow us to associate a soundtrack to a pinpointed event in the full plot description, our assumption is that *similar music will occur for similar plot descriptions.*

We now can associate film plot information to soundtrack song names. On top of this, we acquire a richer description for the songs by considering another major online social resource focused on music information: `last.fm`. This is a social music platform in which music listening behavior is stored and used

for recommendation. On last.fm, users can describe songs in the form of free-text social tags. The public API of last.fm reveals this information, as well as associated information including popularity data.

The information from these two resources is connected by considering what IMDb movies have plot descriptions with soundtrack listings, and what soundtrack songs have social tag descriptions on last.fm. For our current work, we crawled 22,357 unique IMDb movies with plot descriptions, which had at least one soundtrack song with a last.fm tag associated to it. In total, considering the soundtracks of all these movies, 265,376 song tags could be found.

To model the connected information in a way that is useful for automated soundtrack retrieval, we encode it in the form of various search indices for information retrieval scenarios, as described in the following subsections.

3.1 An information retrieval setup

We wish to automatically suggest music that would fit given non-musical contexts. For this, we first need to know what non-musical context is intended. Then, we need to match this information to knowledge from our crawled resources. We model this as an information retrieval problem, in which the desired context will form the query, expressed in vocabulary-unrestricted free textual form. Because of this allowed freedom, we can accommodate queries at many semantic levels, ranging from descriptions of multimedia items (e.g. a video or image), to more abstract narrative descriptions (e.g. a story or a situation).

Following the information retrieval paradigm, the query is used to retrieve relevant documents from one or more search indices, to be detailed below. Each search index considers a collection of documents. The document content is analyzed and matched against the query to assess relevance, while the document key represents the document in a short-hand way. For the creation of the search indices, which will be considered in various ways in our Evaluation Setup, we make use of the standard features of the Apache Lucene search engine library[1].

A full system setup consists of four steps:

1. Take a contextual free-text query description as input;
2. Match the query against a search index containing movie plot descriptions; return song tags for soundtracks belonging to the plot descriptions with highest relevance to the query;
3. Adapt the song tag collection by expanding it with the most frequently co-occurring song tags to the ones already in the collection;
4. Use the adapted song tag collection to query a music database (in which music items are described by song tags).

3.2 Mapping of song tags to movie plots

A movie m can be represented by a full-text plot description p_m, but also by a collection of soundtracks $S_m = \{s_1, s_2, s_3, ...\}$. Each soundtrack song $s_n \in S_m$

[1] https://lucene.apache.org

can in its turn be represented as a collection of corresponding social song tags: $s_n = \{t_1, t_2, ...\}$.

We re-encode each social song tag by mapping it to a collection of movie plot descriptions: $t \mapsto \{p_m\} \quad \forall p_m : t \in s \in S_m$. The corresponding representation is treated as a document for a search engine, with song tags as document keys and the collection of movie plot descriptions as document content. These are indexed using Lucene's standard indexing options, which are based on the tf-idf weighting model. Because of this, stopwords and frequently occurring words across a corpus will naturally get lower weights than terms that are more representative of a particular document in a corpus.

Upon availability of the search index, it is possible to enter any narrative description as a query, upon which a ranking of matched 'documents' is returned in response to the query. Each of these documents will be represented by the document key. Therefore, upon entering an unrestricted free-text narrative description, a ranked list of song tags will be returned. In [8] we already visualized some interesting patterns emerging from this plot-based index; in the current work, we will consider it as part of our evaluations.

3.3 Mapping of song tags to other song tags

In [9], we informally noted that retrieval results appeared to improve when a pseudo relevance feedback step was performed after the narrative querying step mentioned above. The reasoning behind such a step is as follows: if an initial search (for example on a movie plot search index) will yield relevant song tags, other song tags that co-occurred with the relevant song tags in a song may also be relevant. Therefore, it is good to retrieve the most important song tags that co-occur in songs with the tags we already have.

To acquire relevant co-occurring tags, we take a ranked list of suggested song tags, and use these to query an index in which song tags are mapped to other song tags with which they co-occur in a song. Put more formally, if we have a song $s_n = \{t_1, t_2, ...\}$, then $t_i \mapsto \{t_j\} \quad \forall t_i \in s_n, \forall t_j \in s_n, \text{with } i \neq j$. When indexing this associated song tag representation in a similar way as we did before (taking the collection of co-occurring tags as 'document content', and the tag of interest as 'document key'), frequently occurring tags will only be considered as relevant terms when they do not dominate the full corpus.

3.4 Mapping database target songs to song tags

When we ultimately want to associate narrative queries to songs, we need a database of candidate songs. Each song in the database should be represented in such a way that the steps above lead to a querying representation against which the song can be matched. Therefore, each song in the database should be represented as a collection of song tags. Unusually, our aim is not to restrict the vocabulary here to allow for as-flexible-as-possible matching. In [9], we employed tag vocabulary, song titles and song descriptions in a production music database

Aviation, Architectural, Cloudscape, Conservation, Cosplay, Digiscoping, Fashion, Fine art, Fire, Food, Glamour, Landscape, Miksang, Nature, Old-time, Portrait, Sports, Still-life, Street, Underwater, Vernacular, Panorama, War, Wedding, Wildlife

Table 1: Query image collection categories in SRbench.

as contributed by the songs' composers. In the current work, we do not restrict to production music, but make use of a general commercial song database.

4 The SRbench benchmark dataset

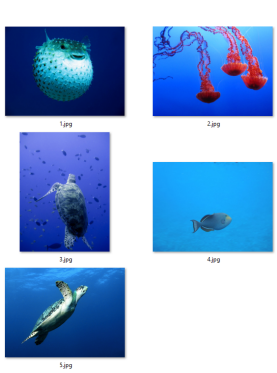
To evaluate the appropriateness of the retrieval mechanisms as outlined in Section 3, we need a dataset of songs, vocabulary-unrestricted queries, and corresponding relevance or recognition assessments. No standardized datasets exist for our problem. Furthermore, since our approach begins with a free-text, contextual expression of the music information need, the system allows for a high degree of flexibility and subjectivity in query vocabulary.

We could design a validation campaign through a user study, although an additional concern may be that free-text querying of music information is not commonly adopted yet in search engines, and therefore might need more detailed instruction or examples to test subjects, risking bias effects. Furthermore, the initial description does not have to be generated by a human per se; for example, we could envision the output of automatic image (or video) captioning systems (e.g. [17]) to also be usable as input to our framework.

As in the current work, we are not interested in the querying mechanism, but rather in the retrieval performance as a consequence of given queries, for our current evaluations, we will employ the existing SRbench dataset [15]. SRbench was established as an evaluation benchmark for soundtrack retrieval in connection to collections of still images. It consists of 25 query photo collections, which were crawled from popular categories in Picasa, as shown in Table 1. Next to this, it offers audio snippets of 470 popular commercial songs in various genres. Through locally and online conducted user assessments, for each of the photo collections, for each pair of songs taken from the dataset, six different people indicated how they rate the fit of each of the songs to the photo collection. The strength of their preference for one song over the other is expressed through a *preference strength* p_s, on an integer scale from 0 (no preference) to 5 (strong preference). A system can then be evaluated by taking (a part of) its output ranking in reponse to a query, and assessing to what extent rankings between pairs in the system-produced ranking match those indicated by the human assessors.

Stupar and Michel [14] measure performance with respect to a top-20 output, according to two measures:

1. Preference precision : $\frac{\#\text{ correctly ordered pairs}}{\#\text{ evaluated pairs}}$.
2. Weighted preference precision: $\frac{\sum p_s \text{of correctly ordered pairs}}{\sum p_s \text{of evaluated pairs}}$.



(a) images in SRbench

(b) Flickr API response

Fig. 1: Images of the 'underwater' category in SRbench, alongside Flickr `tags.getRelated` API response output for the tag 'underwater'.

The measures are calculated considering ratings at three agreement levels: those ratings with 6/6, 5/6 and 4/6 assessor agreement, respectively, on which song in the assessment pair was a better fit to the given image collection. In our current work, we also make use of these measures, as explained in the following section.

5 Evaluation setup

In order to evaluate our approach against SRbench, a few steps should be performed in the evaluation setup. First of all, we should *translate the target song database to a song tag representation*. To this end, we cross-match artist and title names with `last.fm`. We managed to match 439 out of 470 songs. For classical music items, typically one particular rendition (not necessarily the one matched to our metadata) would have the most extensive tag vocabulary; we manually corrected the mapping such that this vocabulary would be used.

Another necessary step is to *translate the 25 image collection queries into textual queries to our search indices*. As can be noted from Table 1, the image categories are very general and abstract. We wanted to obtain a richer description without putting in personal bias. While one could run visual concept detectors for this, we retained the spirit of using social online knowledge resources, employing the API of the Flickr image service. Through the `tags.getRelated` API call, using SRBench category names as query tag input, a larger set of related tags could be obtained as commonly co-occuring in Flickr. As an illustration of the adequateness of this technique, in Figure 1 we show the SRbench images for the category 'underwater' alongside the Flickr API response for `tags.getRelated('underwater')`.

As not all songs could be matched to `last.fm`, our set of 439 songs comprises 5847 song pairs, as opposed to the 6836 song pairs in SRbench for the original

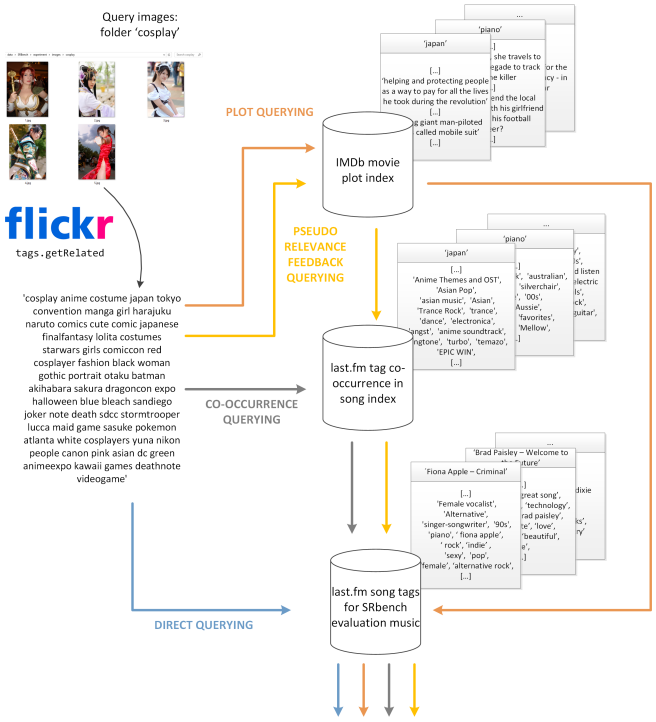
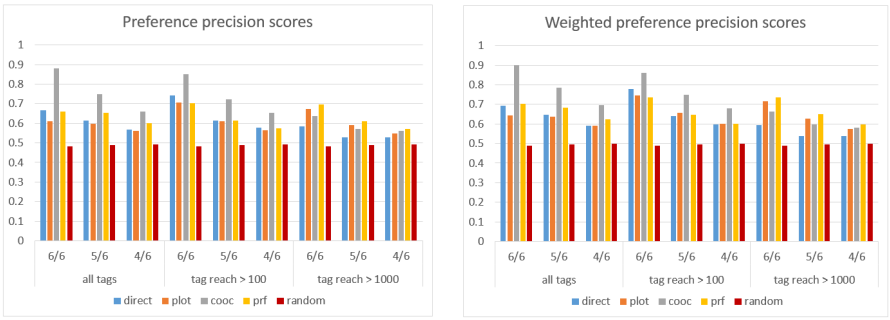


Fig. 2: Illustration of the various experiments performed for evaluation of our approach.

full set of 470 songs. Since SRbench does not include raw ranking outputs for earlier proposed systems, we cannot recompute performance for the approaches in [15] based on a smaller set of song pairs, so direct comparison of evaluation outcomes against those originally reported in [15] is impossible. To still get a good sense of performance, we will consider various configurations within our proposed framework, in the form of four possible setups which also are illustrated in Figure 2:

1. Direct querying of the target song database index (*direct*). This approach is expected to give a baseline of performance of our system, as we do not expect many non-musical terms from the image query collection to coincide with social music tags.
2. Querying of an index associating song tags to movie plots (*plot*). Returned song tags are used as query to the target song database.
3. Querying of a song tag co-occurrence index (*cooc*). Returned song tags are used as query to the target song database.
4. Pseudo relevance feedback querying (*prf*): first query the index associating song tags to movie plots, use the results to query the song tag co-occurrence index, and use those results to query the target song database.



(a) Preference precision (b) Weighted preference precision

Fig. 3: Evaluation results for our various experiments, considering different minimum tag reach frequencies and minimum levels of agreement.

As a final, fully independent baseline, we also consider a *random* approach. For this, we generate 1000 random top-20 outcomes. For each of these outcomes, we calculate the evaluation metrics (unweighted and weighted performance, considering 6/6, 5/6 and 4/6 annotator agreement), averaging over all trials.

In reports on tag metadata (obtained through the `tag.getInfo` API call), `last.fm` distinguishes between 'reach' (the number of individual consumers of a given tag) and 'taggings' (the number of times a tag is used). To test the effect of 'cleanness' of social tags, we test three setups varying the reach frequency of the tags (avoiding frequency bias caused by very actively tagging individuals): (1) all social tags are indexed; (2) only social tags with a reach of at least 100 are indexed; (3) only social tags with a reach of at least 1000 are indexed.

6 Results

The obtained evaluation results for the various experimental setups are plotted in Figure 3. First of all, there is clear distinction between the outcomes obtained for our various system setups in comparison to the random baseline. We generally notice that upon tolerating lower agreement levels (e.g. 4/6 instead of 6/6), performance drops. This is logical, as that would make the query more ambiguous to interpretation. Surprisingly, performance generally is highest when *not* applying reach frequency filtering to tags. Next to this, the baseline of directly querying the song database (*direct*) performs reasonably well, even outperforming movie plot index querying (*plot*) when considering all social tags for indexing. Under this setup, we also see notably high performance on the tag co-occurrence index (*cooc*). However, when enforcing higher minimal reach frequencies, performance on the *direct* and *cooc* approaches drops considerably, while the *plot* and *prf* querying approaches will improve as the song tag corpus gets 'cleaner'.

A conclusion to be drawn from these results is that rarely consumed song tags, despite being part of a large and noisy vocabulary, still may hold some

useful narrative and contextual information to directly match against. Yet as soon as tags get 'cleaner' and more frequent, these tags will disappear from the index. In that case, direct matching against song tags becomes more difficult, and a narrative matching step (*plot* and particularly *prf*) will make more sense. Applying the pseudo relevance feedback stage consistently outperforms querying of the movie plot only, and as such seems to hold the best of both worlds: focusing a song tag corpus based on co-occurrence, but also taking advantage of any relevant narrative contextual associations.

7 Conclusions and future opportunities

In order to associate music to non-musical narrative elements, we proposed to make use of socially established associations from online mass media. In our proposed approach, we start from an unrestricted free-text narrative query expressing the desired context. We modeled folksonomically expressed associations as obtained from the IMDb and `last.fm` as search indices in an information retrieval scenario and considered various search index setups.

Evaluating our approach against the SRbench benchmark dataset shows that social song tags turn out fairly adequate to match non-musical queries against—especially when 'filtered' through a tag co-occurrence index. Future work should examine if co-occurrence has particularly strong effects due to the songs having been used as movie soundtracks, or if this would hold for any corpus with social tagging. Employing matches against full-text movie plots especially performs well when combining this with co-occurrence filtering. Performance effects caused by restricting to higher minimum reach frequencies suggest that tags with low reach may actually hold relevant contextual information. At the same time, this is at the expense of having to deal with a larger and noisier index.

We believe that our work holds promise in several ways. From a disciplinary viewpoint, we made an important step in unifying data-driven approaches with more subjective and qualititative notions on music utility. While current online music service models tend to focus on on-demand consumption of known songs, employing an approach like ours can assist in transforming this into on-demand immersion in known experiences. We therefore hope our work can ultimately offer novice audiences new and serendipitous entrances to lesser-known genres.

Acknowledgements: The research leading to these results has received funding from the European Union Seventh Framework Programme FP7 / 2007–2013 through the PHENICX project under Grant Agreement no. 601166.

References

1. R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma. MusicSense: Contextual Music Recommendation using Emotional Allocation Modeling. In *Proc. 15th ACM Int. Conf. on Multimedia (ACM MM)*, pages 553–556, Augsburg, Germany, 2007.

2. M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, April 2008.
3. A. J. Cohen. How music influences the interpretation of film and video: Approaches from experimental psychology. In R. Kendall and R. W. Savage, editors, *Selected Reports in Ethnomusicology: Perspectives in Systematic Musicology*, volume 12, pages 15–36. Department of Ethnomusicology, University of California, Los Angeles, 2005.
4. N. Cook. *Analysing musical multimedia*. Oxford Univ. Press, New York, USA, 1998.
5. M. Kaminskas and F. Ricci. Contextual music information retrieval: State of the art and challenges. *Computer Science Review*, 2012.
6. F.-F. Kuo, M.-F. Chiang, M.-K. Shan, and S.-Y. Lee. Emotion-based music recommendation by association discovery from film music. In *Proc. 13th ACM Int. Conf. on Multimedia (ACM MM)*, pages 507–510, 2005.
7. C.-T. Li and M.-K. Shan. Emotion-based impressionism slideshow with automatic music accompaniment. In *Proc. 15th ACM Int. Conf. on Multimedia (ACM MM)*, pages 839–842, Augsburg, Germany, 2007.
8. C. C. S. Liem. Mass Media Musical Meaning: Opportunities from the Collaborative Web. In *Proc. 11th Int. Symp. on Computer Music Multidisciplinary Research (CMMR)*, Plymouth, UK, June 2015.
9. C. C. S. Liem, A. Bazzica, and A. Hanjalic. MuseSync: Standing on the Shoulders of Hollywood. In *ACM Multimedia Grand Challenge finalist entry*, November 2012.
10. C. C. S. Liem, M. A. Larson, and A. Hanjalic. When Music Makes a Scene — Characterizing Music in Multimedia Contexts via User Scene Descriptions. *Int. J. of Multimedia Information Retrieval*, 2:15–30, 2013.
11. Z. Lissa. *Ästhetik der Filmmusik*. Henschelverlag, Berlin, Germany, 1965.
12. M. Schedl, E. Gómez, and J. Urbano. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2–3):127–261, 2014.
13. R. R. Shah, Y. Yu, and R. Zimmermann. ADVISOR: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *Proc. 22nd ACM Int. Conf. on Multimedia (ACM MM)*, pages 607–616, Orlando, Florida, USA, 2014.
14. A. Stupar and S. Michel. PICASSO — To Sing you must Close Your Eyes and Draw. In *Proc. 34th Annual ACM SIGIR Conf.*, Beijing, China, July 2011.
15. A. Stupar and S. Michel. Srbench–a benchmark for soundtrack recommendation systems. In *Proc. 22nd ACM Int. Conf. on Information & Knowledge Management (CIKM)*, San Francisco, USA, October 2013.
16. P. Tagg and B. Clarida. *Ten Little Title Tunes — Towards a Musicology of the Mass Media*. The Mass Media Scholar's Press, New York, USA and Montreal, Canada, 2003.
17. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *Proc. 28th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA, June 2015.
18. J.-C. Wang, Y.-H. Yang, I. Jhuo, Y.-Y. Lin, H.-M. Wang, et al. The acousticvisual emotion guassians model for automatic generation of music video. In *Proc. 20th ACM Int. Conf. on Multimedia (ACM MM)*, pages 1379–1380, Nara, Japan, 2012. ACM.