Int J Multimed Info Retr (2013) 2:15–30
DOI 10.1007/s13735-012-0031-3

REGULAR PAPER

When music makes a scene

Characterizing music in multimedia contexts via user scene descriptions

Cynthia C. S. Liem · Martha Larson · Alan Hanjalic

Received: 9 July 2012 / Revised: 1 October 2012 / Accepted: 8 November 2012 / Published online: 24 January 2013

Abstract Music frequently occurs as an important reinforcing and meaning-creating element in multimodal human experiences. This way, cross-modal connotative associations are established, which are actively exploited in professional multimedia productions. A lay user who wants to use music in a similar way may have a result in mind, but may lack the right musical vocabulary to express the corresponding information need. However, if the connotative associations between music and visual narrative are strong enough, characterizations of music in terms of a narrative multimedia context can be envisioned. In this article, we present the outcomes of a user study considering this problem. Through a survey for which respondents were recruited via crowdsourcing methods, we solicited descriptions of cinematic situations for which fragments of royalty-free production music would be suitable soundtracks. As we will show, these descriptions can reliably be recognized by other respondents as belonging to the music fragments that triggered them. We do not fix any description vocabulary beforehand, but rather give respondents a lot of freedom to express their associations. From these free descriptions, common narrative elements emerge that can be generalized in terms of event structure. The insights gained this way can be used to inform new conceptual foundations for supervised methods, and to provide new perspectives on meaningful and multimedia context-aware querying, retrieval and analysis.

Keywords Music information retrieval · Cross-modal connotation · Narrative structure · User studies · Crowdsourcing · Annotation

C. C. S. Liem (✉) · M. Larson · A. Hanjalic
Delft University of Technology, Delft, The Netherlands
e-mail: c.c.s.liem@tudelft.nl

1 Introduction

Music is used in different ways. Besides serving active listening purposes, it can function as background entertainment alongside everyday human activities. Next to this, it also frequently occurs as an important reinforcing and meaning-creating element in multimodal human experiences. Many social occasions, ranging from parties to more ceremonial events such as weddings and funerals, would not be the same without music. In many digital audiovisual productions, music plays an essential role as well, being exploited for setting the atmosphere of a visual scene, characterizing key characters in a scene, implying situations that are not directly visible to the audience, and much more.

In the present-day Web era, creating and sharing such productions is no longer the exclusive privilege of professional creators, as can be seen from the continuously rising usage numbers of video sharing sites such as YouTube[1]. Everyday users are increasingly interested in capturing personally significant situations they encounter on the go, and in sharing these with a world-wide audience. Especially if the captured video material would be of low quality, a fitting music soundtrack can greatly impact and clarify the video's significance (see Fig. 1 for an example).

1.1 Connotative associations in multimedia contexts

In both the cases of enriching real-life social occasions, as well as influencing the message of a digital audiovisual production, music will not be considered in the form of isolated songs. Instead, it is placed in a context that is not necessarily musical itself. The context may be time-dependent, and

[1] http://www.youtube.com/t/press_statistics, accessed December 13, 2012.

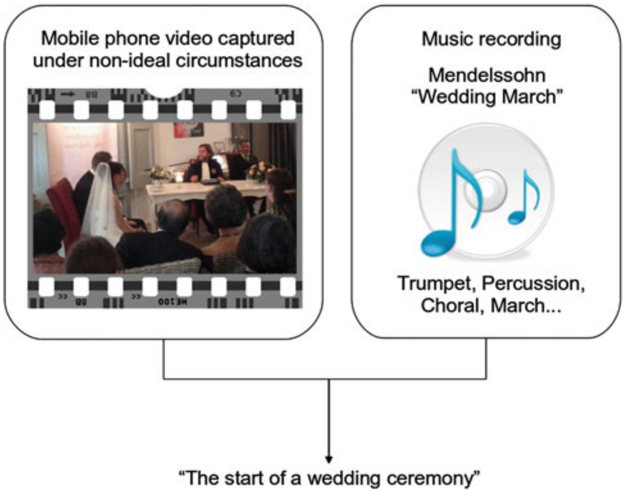

Mobile phone video captured under non-ideal circumstances
Music recording
Mendelssohn "Wedding March"
Trumpet, Percussion, Choral, March...
"The start of a wedding ceremony"

Fig. 1 Certain social events have grown to imply certain types of music and vice versa. This can be exploited to clarify the message of multimedia productions

is composed of *multimodal* information: information which is relevant to different parts of the human sensory system. When music is placed in this context, a situation arises in which a combination of information in different modalities and formats provides added value in comparison to the original context. In other words, music will now be part of a *multimedia context*.

The added value that music provides to a context can work in a *cross-modal* way, in which information relevant to one part of the human system, such as the auditory channel, influences the perception of information relevant to another part of this system, such as the visual channel. In terms of meaning and significance, under these circumstances, music can have influence outside its own domain. As soon as this becomes conventional, a non-musical context may directly get associated with specific types or properties of music, and vice versa. For example, a bugle call typically has grown to evoke militaristic images, while a lullaby rather evokes images of infants. This way, outside of *denotative* meaning dealing with purely musical properties, music also gets a referential *connotative* meaning layer [29].

Indeed, in multimedia productions, cross-modal connotative effects and associations are heavily exploited. Since the early years of the moving picture, multimedia professionals have been trained to analyze the techniques to achieve this [17,26], and to actively exploit these themselves when creating their own productions [20,34].

Due to exposure in daily life and mass media, non-professional lay users may be capable of recognizing prototypical associations between music and contextual information in other modalities as well. While it is unrealistic to assume that these users are capable of describing their musical information needs for a desired non-musical context in terms of musical or signal characteristics, it is likely that they can indicate high-level characteristics of the envisioned multimedia result. In fact, since low-level signal properties cannot cover all aspects of meaningful associations, and since the optimal musical match will differ per case, it would even be more realistic to describe an information need in terms of the envisioned multimedia result, rather than in terms of purely musical characteristics. This point is illustrated in Fig. 2: based on low-level timbral feature properties, it would be hard to distinguish between the three displayed songs (e.g., they all contain trumpets and percussion). However, the associations triggered by the songs are very different. The narrative context of the user query influences the relevance of these associations. For example, 'Duel of the Fates' would be the most appropriate match if the user envisions an epic battle scene. On the other hand, if the user would wish to show the rivalry between his pets in a funny or ironic way, the Wedding March could actually become a more suitable match.

1.2 Contributions and outline

In this article, we take a novel approach to multimedia context-oriented annotation. Focusing on production use cases, we investigate the feasibility of characterizing music via high-level descriptions of *its context and intended use* (envisioned multimedia results that fit to it) rather than in terms of *modality-specific content* (musical feature descriptors). We do *not* assume the annotation vocabulary to be known beforehand. Instead, we solicit free-form text responses for our survey, giving full control over the annotation vocabulary and its typical usage to the survey respondents, and inferring the characteristics of the vocabulary and its usage only after the responses are received.

The results of our work push the music and multimedia information retrieval fields forward in two ways:

- As we show, contextual multimedia scene descriptions can reliably be matched to the music fragments that triggered them. This opens entirely new perspectives on cross-modal connotative annotation and querying strategies.
- The conceptual levels at which descriptions are generalizable, as well as the self-reported reasons for giving the descriptions, provide insights into criteria that caused cross-modal connotative associations to emerge. These can inform the design and realization of supervised automated methods and ground truth at the multimedia signal data level, such that practical systems can be realized that take these associations into account.

This article is structured as follows. We start with giving an overview of related work in Sect. 2. We then present our

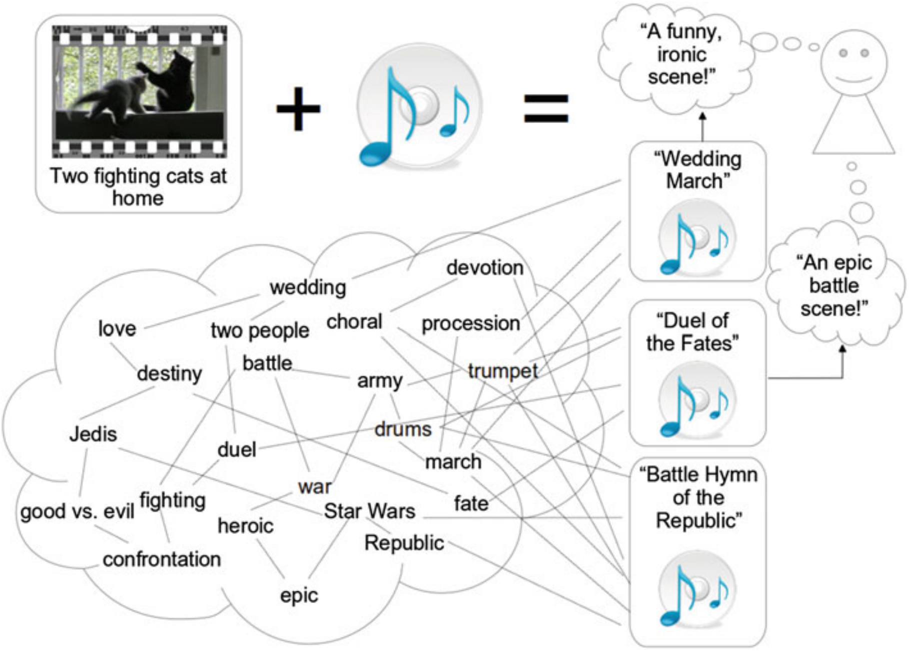


Fig. 2 An example of potential connotative meaning-formation in the intended multimedia production use case. We have a user video and three potential soundtrack candidates, for which several possible meaning associations are indicated with *dashed lines*. Due to these associations, the choice of soundtrack will influence a user's interpretation of the resulting multimedia scene

experimental setup in Sect. 3. We proceed by describing statistics related to the crowdsourcing mechanisms in Sect. 4. Section 5 analyzes the results of a task testing the recognizability of correspondences between multimedia descriptions and the music fragments that triggered them. Section 6 focuses on the analysis of the obtained free-form text survey responses, investigating which narrative elements emerge and at which conceptual levels descriptions are generalizable. Finally, Sect. 7 provides a conclusion and outlook to follow-up opportunities to our findings in the music and multimedia information retrieval fields.

2 Related work

Our work in this article is an exploratory user study. In the music information retrieval (Music-IR) domain, the importance of user-centered versus system-centered approaches has attracted attention for several years [9,24], although actual adoption of user-centered approaches still is rare. Recent overviews on the types of user studies that have been conducted in the field and their impact can be found in Weigl and Guastavino [44] and Lee and Cunningham [23]. Most of the few existing Music-IR user studies have been focused on information needs, search strategies, and query analysis topics (e.g., [8,10,22]). In our case, while our work ultimately relates to search and retrieval, we will first and foremost consider a scenario of data description.

In terms of the domain under study, the work of Inskip et al. [14,15], dealing with discourse in search and usage of pre-existing commercial music, is close. However, this work has been focusing on professional stakeholders only and was more oriented towards identifying typical language usage in decision-making and the production pipeline.

Looking at previous data description work, well-known human-tagged Music-IR datasets such as CAL500 [40] and MagnaTagATune [21] strongly focused at having the tag annotations satisfying inter-annotator agreement criteria, but did not explicitly analyze user motivations behind the resulting annotations. Only very recently, work by Lee et al. [24] appeared in which the rationale behind assigned musical mood labels was investigated more deeply. In all these cases, music is treated as an isolated information source, not exceeding the earlier-mentioned idea of pleasant background diversion accompanying everyday human activities.

The notion of music in *context* has been emergent in recent years. Schedl and Knees [35] mention cultural context information, but this context information (such as web pages on performing artists) once again remains largely oriented towards the idea of music in an isolated or accompaniment function. Kaminskas and Ricci [18] give an overview of work in different categories of music context, including multimedia. These approaches are strongly data-driven, while deeper insights into the underlying mechanisms establishing context would be desirable.

Recent work considering cross-modal relations between music and contextual, external information includes Refs. [4, 11,19,25,38]. These either fix the cross-modal association domain to be purely affective, or directly match signal features in different modalities to each other. In our work, we will adopt the view from Refs. [19,38] on cinema as prototypical multimedia form. However, we want to look beyond pure signals and fixed domains, at *semiotics* rather than *semantics*, and *connotation* rather than *denotation*. Ideas on semiotic hypermedia and multimedia have been pioneered in [6,30]. A recent rare work on connotation was done by Benini et al. [3], in which annotations in a connotative space was shown to produce higher user agreement than emotional label tagging.

Semiotic aspects have been a topic of interest in the field of musicology for several decades, one of the first pioneers being Nattiez [31]. While Nattiez argues that music lacks the communicative power to picture concrete events, it still is capable of suggesting general outlines of these events. In an experiment very similar to the way we will set up our description survey, Nattiez played a piece of classical symphonic program music to secondary school students, asking them to describe the story the piece was telling. While the original program of the piece did not appear, certain categories of stories clearly occurred more frequently than others. In addition, the study showed that not all elements of formal musical structure were being connected to stories, and that less relevant events in relation to the musical argumentation (e.g., a final chord) are nonetheless being picked up as relevant to the story descriptions.

An interesting aspect of present-day musicological thought is that it has shifted away from absolutist, positivist views (assuming an objective truth 'in the music data') towards subjective, contextual, and culturally specific aspects of music [7], up to the point that 'the music itself' is nowadays even considered a taboo concept [45]. Because of this view, musicology has strongly been studying contextual topics in which music plays a more significant role than mere accompaniment. Film music and mass media studies also have been a long-time interest in the field. A very useful categorization of film music functions has been made by Lissa [26], mentioning that music in film can be used to indicate movement, stylize real sounds, represent space, represent time, communicate meaning through deformation of sound, provide a commentary external to the narrative, serve as music within the narrative, indicate psychology of actors, provide a source of empathy, function as a symbol, anticipate action, and serve as a formal unifying factor.

Influenced by Lissa, Tagg [39] proposed a sign typology for music involving sonic, kinetic, and tactile *anaphones* (sounds that were metaphorical for sonic, kinetic, and tactile sensory qualities), *genre synecdoche*, in which a small glimpse of an exotic genre will evoke the full exotic context, *episodic markers* announcing musical episode transitions, and *style indicators*. In [39], Tagg and Clarida also report findings from a decade-long study, in which association responses to ten different well-known theme and title songs were collected and categorized. Major differences with our approach are that immediate, very short associations were sought (respondents only had a few seconds to write these down) while we solicit more elaborate user input.

Following our vision of free-form text descriptions of multimedia productions, the results will be text descriptions of visual narratives, needing to be analyzed in terms of *narrative structure* [1,2,5]. In our analyses, we will adopt a simple, pre-theoretical, and widely-accepted dictionary definition of a narrative as "*a spoken or written account of connected events; a story*" [36]. This definition motivates us to take events as basic building blocks of a narrative, and turn to event theory for a basic typology of events as they are most naturally conceptualized in human semantic systems. More specifically, we will build upon the work of Vendler [42] regarding the structure of events in terms of classes of verbs. Four classes are differentiated on the basis of temporal structure: *states* have no internal structure, *activities* involve internal change but no endpoint, *achievements* involve an instantaneous culmination or endpoint and *accomplishments* involve a build up period, and then a culmination. Collectively these classes are referred to as event classes. A basic distinction can be made between culminating events (achievements and accomplishments) and non-culminating events (states and processes).

Our chosen methodology for soliciting user responses is that of *crowdsourcing*, an approach which is currently emerging in the Music-IR field, a.o. in [24,27,41]. In multimedia information retrieval, previous crowdsourcing work typically focuses on collecting annotations describing the explicitly depicted content of multimedia. For example, Nowak and Rüger [32] employ crowdsourcing to collect information about images and focus on labels for visual concepts. Such a task targets collecting 'objective' annotations, which can be considered correct or incorrect, independent of the personal view of the annotator. In contrast, in this work we use crowdsourcing to carry out a highly open-ended and creative task for which no independent gold standard can be considered to exist. In this way, we take a step beyond the work of Soleymani and Larson [37], in which larger tasks than microtasks were already experimented with and reported on, and Vliegendhart et al. [43], in which crowdsourcing tasks requiring a high imaginative load were supported. Because we are interested in eliciting associations made by humans, our use of crowdsourcing also shares commonalities with its use for behavioral studies, as discussed by Mason and Suri [28].

3 Experimental setup

3.1 Infrastructure: Amazon Mechanical Turk

The user studies presented in this article were conducted using the infrastructure of the Amazon Mechanical Turk[2] (AMT) crowdsourcing platform. AMT is an online work marketplace on which *requesters* can post *Human Intelligence Tasks* (HITs), which can be taken up by *workers*. Each HIT

[2] http://www.mturk.com, accessed December 13, 2012.

Part 1: General questions on the music fragment

1. How familiar are you with this music fragment?
   [*very unfamiliar - somewhat unfamiliar - somewhat familiar - very familiar*]
2. Please characterize the music fragment.
   (a) How positive or negative do you consider the message of this fragment to be?
   [*very negative - somewhat negative - neutral - somewhat positive - very positive*]
   (b) How active or passive do you consider this fragment to be?
   [*very passive - somewhat passive - neutral - somewhat active - very active*]

Part 2: Describe a cinematic scene to which this music fragment can be a soundtrack
Imagine you would be a film director, preparing your greatest piece of cinema work so far. You have a huge budget for this, meaning you could hire a world-class crew to support you, and you don't have any limitations regarding filming locations, costumes, props etc. either. The award-winning soundtrack composer perfectly understands what you want to express with your movie - and came up with this fragment as a musical sample to go with a certain scene in your movie. In the following questions we would like to learn from you what the characteristics of this scene would have been. Please use your imaginative power and be as creative and illustrative as you can.

3. What is happening in the scene?
   [*free-form text input*]
4. If you had any specific film genre in mind of which this scene would be a part, please mention it here.
   [*free-form text input*]
5. Who are the key characters in the scene?
   (a) If you indicated any key characters in your scene description above, please introduce them here (e.g., 'an old woman', 'James, a 40-year old father', 'an alien from Mars', 'a pit bull terrier', 'Napoleon Bonaparte'...).
   [*free-form text input*]
   (b) What do these characters look like? What kind of persons/creatures are they? Are they thinking or feeling anything specific? If there are multiple characters, do they relate to and interact with each other? If so, in what way? (e.g., 'James spends most of his time at the office. He doesn't like that and wishes to spend more time with his kids.', 'The pit bull terrier is evil and very aggressive. He does not get along with the docile cat next door.' etc...)
   [*free-form text input*]
6. What is the general setting of the scene?
   (a) Where does the scene take place? (e.g., 'China', 'in the mountains', 'in a Tuscan Medieval convent')
   [*free-form text input*]
   (b) When does the scene take place? (e.g., 'in 1815', 'in Summer', 'during the Second World War', 'at midnight')
   [*free-form text input*]
   (c) Please give some indications of what the setting visually looks like (e.g., 'The sun is shining brightly over the convent. It is a recently restored building, that attracts a lot of tourists during the Summer season.')
   [*free-form text input*]
7. Please describe why you chose the scene, characters and setting described in the previous questions for the music fragment.
   [*free-form text input*]

Fig. 3 Questions used for the cinematic scene description survey

typically is a microtask, soliciting a small amount of human effort in order to be solved, and yielding a small amount of payment to a worker who successfully completes it. Our choice for this platform is a.o. supported by findings in previous work by Paolacci et al. [33], in which experiments in the area of human judgment and decision-making were performed both on AMT and using traditional pools of subjects, and no difference in the magnitude of the effects was observed.

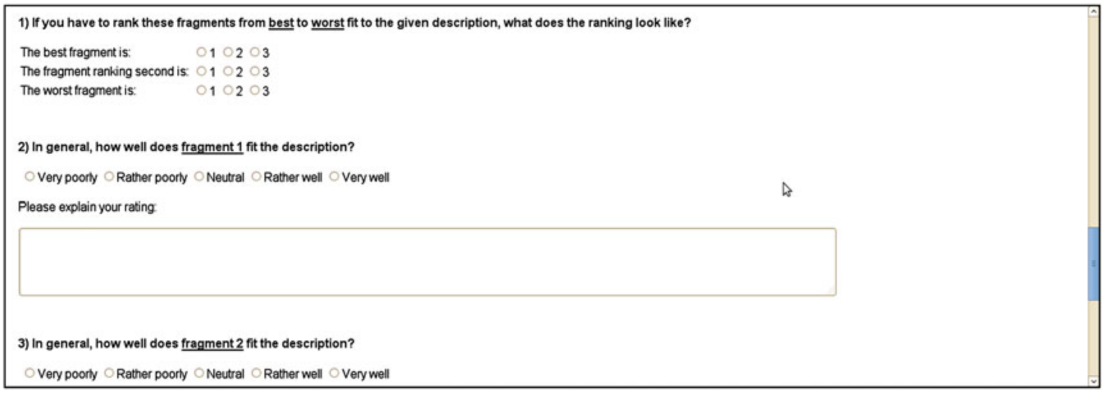
A requester designs a HIT template and separately specifies the input data to be used with this. Furthermore, for one HIT, a requester can release multiple *assignments*, requiring that multiple different workers carry out the same task on the same input data. HITs are typically released in *batches*, grouping the assignments of multiple HITs with the same template. This way, upon completing a HIT, a worker can easily proceed to taking up another HIT of the same type, but with other input data. All completed work has to be reviewed by the requester and approved.

If a worker did not complete a task in a satisfactory way, the requester may reject the work and repost the task. A requester can limit the access to HITs by imposing requirements on the worker's status. Besides requirements based on general worker properties, such as a worker's approval rate on earlier tasks of the requester, a requester can also manage the access to HITs by soliciting and handing out custom qualifications to workers.

In the following two subsections, we will present the core of a description survey and a subsequent music ranking task on which the work in this article is based. After this, we proceed by describing the data that were used for our experiments, and specifying our task runs in more detail.

3.2 Cinematic scene description survey

As the basis for the work reported in this article, a HIT template was designed for a survey. With the survey, we wanted to identify suitability criteria for connections between music and narrative videos, described in spontaneous, rich, and intuitive ways. Therefore, we asked survey respondents to describe an *idealized cinematic situation, for which a given music fragment would fit as a soundtrack*. The choice for the cinema genre was made, since cinematic scenes are considered to be a prototypical case in which music soundtracks and visual scenes are connected and jointly create new connotative meaning. Besides, cinematic scenes are works of fiction, and thus allow for unlimited creative freedom in their conception. In order to help respondents in elaborating on their ideas, and facilitate later analysis, the solicited description was broken down into separate free-form text questions related to the scene, its characters, and its setting. The full breakdown of all questions of the survey, together with the examples and instructions that accompanied them, are shown in Fig. 3.



1) If you have to rank these fragments from best to worst fit to the given description, what does the ranking look like?

The best fragment is:	○1 ○2 ○3
The fragment ranking second is:	○1 ○2 ○3
The worst fragment is:	○1 ○2 ○3

2) In general, how well does fragment 1 fit the description?

○ Very poorly ○ Rather poorly ○ Neutral ○ Rather well ○ Very well

Please explain your rating:

3) In general, how well does fragment 2 fit the description?

○ Very poorly ○ Rather poorly ○ Neutral ○ Rather well ○ Very well

Fig. 4 Excerpt from the music ranking HIT template

While our description survey has similarities to the earlier cited work of Nattiez [31], the types of solicited descriptions differ fundamentally. In the case of Nattiez, the solicited description can be considered a direct visualization or explanation of the music. In our case, we assume the opposite case: the music explains the described cinematic situation. In terms of a causal relationship, the description is the cause and the music the effect, which is common practice in the musical scoring of cinematic productions [34].

3.3 Music ranking task

Ultimately, we want to place the work of this article in a multimedia Music-IR context. Because of this, it is important to not just solicit descriptions for music fragments, but also verify if music fragments can be realistically recognized and retrieved based on the same type of descriptions. To this end, next to the previously described survey, we designed a HIT in which a cinematic scene description that had been previously supplied would be given to a worker. The worker would then be asked to rank and rate three given music fragments according to their fit to the given description. A screenshot of part of the corresponding HIT template is given in Fig. 4.

The description was presented in the original breakdown form as presented in the previous subsection, e.g., separating key character and location descriptions from overall scene descriptions. We only omitted the field in which the original describer explained why he chose the particular scene that was described.

After the given cinematic description, the worker was provided with three music fragments. These fragments were a randomly ordered triplet, out of which one song was the music fragment that had originally triggered the description (from now on indicated as the 'stimulus fragment'). The other two fragments were randomly chosen without replacement from a large royalty-free production music database, which will be described in more detail in Sect. 3.4. The task of the worker was to rank the three fragments according to their fit to the given description. This was done both through explicit indication of the rank positions, as well as through ratings of the perceived fit of each fragment on a 5-point scale. Finally, the worker had to give an explanation for the given ratings. The input has some redundancy: from the 5-point ratings, a ranking can be devised already. However, this redundancy was conscious and served to ensure answer robustness.

3.4 Data

A music dataset was created with 1,086 songs from three royalty-free production music websites[3]. Royalty-free music has the advantage of being legally playable on the Web and reusable; production music has the advantage of having explicitly been written to be used in multimedia contexts.

To prevent listener fatigue, we only consider dataset items with durations between 30 s and 5 min for our task runs. We chose not to further standardize the fragments lengths, since this would require alterations modifying the original musical discourse. The exact number of songs used differed per task, and will be explained in the following subsection.

3.5 Task run specifications

With the tasks described in Sects. 3.2 and 3.3 as basis, we performed several batches of experimental runs: (1) an initial description survey run, (2) a description HIT run yielding qualifications for follow-up HITS, (3) a follow-up description HIT run, and (4) a follow-up music ranking HIT run.

3.5.1 Initial survey run

The first released version of our description survey was more extensive than our outline of Sect. 3.2: apart from a description of a cinematic situation, respondents also had to

[3] http://www.incompetech.com (Kevin MacLeod), http://www.danosongs.com (Dan-O) and http://derekaudette.ottawaarts.com/index2.php (Derek R. Audette), accessed December 13, 2012.

describe a real-life situation they once experienced to which the music fragment would fit. Furthermore, basic demographic information was asked [gender, age, country, level of (music) education and frequencies of listening to music, watching web videos, and watching movies]. Finally, since the solicited real-life situation description could be privacy-sensitive, respondents had to explicitly permit the reuse and quoting of their responses.

As the survey was expected to require a much longer completion time (over 15 min) than a regular AMT HIT, we, at first, were concerned about releasing it as such, and instead put up a HIT batch in the AMT Sandbox environment. This would allow us to benefit from the AMT infrastructure, but to handle the recruitment of respondents ourselves. A call for survey participation was disseminated intensively: through a snowball procedure, professional and personal contacts were written and asked to propagate the message by forwarding the message to their own (inter)national professional and personal contacts. Next to this, paper flyers and announcements were distributed on-campus in The Netherlands, and announcements were placed on six different web forum communities in which role-playing was strongly featured. Through this, a large and demographically diverse audience was quickly reached.

While our dissemination strategy reached hundreds of people, the actual uptake was not as large as we expected. Multiple contacts forwarded our call, but did not feel obliged or eligible to complete a survey themselves. In addition, unfamiliarity with the AMT platform made several potential respondents reluctant to set up a login account. Upon receiving this feedback, we offered interested people the possibility of using a lab-created account which would be fully independent of their personal information.

For the task, we put up surveys for 200 music fragments meeting the specifications given in Sect. 3.4, with three assignments per fragment. After 20 days, 65 completed surveys had been received, from respondents from 13 different countries, on 57 unique music fragments[4]. We had reuse permission of input for 55 of these fragments.

Judging from the responses, respondents found it very hard to envision and describe real-life situation descriptions, but rarely had problems to do this for cinematic scenes. For this reason, in follow-up runs of the survey we did not focus on the real-life situation anymore.

3.5.2 Qualification run

Following the experiences above, we moved to the regular AMT platform after all. At first, we piloted a HIT following the specifications of Sect. 3.2; in addition, the same demographic information as described in Sect. 3.5.1 was asked. Any worker with a HIT Approval rate of at least 0.9 could perform the task. The reward for successfully completed HITs was $0.09.

We first piloted a HIT task for one song (‘Exciting Trailer’ by Kevin MacLeod, which was not used in the initial survey), intended to run continuously for several days to harvest as many responses as possible. Therefore, 150 assignments were released for the HIT. Within a few days, it became clear the uptake was much larger than in the case of our initial survey run (more information on this will follow in Sect. 4). Several workers indicated that they had not seen a task like ours before on AMT, but despite the longer required working time and the creative load of the task they indicated to enjoy doing the task. Following this enthusiastic response, we released similar HITs for two more songs that were not used in our initial survey: ‘Mer Bleue Boogie’ by Derek R. Audette and ‘Origo Vitae’ by Dan-O. Once again, these had 150 assignments each, and were intended as opportunities for harvesting many responses for the same song.

Any completed HIT for which the free-form text responses were at a sufficient level of English, and for which there were indications that the music fragment had a relation to the given scene description (as opposed to answers in which workers just described their favorite actors or movie scenes), would grant the corresponding worker a qualification for a follow-up HIT batch: either another description batch as detailed in Sect. 3.5.3, or a batch of the ranking task detailed in Sect. 3.5.4.

In total, qualifications were granted to 158 workers. Upon earning a qualification, a worker was e-mailed with a notification. Workers were distributed over batches such that the demographics distributions in all batches would be similar. Qualifications were granted such that a worker would not be able to work on more than one batch at the same time. If the batch to which a worker was assigned finished without the worker having worked on it, he was reassigned to another open batch and notified.

3.5.3 Description task

The qualification-requiring description HIT followed Sect. 3.2 and Fig. 3 exactly. Workers received $0.19 per successfully completed HIT. Input data consisted of the 55 fragments which were described in our initial survey run, and for which reuse and quoting permission have been granted. Due to a technical issue, the results for one music fragment had to be invalidated after finalization of the batch, causing 54 remaining music fragments to be considered in our further analysis in Sect. 6. Per HIT, three assignments were released. However, some songs would ultimately get four different descriptions out of this HIT, due to the re-release of a small number of

[4] Upon the release of a HIT batch, the AMT platform randomizes the order in which assignments are presented to workers, regardless of how many assignments already have been completed for a given HIT.

assignments to a motivated worker who had an audio player incompatibility problem.

3.5.4 Ranking task

The music ranking HITs exactly followed the specification given in Sect. 3.5.4. Once again, workers received \$0.19 per succesfully completed HIT. For the cinematic scene descriptions for which a worker should rank three music fragments according to their fit, we used descriptions for the 55 fragments for which reuse and quoting permission have been granted in our initial survey run. Due to the same technical issue as mentioned in Sect. 3.5.3, the results for one music fragment had to be invalidated after finalization of the batch, causing 54 remaining music fragments to be considered in our further analysis in Sect. 5.

In comparison to the original survey answers, a few minor clean-up changes were made. First, answers that had been given in Dutch (which was allowed for our initial survey, for which recruiting started in The Netherlands) were translated to English. Second, basic spelling correction was performed on the answers. Finally, we removed concrete music timestamp references from descriptions ("*The beginning (0–33 s): Big overview shots of the protagonist*"), to avoid the HIT becoming a timestamp matching task.

The two other fragments to offer with the stimulus fragment were randomly chosen without replacement from our production music database, only considering fragments with a duration between 30 s and 5 min that had not been used already in our other HIT batches. We released this HIT in three separate batches, each requiring a different qualification to keep the worker pools mutually exclusive, with three assignments per HIT. This choice for multiple batches had two reasons: first, having multiple isolated batches ensures a larger minimum number of required workers (in this case, it meant that 3 × 3 workers were needed per provided description). Second, since we generated different triplets for each batch, a stimulus fragment would ultimately be compared against $3 \times 2 = 6$ random fragments.

4 Crowdsourcing statistics

General crowdsourcing statistics for all previously described task runs are shown in Tables 1 and 2. As can be seen in these tables, for almost every task run a few HITs were rejected. For our description tasks (initial survey, qualification task, and qualification-requiring description task), we rejected assignments with blank or nonsense responses in the free-form text fields. For the music ranking batches, we rejected assignments with blank or nonsense motivations for the chosen ranks and ratings. In all cases, rejected assignments were republished for other workers to complete.

For the batches run on the formal AMT platform, only three workers (two in Batch A of the ranking task and one in the description task) were found to challenge the system by consistently copying over a few uninformative answers as assignment responses. In other cases, rejections involved incomplete responses that largely appeared to have resulted from unintentional human error, since the workers causing them typically performed well on other assignments. It is striking to note that the largest absolute number of rejected assignments occurred in our initial survey run. As it turned out, these assignments were performed by a small number of respondents who did not grasp the importance of providing a

Table 1 Statistics for the initial and qualification survey tasks

	Running time (days)	No. of workers	Average worker age	No. of approved assignments	No. of rejected assignments	No. of qualifications granted	Median completion time (approved only)
Initial run	20	49	32.6	65	23	N/A	19 min 48 s
Qual. 'Exciting Trailer'	15	133	27.8	130	3	108	9 min 16 s
Qual. 'Mer Bleue Boogie'	11	58	26.7	55	3	40	9 min 54 s
Qual. 'Origo Vitae'	11	65	26.6	63	2	47	7 min 52 s

Table 2 Statistics for the qualification-requiring follow-up tasks

	Running time (until completion)	No. of qualified workers	No. of actual workers	No. of approved assignments	No. of rejected assignments	Median completion time (approved only)
Ranking Batch A	10 days 15 h 10 min	43	11 (25.6 %)	165	20	7 min 33 s
Ranking Batch B	3 days 8 h 4 min	31	12 (38.7 %)	165	7	4 min 42 s
Ranking Batch C	1 days 1 h 53 min	21	12 (57.1 %)	165	0	3 min 50 s
Description	10 days 5 h 5 min	89	32 (36.4 %)	173	18	8 min 23 s

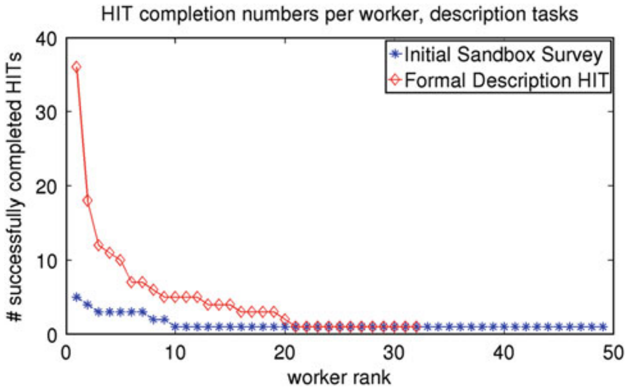


Fig. 5 HIT effort distribution for the original survey and qualification-requiring description tasks. The *horizontal axis* represents the rank of workers, when sorted by the number of successfully completed HITs

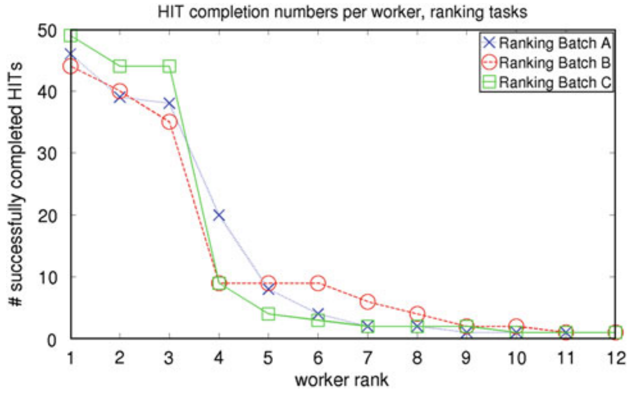


Fig. 6 HIT effort distribution for the three ranking task batches. The *horizontal axis* represents the rank of workers, when sorted by the number of successfully completed HITs

complete response, leaving all free-form text questions blank because they found them too hard to answer.

Selective worker behavior regarding input data can be noted in Table 1 in the number of approved assignments. Out of the three possible qualification music fragments, ‘Exciting Trailer’ received more than twice as many responses as the other two fragments. This cannot be fully explained by the longer running time of the fragment’s HIT, and may have been due to a general preference for this fragment.

From Table 2, insight can be obtained in the return rate of qualified workers. As can be seen, while many workers indicated they very much liked their qualification HIT, this did not guarantee that they also returned to do more tasks. Our found percentages of qualified workers that returned to perform more HITs are consistent with earlier reported results in [37] and underline the need to accommodate for a substantially larger pool of potential workers, if a certain minimum worker pool size should be met in order for a HIT batch to complete successfully.

Finally, information on effort distribution within batches is shown in Figs. 5 and 6. For all tasks performed on the regular AMT platform, it can be noted that most of the work is performed by a relatively small amount of motivated workers. This effect is the strongest for the music ranking batches, which all had three highly active workers performing the majority of available assignment tasks (although no single worker managed to perform HITs for all the data in the batch). The qualification-requiring description task required longer completion times and more personal, creative input, which makes it harder for a worker to build up an efficient routine for this task. Indeed, the effort distribution curve for this task as shown in Fig. 5 is much smoother than in the case of the music ranking batches shown in Fig. 6.

5 Music ranking task results

In order to test the recognizability of stimulus fragments given a cinematic description, our analysis will address two questions:

1. To what extent are the provided rankings and ratings consistent across workers?
2. Is the perceived fit to the description larger for the stimulus fragment in comparison to random fragments?

5.1 Rating consistency

As a measure for inter-rater consistency in case of more than two raters per item, the Fleiss kappa [12] is frequently employed (in the Music-IR field this was, e.g., done in [16]). However, to give a statistically valid result, the measure assumes that different items are rated by different sets of raters. For the AMT crowdsourcing setup, this assumption will not hold, since one worker may work on as many assignments as the number of HITs, thus causing large and unpredictable variation in the rater distribution over different items. This variation also makes it difficult to reasonably normalize 5-point scale rating scores per worker to prevent polarization biases. In our analysis, we therefore will treat the given ratings not only as absolute numbers, but also as indicators of relative orderings.

In order to get a measure of rating consistency, we chose to take a similar approach to Urbano et al. [41], in which crowdsourced worker inter-agreement was measured based on preference judgments for melody similarity. In preference judgments HITs, a worker was provided with a melody and two variations. The worker then had to indicate which of the two variations was more similar to the given melody. If n workers are assigned per HIT, $\binom{n}{2}$ worker pairs can be chosen out of this worker pool.

For each pair, an agreement score was computed. If both workers prefer the same item, two points are added to an agreement score. If one of the workers indicated that both variations were equally (dis)similar to the melody, one point is added to an agreement score. If the workers prefer different items, the agreement score is not increased. The agreement scores for all individual HITs are summed, and divided by the maximum obtainable score of $\binom{n}{2} \times 2 \times \#$ HITs.

In our case, for every assignment in every HIT we converted the three fragment ratings of a worker into an analogous form, by transforming them into $\binom{3}{2}$ pairs of orderings. For example, if three fragment ratings are [5; 2; 3], we now encode them as [$5 > 2;\ 5 > 3;\ 2 < 3$], reflecting their relative ordering rather than absolute rating values, which from a retrieval perspective is reasonable. These are then turned into agreement scores by comparing how pairs of workers judged the relative ordering, in the same way as described above.

Out of the 3 batches × 54 fragments × 3 assignments × 3 ratings per assignment = 1,458 ratings in total, 15 ratings were missing in our data. Out of these, for 12 missing ratings (0.82 % of the total), the workers had given sufficient information via other input fields to clarify their view on the fragments, showing the benefit of our redundancy mechanisms. In three cases (0.21 %), a rating was missing due to an audio loading issue.

Since our chosen technique to compute agreement scores does not allow for such missing values, we applied a data imputation procedure before performing the computation. If the unrated fragment was indicated as the best fit, it got the rating score of the second ranked item, regardless if the worker indicated the first item to be a much better fit. Similarly, if the unrated fragment was indicated as the second best fit, it got the rating score of the worst ranked item. If the unrated fragment was indicated as the worst fit, it got the rating score of the second ranked item, regardless if the worker indicated the worst item to be much worse than the second ranked item. In case there was an audio problem, the rating for the corresponding item would be -1, causing the item to automatically be ranked last. This is a conservative approach, such that the results obtained on data to which this policy is applied can be considered to be lower bounds to the actual results. The agreement scores obtained for the three batches are given in the first column of Table 3. As can be seen, the scores indicate high agreement between the workers regarding the ordering of fragments.

Table 3 Rating agreement score and rank attributions (in %) for the stimulus fragments

	Rating agreement	Rank 1	Rank 2	Rank 3
Batch A	0.7346	83.33	10.49	6.17
Batch B	0.6934	69.14	22.22	8.64
Batch C	0.6317	72.84	14.81	12.35
Overall	0.6866	75.10	15.84	9.05

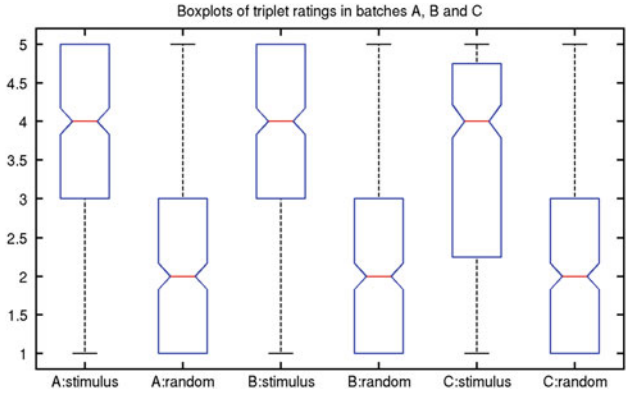


Fig. 7 Boxplots of the fragment fit ratings (without data imputation on missing values) for the three HIT batches. For every batch, ratings are plotted for the original stimulus fragment and the randomly chosen fragments

5.2 Stimulus fragment versus random fragments

The agreement measure across rating orderings gives only an indication of intra-rater ranking ordering behavior. In order to verify whether the stimulus fragment actually is being ranked higher than the random fragments it occurs with, we consider the frequencies of rank occurrences for the stimulus fragment. As shown in Table 3, in a large majority of cases, the stimulus fragment is indeed ranked first for all three batches. Looking across batches at the nine rankings that are available for each song, there always are at least three workers considering the stimulus fragment as the best fitting fragment to a description.

Finally, to still consider the absolute rating numbers across batches, boxplots for the ratings of the provided music triplets in all three batches are shown in Fig. 7. These are provided per batch, starting with a boxplot of the stimulus fragment, and followed by a boxplot of the ratings obtained for the randomly chosen songs. Here, we did not apply any data imputation procedures, so missing rating values were omitted. Applying the Kruskal–Wallis test to the ratings with a subsequent Tukey–Kramer Honestly Significant Difference Test, it turns out that at $p < 0.05$, the three stimulus fragment groups differ significantly from the three random fragment groups ($p \approx 0$). Thus, it can be concluded that stimulus fragments are rated differently from randomly chosen fragments, and, as our measures regarding ordering and ranking showed, that the stimulus fragments are clearly considered as better fits to the description than the random fragments.

6 Common elements: analysis of description responses

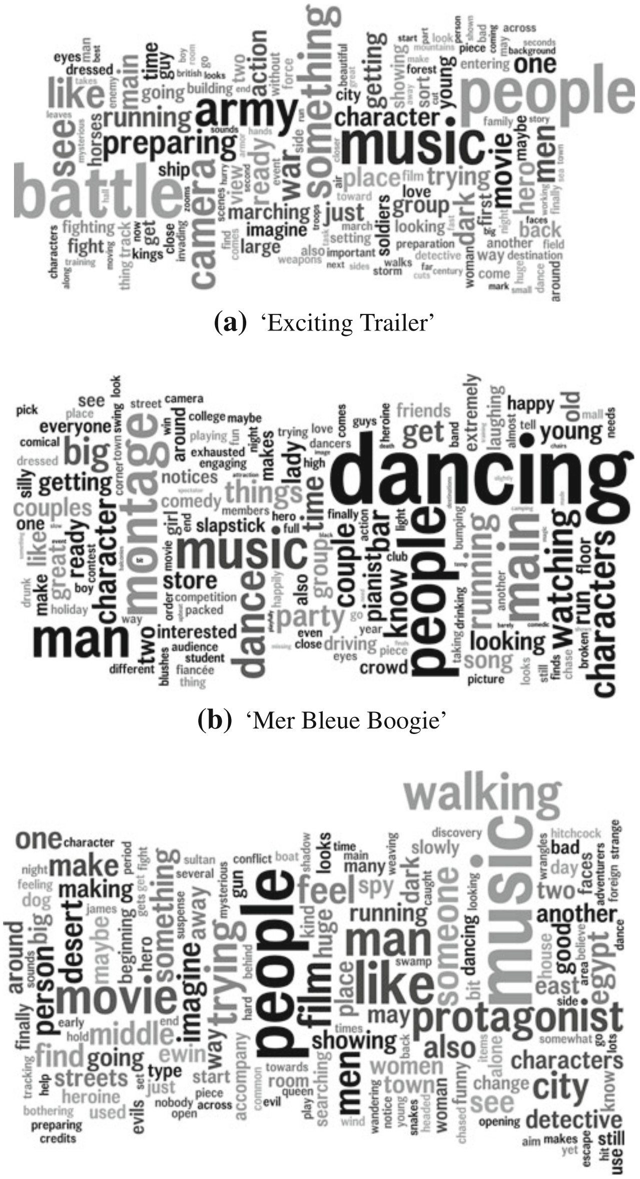
We now arrive at the essential focus of our work: investigating characteristics of the given contextual description responses. Given the amount of individual freedom we granted for these descriptions, analyzing and identifying such characteristics is not straightforward. Nonetheless, in this section we strive to provide deeper insights into them by addressing three questions:

1. Do common narrative elements emerge for a certain song, as soon as we receive a sufficient amount of contextual descriptions for it?
2. What types of common narrative elements can be identified in general for this type of description?
3. Are users capable of indicating why they chose particular scene descriptions for particular songs?

6.1 Many descriptions of the same music fragment

Through our qualification tasks, we acquired a relatively large number of descriptions for several music fragments. Word clouds for obtained scene and location descriptions for the fragments are shown in Figs. 8 and 9. As can be seen, for different fragments, there are different vocabularies and common words. Looking at the full answers the following generalizations and emerging categorizations can be made:

1. 'Exciting Trailer' (108 qualified responses) is a symphonic piece for which the composer stated that "*A militaristic snare drum march begins this piece, reminiscent of Eastern Europe during World War II*". The fragment gives the majority of worker respondents a sense of *preparation for a situation involving power*. In terms of actors, emerging categories are *fighters* (26.85 %) and *heroes* (24.07 %). A majority of respondents (64.81 %) situates the described scene outdoors, and from the 44 respondents who went as far as indicating a geographic location for the scene, frequently mentioned regions are *Europe* (47.92 %) and the *USA* (35.41 %).
2. 'Mer Bleue Boogie' (40 qualified responses) is a boogie-woogie piano piece, described by its composer as "*A very up-tempo piano boogie. Heavy swing on this one. It should have your toes tappin*"'. To the worker respondents, the fragment strongly evokes *dancing/party scenes*. In terms of actors, the most frequently mentioned categories are *couples/duos* (37.5 %) and *sole male protagonists* (20 %). A majority of respondents (65 %) situates the described scene indoors. It is striking that the mentioned location categories are very strongly *urban* (91.43 %), and out of the 14 respondents who went as far as indicating a geographic location for the scene, 13 (92.86 %) situate the scene in the *USA*.
3. 'Origo Vitae' (47 qualified responses) was described by its composer as being "*mysterious and intense*". To the worker respondents, it evokes mysterious, unknown and sometimes unpleasant situations. The most frequently mentioned actor category is that of *adventurers* (21.28 %). A majority of respondents (76.6 %) situates the described scene outdoors. In relative terms, this fragment evoked the largest number of geographic area responses (31 respondents), with 15 (49.38 %) of the respondents situating the scene in the *Middle East*.



(a) 'Exciting Trailer'

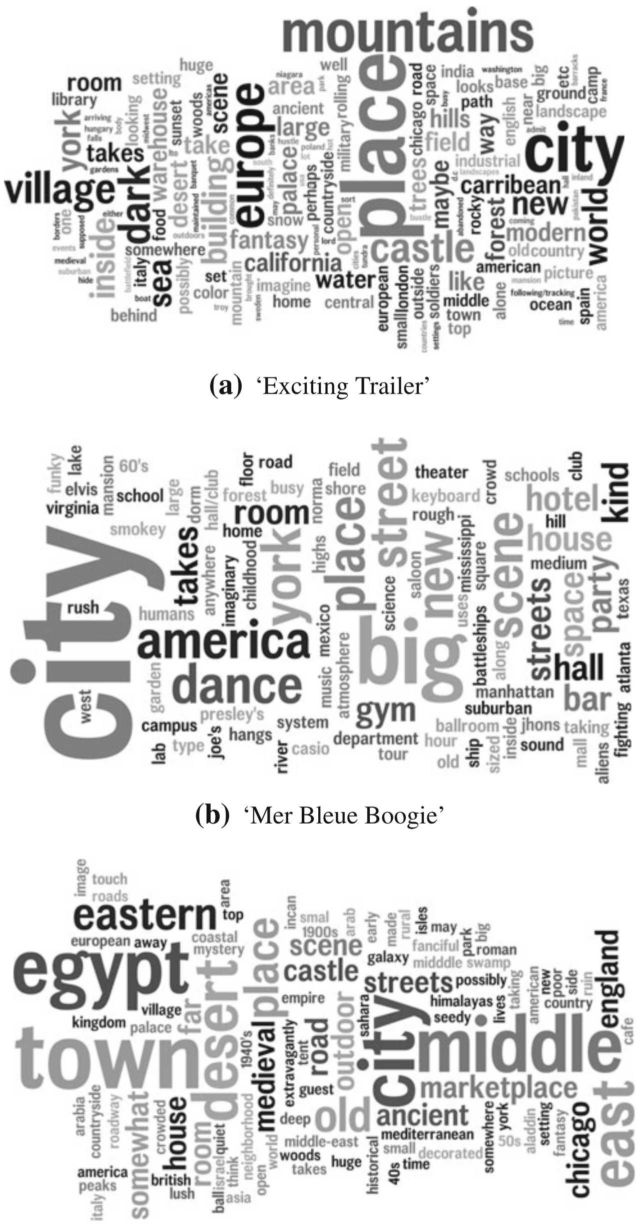
(b) 'Mer Bleue Boogie'

(c) 'Origo Vitae'

Fig. 8 Word clouds for qualification fragment scene descriptions by qualified workers. Common English stop words were removed. In addition, the frequently occurring word 'scene' was removed from the word clouds, since it never occurred as an element of the actual scene

6.2 Generalizing over more fragments: event structure

Already from the responses to our initial survey, we noted apparent agreement between respondents at the level of linguistic event structures. In particular, we identified four



(a) 'Exciting Trailer'

(b) 'Mer Bleue Boogie'

(c) 'Origo Vitae'

Fig. 9 Word clouds for qualification fragment location descriptions by qualified workers. Common English stop words were removed

event structure classes, motivated by the Vendler event typology on events types that can be expressed by human language [42] (examples in parentheses are taken from the survey responses):

- Class 1: Activity or state with no goal ("*A group of people working in a factory. There is steam everywhere. They work hard, but they are not negative about that*");
- Class 2: Activity with underspecified goal and no certainty of achieving the goal ("*A scene of awaiting. Something is going to happen*");
- Class 3: Activity with well-specified goal and no certainty of achieving the goal ("*Warriors are moving fast in the darkness and trying to sneak into their enemies' campsite*");
- Class 4: Activity with well-specified goal and certainty of achieving the goal ("*Someone is walking in a dense forest and finally arrives at village where the inhabitants are dancing slowly*").

To verify the validity of these structure classes, we coded all responses belonging to the 54 music fragments that were also used in our ranking tasks, both from the initial survey run and the qualification-requiring description task. In doing this, we took the perspective of the protagonist at the narrative level that Bal [1] calls the *fabula*: the level consisting of the narrative as a pure chain of events, without any 'coloring' of it towards the audience. For 44 out of 54 fragments, a majority of respondents described a story of the same event structure class. For these cases, we studied the amount of agreement by framing the problem as a classification problem, in which the majority vote on the event structure class would be considered as the actual class. By counting actual class occurrences for every majority vote class, we can construct a 'confusion matrix' C, which for our fragments was

$$C = \begin{bmatrix} 49 & 7 & 12 & 12 \\ 4 & 15 & 3 & 4 \\ 5 & 7 & 43 & 5 \\ 5 & 3 & 0 & 12 \end{bmatrix}$$

where $c_{m,n}$ indicates the number of descriptions for event structure class n, when a majority of respondents described a story of event structure class m.

The most striking result is that class 1 frequently gets mixed with the other classes. This may be because of the way our descriptions were solicited: even when not feeling any goal to be pursued in the music, a respondent could still have been triggered by our questions such that a full story was conceived with a goal in it. However, this situation occurs much less the other way around. If a music fragment invokes a scene with a goal for the majority of respondents, this majority will be large. In addition, respondents are strongly agreeing on uncertainty of reaching a goal (classes 2 and 3).

In order to see how the classes are balanced if many responses are available for a song, we return to the qualification fragments again. The results are shown in Table 4. For 'Mer Bleue Boogie', the most frequently occurring class is class 1, but similarly to the confusion matrix, we notice that this is no strong majority either. On the other hand, the other two fragments, in particular 'Exciting Trailer', do have a strong majority class. 'Origo Vitae', which was supposed to be a mystery fragment, has considerably more 'vague goals' (class 2) than the other fragments, or the general trend observed in the confusion matrix.

Table 4 Event structure class occurrences (in %) for the qualification fragments

	1	2	3	4
'Exciting Trailer'	19.62	3.74	57.94	18.69
'Mer Bleue Boogie'	42.5	5.0	37.5	15.0
'Origo Vitae'	8.5	23.4	48.94	19.15

6.3 Self-reported reasons for descriptions

Are workers capable of indicating why the music made them choose their descriptions? This information is relevant to investigate salient high-level description features that can be used for query construction and to which low-level feature extractors should be mapped.

In order not to bias against songs, we once again considered our set of 231 descriptions for 54 music fragments. We categorized all reported motivations, starting from Tagg's earlier-mentioned musical sign topology [39], and adding any other emergent categories. This led to 12 categories:

1. Sonic anaphone: "*The high tones sound like fluttering birds, the lowerer (sic) tones sound like a grumbling character*".
2. Kinetic anaphone: "*The bass line is paced and deliberate. It gave me the sense of someone sneaking around*".
3. Tactile anaphone: "*The music brings to mind picture of something moving slowly, uniformly and languidly like a falling night, setting sun or rising sun, sluggish waters, billowing smoke etc. It also has a touch of romance and warmth. Thus a loving couple in a moonlit night walking slowly in a world of their own came naturally to mind*".
4. Genre synecdoche: "*The saxophone prominence leads me to believe that this film is set in a past generation*".
5. Episodic marker: "*The ascending notes which reach a peak and then decline*".
6. Style indicator: "*The music definitely seemed like something you would hear in a club*".
7. Temporal development: "*It has a beat which (sic) some energy in it, but sometimes it halts, and is there (sic) space for the serious part of a conversation, or just a normal diner (sic)*".
8. Character trait/psychology of actor: "*It is rock enough to make me think of a biker dude*".
9. General quality/atmosphere: "*Because the fragment seems to possess some kind of cool feeling: it is not too fast paced but neither is it too slow. Also the sounds/pads used seem to possess this feeling*".
10. Seen before: "*Because the fragment reminds me of a robbery in today's popular action movies*".
11. Personal relation: "*It reminds me of my jolly school days*".
12. Intuition: "*Just something that popped in my head the minute I heard this music*".

The answer distribution over these categories is shown in Table 5. As it turns out, and was hypothesized at the start of our work, respondents can generally not explain their description reasons at the level of Tagg's musical sign topology (constituted by categories 1–6), except in case of the kinetic anaphone category. At the same time, from an application-oriented requirements elicitation perspective, it is good to see that the very individualistic category 11, and the uninformative category 12 only amount to <11 % of the represented categories. Three meta-categories stand out:

1. *Temporal aspects* (categories 2 and 7, can include 5): temporal developments, movement indicators and episodic changes. In our free-form setting, workers tended to talk much more in terms of characteristics of different episodes that they observed ("*the parts where there are constant harmonies in the background need a more quiet scene than the parts where there is only the beat*", "*At first, I thought that perhaps a human/hacker type of scene would be appropriate, but it does not explain the breaks in the music*") than in terms of markers announcing an episode change. Similar to the findings in [31], workers appear to be highly sensitive to temporal developments that indicate a sound change, but from a musical structure analysis or discourse point would not be significant: e.g., "*It has blunt pauses in-between. The end is also blunt and sounds like 'this just does not work'*", "*The end is a possitive (sic) note and it is as though somebody realises something*".
2. *Psychology, quality and atmosphere* (category 9, can include 8): categories relating to Lissa's functions [26] of indicating a character's psychology, and providing empathy to the viewer. These will be closely related to affective aspects.
3. *Previously seen examples* (categories 4, 6 and 10): knowledge of existing films and styles are steering the evocation directions, and workers are aware of this.

6.4 Further notions

There were two more aspects in the received descriptions that we found worth mentioning in this section. While these are

Table 5 Category occurrences (in %) for self-reported scene description reasons

1	2	3	4	5	6	7	8	9	10	11	12
5.62	10.06	1.12	7.87	0.84	8.15	12.64	7.58	21.63	12.92	5.06	5.62

currently anecdotical, they seem to conform to theories on cognition, expectation, and familiarity as, e.g., investigated by Huron [13]. As such, they show interesting opportunities for cross-disciplinary future work.

6.4.1 Cultural influences

Connotative meaning is largely built on cultural conventions. Overall, our results (which largely reflected Western mass media culture) showed good generalizability measures, indicating that characteristics of this culture are widely recognized, also by workers who are not from countries of this culture. However, there was one example in which culturally specific influences seem to be present: 'Origo Vitae'. While the Middle East was a popular geographical region, none of the 8 workers from India mentioned this, or any 'exoticism' at all. In music-theoretic terms, the piece features a lowered second degree in comparison to a harmonic major scale, which has grown to be an Arabic cliché in Western art music, but may not be recognized as such by others.

6.4.2 Major versus minor ≠ happy versus sad

A common conception regarding the relation of musical keys to affective properties is that major keys imply positive feelings and minor keys negative feelings. In our survey responses, we observed a few situations in which this situation was more nuanced.

For three quiet major-key songs, workers agreed on characterizing the song as positive to very positive. However, in the scenes described with it, feelings were mixed: e.g., "*The father dies slowly after having a perfect time with his son. His son gets sad, but at the same time experiences relief due to having the last chance to see him after being apart for 30 years*", "*The young man in the first scene would be a man with a bright smile, an almost glowing optimism about him, everything in his demeanor a happy and positive. The second scene of this boy would be a older boy, his eyes more wrinkled around the edges, the rings under his eyes deeper, a troubled yet stern look on his face, the face of a man who has stayed up long nights worrying and many times his eyes witnessing the horrible acts war exposes one to, slowly rending tears in this mans happiness and his soul, leaving him a dry and unhappy husk of the man he once was*".

Furthermore, there were two cases of minor-key silent film piano music, which were intended by the original composer as indeed belonging to negatively valenced scenes: "*Short Theme in two tempos for your bad guy*" and "*this staccato piano piece is the tense setting for the classic silent film scene: the heroine is tied to the train tracks, the steam engine train plugging along towards her*". However, workers do not feel this at all anymore, and rather trigger on the staccato piano playing style, which is considered as bright and uplifting: "*Harry is bulky and brainless. He is always confused. Even if he wants to do the job in a way he feels proud off, he just cant (sic) make it. Poor thing*", "*A few young children are playing hide and seek*". If any connections are made with existing movies, these are to comedy and slapstick genres (e.g., Laurel and Hardy, The Addams Family).

7 Conclusion and outlook

We presented a study of narrative multimedia descriptions that people connect to music fragments. In order to obtain a sizable number of participants to our study, we employed a crowdsourcing strategy for recruitment. As it turned out, running our highly open-ended task on a mainstream crowdsourcing platform did not present any disadvantages over a more controlled recruiting approach. In contrast, gathering input was more efficient, and there were no significant drawbacks in terms of user input quality.

When provided with earlier obtained descriptions, workers were able to reliably recognize the stimulus fragment that evoked the description. When multiple descriptions were gathered for the same music fragment, fragment-specific profiles emerge in terms of actor, location and story types. Based on the provided worker input, event structure classes that consider the absence or presence of a goal and the certainty of achieving this goal were deduced as a conceptual level at which different free-form worker descriptions will be generalizable. In most of the cases studied, there was a majority preference for an event structure class, and in case the preference considers a narrative involving a goal, this majority is strong. This suggests that automated analysis methods should be conceivable that can map to such goal-oriented classes.

For our current fragment ranking task, our experiments contained a stimulus fragment versus two randomly chosen fragments. This choice to randomly pick the alternative fragments was made to allow for an objective experimental situation, without any potential selection bias. Following the indications regarding salient narrative themes and event structures resulting from our description task, follow-up experiments are imagineable in which the fragment ranking task is repeated, but the alternative fragments are picked in a (semi-)supervised way, considering these findings from our description task.

As we found, workers are not very good at explaining their connotative associations in terms of musical characteristics. Nonetheless, their reports reveal sensitivity to associations between temporal development and musical movement, to affective and psychological effects, and to existing prototypical examples. Two remarks should be made here that need more investigation in the future:

- As for the first self-reported association mechanism of temporal development and movement, the features on which workers trigger seem at a higher level than mere beat or tempo tracking (and related to sound timbre), but at a lower level than formal music structure analysis.

– Future work in this area in the fields of affect and general learning should take into account that connotative associations may involve several signal-unrelated steps ("*I enjoyed this piece of music. It seemed 80 s adn (sic) joyful, so that made me think of a romcom's ending. It seemed a little spacy and techy, so that made me think of Geek Romance*"), and that, while consensus is reached on the general affective content of a music fragment, multiple layers of affect will get involved as soon as the multimedia context is involved ("*I thought the music sounded kind of wistful, but on the same hand it also had some sense of pride in it*").

Our approach to the problem took a cross-disciplinary view inspired by current thought in linguistics and musicology. For several decades, the field of musicology has already been acknowledging and studying the importance of connotative meaning, albeit not from a data-oriented perspective, avoiding any absolutist and positivist approaches. With the rise of the Social Web, we are currently able to take a data-oriented perspective that is not necessarily absolutist or positivist, but rather driven by cultural communities and, as we showed in this article, connotation-aware.

Our findings that the connotative connections between free-form and spontaneous descriptions of visual narrative and musical information are strong enough to be recognizable and generalizable open up new perspectives for multimedia-oriented music querying and retrieval. Music queries in this context do not need to be confined to musical vocabulary, but can be constructed in a user-friendly narrative form created under consideration of the envisioned end result. Such queries constitute versatile and sophisticated multimedia messages, and, as our results suggest, the goal of deepening out the connotative layer underlying these messages is feasible to pursue in the near future.

Acknowledgments Cynthia Liem is a recipient of the Google European Doctoral Fellowship in Multimedia, and this research is supported in part by this Google Fellowship. We are grateful to the composers of the royalty-free production music fragments for their willingness to share their creative work, to the anonymous reviewers for their suggestions for improvement, and to Vera Liem for her help in coding the survey responses.

References

1. Bal M (2009) Narratology—introduction to the theory of narrative, 3rd edn. University of Toronto Press, Toronto
2. Barthes R (1977) Image music text. Hill and Wang, New York
3. Benini S, Canini L, Leonardi R (2011) A connotative space for supporting movie affective recommendation. IEEE Trans Multimed 13(6):1365–1370
4. Cai R, Zhang C, Wang C, Zhang L, Ma W-Y (2007) MusicSense: contextual music recommendation using emotional allocation modeling. In: Proceedings of the 15th annual ACM international conference on multimedia, Augsburg, Germany
5. Chatman SB (1980) Story and discourse: narrative structure in fiction and film. Cornell University Press, Ithaca
6. Colombo C, Del Bimbo A, Pala P (2001) Retrieval of commercials by semantic content: the semiotic perspective. Multimed Tools Appl 13:93–118
7. Cook N (1998) Music—a very short introduction. Oxford University Press, New York
8. Cunningham SJ, Reeves N, Britland M (2003) An ethnographic study of music information seeking: implications for the design of a music digital library. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries (JCDL '03), Houston, USA
9. J. Downie S, Byrd D, Crawford T (2009) Ten years of ISMIR: reflections on challenges and opportunities. In: Proceedings of the 10th International Society for Music Information Retrieval conference (ISMIR 2009), Kobe, Japan
10. Downie JS, Cunningham SJ (2002) Toward a theory of music information retrieval queries: system design implications. In: Proceedings of the 3rd international conference on music information retrieval (ISMIR 2002), Paris, France
11. Feng J, Ni B, Yan S (2010) Auto-generation of professional background music for home-made videos. In: Proceedings of the 2nd international conference on internet multimedia computing and service (ICIMCS), Harbin, China
12. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
13. Huron D (2006) Sweet anticipation: music and the psychology of expectation. MIT Press, Cambridge
14. Inskip C, MacFarlane A, Rafferty P (2008) Music, movies and meaning: communication in film-makers' search for pre-existing music, and the implications for music information retrieval. In: Proceedings of the 9th International Society for Music Information Retrieval conference (ISMIR 2008), Philadelphia, USA
15. Inskip C, MacFarlane A, Rafferty P (2010) Upbeat and quirky, with a bit of a build: interpretive repertoires in creative music search. In: Proceedings of the 11th International Society for Music Information Retrieval conference (ISMIR 2010), Utrecht, The Netherlands
16. Jones MC, Downie JS, Ehmann AF (2007) Human similarity judgments: implications for the design of formal evaluations. In: Proceedings of the 8th international conference on music information retrieval (ISMIR 2007), Vienna, Austria
17. Kalinak KM (1992) Settling the score: music and the classical Hollywood film. University of Wisconsin Press, Madison
18. Kaminskas M, Ricci F (2012) Contextual music information retrieval and recommendation: state of the art and challenges. Comput Sci Rev 6(2–3):89–119
19. Kuo F-F, Chiang M-F, Shan M-K, Lee S-Y (2005) Emotion-based music recommendation by association discovery from film music. In: Proceedings of the 13th annual ACM international conference on multimedia, Singapore, Singapore, pp 507–510
20. Lang E, West G (1920) Musical accompaniment of moving pictures—a practical manual for pianists and organists and an exposition of the principles underlying the musical interpretation of moving pictures. The Boston Music Company, Boston
21. Law E, Von Ahn L (2009) Input-agreement: a new mechanism for collecting data using human computation games. In: Proceedings of ACM CHI 2009, Boston, USA
22. Lee JH (2010) Analysis of user needs and information features in natural language queries seeking user information. J Am Soc Inform Sci Technol 61(5):1025–1045

23. Lee JH, Cunningham SJ (2012) The impact (or non-impact) of user studies in music information retrieval. In: Proceedings of the 13th International Society for Music Information Retrieval conference (ISMIR 2012), Porto, Portugal
24. Lee JH, Hill T, Work L (2012) What does music mood mean for real users? In: Proceedings of the iConference, Toronto, Canada
25. Li C-T, Shan M-K (2007) Emotion-based impressionism slideshow with automatic music accompaniment. In: Proceedings of the 15th annual ACM international conference on multimedia, Augsburg, Germany
26. Lissa Z (1965) Ästhetik der Filmmusik. Henschelverlag, Berlin
27. Mandel MI, Eck D, Bengio Y (2010) Learning tags that vary within a song. In: Proceedings of the 11th International Society for Music Information Retrieval conference (ISMIR 2010), Utrecht, The Netherlands, pp 399–404
28. Mason W, Suri S (2012) Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Method 44(1):1–23
29. Meyer LB (1968) Emotion and meaning in music. The University of Chicago Press, Chicago
30. Nack F, Hardman L (2001) Denotative and connotative semantics in hypermedia: proposal for a semiotic-aware architecture. New Rev Hypermedia Multimed 7(1):7–37
31. Nattiez J-J (1973) Y a-t-il une diégèse musicale? In: Faltin P, Reinecke H-P (eds) Musik und Verstehen – Aufsätze zur semiotischen Theorie., Ästhetik und Soziologie der musikalischen Rezeption Arno Volk Verlag, Cologne, Germany, pp 247–257
32. Nowak S, Rüger S (2010) How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on multimedia information retrieval, MIR '10. ACM, New York
33. Paolacci G, Chandler J, Ipeirotis PG (2010) Running experiments on Amazon Mechanical Turk. Judgm Decis Mak 5(5):411–419
34. Prendergast RM (1992) Film music: a neglected art—a critical study of music in films. Norton, New York
35. Schedl M, Knees P (2009) Context-based music similarity estimation. In: Proceedings of the 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS 2009), Graz, Austria
36. Soanes C, Stevenson A (eds) (2008) Concise Oxford English Dictionary, 11th edn. Oxford University Press, NY
37. Soleymani M, Larson M (2010) Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In: Proceedings of the SIGIR workshop on crowdsourcing for search evaluation (CSE 2010), Geneva, Switzerland
38. Stupar A, Michel S (2011) Picasso—to sing you must close your eyes and draw. In: Proceedings of the 34th annual ACM SIGIR conference, Beijing, China
39. Tagg P, Clarida B (2003) Ten little title tunes—towards a musicology of the mass media. The Mass Media Scholar's Press, New York/Montreal
40. Turnbull D, Barrington L, Torres D, Lanckriet G (2008) Semantic annotation and retrieval of music and sound effects. IEEE Trans Audio Speech Lang Process 16(2):467–476
41. Urbano J, Morato J, Marrero M, Martín D (2010) Crowdsourcing preference judgments for evaluation of music similarity tasks. In: Proceedings of the SIGIR workshop on crowdsourcing for search evaluation (CSE 2010), Geneva, Switzerland
42. Vendler Z (1967) Linguistics in philosophy. Cornell University Press, Ithaca
43. Vliegendhart R, Larson M, Kofler C, Eickhoff C, Pouwelse J (2011) Investigating factors influencing crowdsourcing tasks with high imaginative load. In: Proceedings of the WSDM workshop on crowdsourcing for search and data mining (CSDM 2011), Hong Kong, China
44. Weigl DM, Guastavino C (2011) User studies in the music information retrieval literature. In: Proceedings of the 12th International Society for Music Information Retrieval conference (ISMIR 2011), Miami, USA
45. Wiering F, Volk A (2011) Musicology. Tutorial slides. In: Proceedings of the 12th International Society for Music Information Retrieval conference (ISMIR 2011), Miami, USA