On the Automatic Identification of Music for Common Activities

Yadati, Karthik; Liem, Cynthia C.S.; Larson, Martha; Hanjalic, Alan

DOI
10.1145/3078971.3078997

Publication date
2017

Document Version
Final published version

Published in
Proceedings of the 2017 ACM International Conference on Multimedia Retrieval

On the Automatic Identification of Music for Common Activities

Karthik Yadati
Delft University of Technology
The Netherlands
n.k.yadati@tudelft.nl

Cynthia C. S. Liem
Delft University of Technology
The Netherlands
c.c.s.liem@tudelft.nl

Martha Larson
Radboud University & Delft University of Technology
The Netherlands
mlarson@science.ru.nl

Alan Hanjalic
Delft University of Technology
The Netherlands
a.hanjalic@tudelft.nl

ABSTRACT

In this paper, we address the challenge of identifying music suitable to accompany typical daily activities. We first derive a list of common activities by analyzing social media data. Then, an automatic approach is proposed to find music for these activities. Our approach is inspired by our experimentally acquired findings (a) that genre and instrument information, i.e., as appearing in the textual metadata, are not sufficient to distinguish music appropriate for different types of activities, and (b) that existing content-based approaches in the music information retrieval community do not overcome this insufficiency. The main contributions of our work are (a) our analysis of the properties of activity-related music that inspire our use of novel high-level features, e.g., drop-like events, and (b) our approach's novel method of extracting and combining low-level features, and, in particular, the joint optimization of the time window for feature aggregation and the number of features to be used. The effectiveness of the approach method is demonstrated in a comprehensive experimental study including failure analysis.

CCS CONCEPTS

•Information systems →Music retrieval; •Human-centered computing →*Social recommendation; Social media; Social tagging;*

KEYWORDS

Music recommendation; Activity; Workout music; Relax music; Study music

ACM Reference format:
Karthik Yadati, Cynthia C. S. Liem, Martha Larson, and Alan Hanjalic. 2017. On the Automatic Identification of Music for Common Activities. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078997

ICMR '17, Bucharest, Romania

978-1-4503-4701-3/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078971.3078997

1 INTRODUCTION

In addition to "just" listening to music as a part of our leisure, we can also use music to facilitate our daily activities. For example, listening to appropriate music while studying can help us focus better [3]. While there exist online music services specialized in this direction (e.g., Focus at will and Brain FM), the mechanisms underlying their offerings are either automatic music generation or fully manual curation. Moreover, they only cover a limited scope of activities (e.g., focus only). A recent study on using "music as a technology" [5] to accomplish a goal highlights how little research has been done on developing music recommender systems for daily activities. Our paper delves into this area and investigates the main challenges in automating the process of identifying existing music for different daily activities.

The first challenge we face is deriving a list of daily activities for which music is sought. Instead of simply predefining this list, as is typically done in the existing work (e.g., [18]), we mine a social media sharing platform (YouTube) to derive a list of popular activities i.e., those activities that are found to have frequent mention when searching for activity-related music. Our mining approach is similar to that taken by recent work [9] on identifying common user-intent categories in online video search.

The second challenge we address in this paper is to find an appropriate approach to automatically recognize activity-related music categories. We start by looking at the available textual metadata that we typically find associated with the music tracks that are posted on YouTube and promoted as being suitable for a particular activity. Specifically, the fact that the titles of these music tracks in many cases contain genre/instrument information, leads us to investigate the usability of this specific information first.

Since our findings indicate that relying on music genre or presence of particular instruments is not a reliable approach to link a music track to a particular activity, we proceed by investigating how to develop an approach that works well for the posed problem. For this purpose, we look into the existing content-based approaches in the field of music information retrieval (MIR) and encounter several issues that need to be addressed. The first issue is the temporal variation of musical content in a given track. We hypothesize that the time resolution for feature extraction in standard MIR tasks is not appropriate to capture the variations for activity-based music classification. Traditional MIR tasks typically extract features at a resolution of 10-100 ms and then aggregate them over the entire music track (e.g., for emotion detection [8]) or a segment sampled from the track (e.g., for genre recognition [8]). In our paper, we

propose to aggregate features over windows of different time resolution and identify the temporal resolution that can give optimal classification performance.

A related issue we address in this paper is how to represent a music track in the feature space in order to enable effective activity-based classification. We take as our starting point a standard set of low-level features that can be extracted from the music signal. Additionally, we also consider some other sources of information that we, either intuitively or through exploratory experiments, found relevant for the task. Specifically, we consider different dimensions of affect (arousal, valence and tension [12]) and the presence of events like onsets and drops [19]. We encode this information in an additional set of high-level features. Finally, we design a classifier with which we investigate the possibility of identifying different activity-related music categories, and the usefulness of different low- and high-level features for the task. Specifically, in parallel with optimizing the time window for feature aggregation as explained above, we also optimize the number of low-level features to be used.

In summary, the main contribution of this paper consists of the answers to the following research questions:

(1) *RQ1: Which activity categories are popular?* We mine music on YouTube to derive common categories of activities. By analyzing the textual metadata related to the activity-related music tracks, we identify the top-3 activity categories to focus on (Section 3).
(2) *RQ2: Is genre or instrument information helpful in predicting an activity-related music category?* This research question is addressed in Section 4 by using the textual metadata of the music tracks, and in particular the presence of genre/instrument related keywords.
(3) *RQ3: How to automatically identify music for a specific activity?* In Section 5, we investigate two aspects to dealing with this question, viz. the temporal resolution at which we should aggregate features and the types of features that would be helpful for the task.

The contributions listed above are presented after an analysis of the existing work on automatically associating music with daily activities, as well as different feature extraction strategies in Section 2. Experimental results assessing the performance of our proposed classification method and a failure analysis are presented and discussed in Section 6. Section 7 concludes the paper.

2 RELATED WORK

In this section, we look at two different aspects dealt in this paper: associating music with activities and different feature extraction strategies used in the MIR literature.

2.1 Associating music with activities

Wang et al. proposed a method that associates music with specific activities [18]. The authors use a predefined list of activities: running, walking, sleeping, working, studying and shopping, for which they recommend music. Sensors on the mobile phone are used to infer whether the user is in the middle of one of these activities, and then suitable music is recommended based on an analysis of low-level features extracted from the signal. To train the recommender system, playlists for specific activities are collected from an online music sharing platform. Next, a subset of 1200 songs is picked from these playlists and manually labeled with one or more activities as tags. A classification problem is then set up where a model is trained for each activity based on the mean and standard deviation of low-level features extracted from a 512 sample frame extracted every 30 seconds of the song. Wang et al. use the following features for classification: Zero crossing rate, Centroid, Rolloff, Flux, Mel-Frequency Cepstral Coefficients (MFCC), Chroma, Spectral Flatness Measure (SFM) and Tempo. The trained model then predicts activity-based tags for new songs. Similar work is reported by Dias et al. [6], where the system "Improvise" is designed to associate music with daily activities mentioned above.

In our approach, we focus on the categories of activities derived from social media data and base the classification process on a novel feature extraction approach that, as we will explain and demonstrate experimentally, is more suitable for the task.

2.2 Feature extraction

Typical MIR tasks, like genre recognition, mood classification or instrument recognition, have been addressed frequently in the past [8]. Characteristic for these tasks is the way of extracting audio features, namely at the frame (time interval) level and with typically rather small frames, e.g., 10-100 ms for timbre features. In order to extract temporal features, like rhythm, a larger time window with a couple of seconds in length is used. Recently, the research community working on extracting emotion from music argued for using longer time windows and tracking emotions over "emotionally stable" segments [1], [15]. We take this discussion a step further by investigating segments of differing lengths while aggregating features for music-to-activity mapping. Additionally, we enrich the set of common audio features by new high-level features that we find especially useful for the task.

3 WHICH ACTIVITIES ARE POPULAR?

In this section, we address RQ1, i.e., we identify types of activities during which users commonly listen to music. Many daily activities are potential occasions for listening to music. A priori, examples include commuting, taking a shower, cooking, cleaning the house, studying or working out. However, compiling an exhaustive list of music-accompanied activities would require difficult-to-acquire behavioral information. For this reason, we focus on activities that are publicly mentioned, and can be assumed to be important to a substantial portion of general population. We turn to social media platforms as an information sources. Specifically, we analyze textual metadata on YouTube for common mentions of activities, which we take as providing indication of their popularity and wide-spread importance to users seeking music online.

When listeners are searching music for specific activities, we assume that search queries could take on various common forms, e.g., "Music for *", where the wildcard * could refer to a specific activity (e.g., studying, workout or jogging). Our metadata analysis is based on the observation that this query consists of a conjunction or a preposition connecting the other two words. In order to construct queries that would allow us to identify common activities,

we looked at all possible prepositions[1] and conjunctions[2] that can follow the word "music". In this way, we arrived at five different word pairs: "music for", "music to ", "music when ", "music while" and "music during". By enclosing the word pairs in quotations, we created a query that could be matched with track metadata (i.e., title and description).

YouTube is a rich source of music, and offers a wealth of music intended for different activities, e.g., Study[3] or Workout[4]. In general, such music takes form of long tracks with durations typically exceeding 30 minutes. We use the queries just discussed to identify these tracks on YouTube. For each of the 5 queries, we follow these steps to collect the tracks and the corresponding metadata:

(1) Using a web crawler, go through all the pages returned by YouTube for a given query and collect the unique identifiers of the videos as well as the titles.
(2) Download the mp3 audio of the videos and the corresponding metadata, e.g., title, description and likes.
(3) Remove duplicates in the search results and also remove the results that are not music tracks.

We accumulated a total of 2589 music tracks from YouTube and their textual metadata using our search queries. We used the titles of the collected music tracks to identify the most frequently occurring activities. We rely on the title of a track because it appeared to be the most informative about the music-type of the track. For example, the title "Workout Music - Best Workout Rock Music 2016 for GYM and Fitness" indicates that this track can be used while working out in the gym and it contains rock music released in the year 2016. We preprocess the titles by changing them into lower case, converting the -ing forms to their root words (e.g., studying is changed into study) and removing unicode characters, the standard English stop words, genre-related keywords, and numbers (e.g., years).

After this pre-processing, we counted the most frequently occurring terms in the titles, and arrived at the following top-3 activity-related keywords: "relax", "study" and "workout". Note that these keywords can be seen more as activity categories rather than single activities. Examples of single activities, e.g., for study music, include keywords like "work" and "office". Similarly, we find keywords like "run" and "exercise" in the titles of workout music. Our response to RQ1, is the top-3 activity categories, which we will focus on in the remainder of this paper. To provide an impression of these categories, we provide a list of associated keywords:

- *Relax*: relax, calm, soothe, peaceful, chill, meditation, stress relief, sleep
- *Study*: study, focus, concentration, office, work
- *Workout*: workout, training, exercise, gym, run

Our final dataset contains a total of 1272 (49%) *Relax* tracks, 567 (22%) *Study* tracks and 450 (17%) *Workout* tracks. The remaining 300 (12%) tracks were found not to belong to any of the above three categories of activity-related music, despite the presence of the relevant keywords. Although this set of tracks is not used as a classification target, we keep it as *Others* and use it for analysis later in this paper. In order to check for bias towards a particular Internet source, we also inspected the names of the channels from which the tracks were collected. We observed that the *Relax*, *Study* and *Workout* tracks were collected from 12, 10 and 9 different channels respectively, which gives a reasonable diversity of sources.

As a supplement to the keyword information above, Figures 1 - 3 show word clouds, which visualize the term clusters corresponding to our activity categories, which were generated using the titles of the tracks, as described above. Common stop words, numbers, urls have been removed and stemming has been applied. The word clouds allow us to observe the difference in terms that charaterize each of the three main activity categories we found. The word cloud for relaxation music contains keywords like "relax", "calm", "sleep", "heal", "meditate", "calm", "zen", "relief" and "lullaby". Similarly, "workout", "training", "gym", "fit" and "running" are the most frequently used keywords in the titles of workout music (Figure 3). Additionally, we observe the word cloud for the tracks not belonging to any of the above three categories labelled as "Others" in Figure 4. Observing Figure 4, we can say that there is a lot of music for babies, playing games, pets (Dogs) etc.

4 IS GENRE OR INSTRUMENT INFORMATION ENOUGH?

In this section, we address RQ2 and investigate whether genre or instrument information is helpful for predicting music for the top-3 activity categories identified in the previous section. For this investigation we do not develop nor implement any existing genre or instrument detection method. Rather we rely on the textual metadata carrying information about the music genre or instruments present in the titles and descriptions of the music tracks we crawled from YouTube. Our hypothesis is that if the link between the genre- or instrument-related textual metadata and a particular activity category is unambiguous, then it is meaningful to focus on the development, implementation and optimization of the corresponding content-based methods and algorithms as the means to solve the activity-related music classification problem.

The next question to answer is whether the specific genre- and instrument-related terms found in the term clusters are also distinctive per activity category. In order to answer this question, we pick the genre- and instrument-related keywords from the titles of tracks in each of the four term clusters and arrange them in Table 1. Please note that there is no particular order in which the genres or instruments are laid out in the table. Since the dataset contains both electronic and acoustic music, we list the instruments found only in acoustic music. Observing the table, we can say that investigating genre or instrument is not enough to associate music to activities. We can see that genres like classical music, electronic music and ambient music are present in three of the four clusters. In particular, house music is present in all the four clusters, thus also in the *Others* cluster consisting of the tracks for a wide range of activities other than the three targeted in this paper. Similarly, piano, guitar and violin are present in three of the four clusters.

We now take a look at example music tracks for genres and instruments that are common between different activity categories. First, we compare genres in *Relax* and *Study* categories and pick one of the common genres present in both the categories: *Trance*.

[1] https://www.englishclub.com/grammar/prepositions-list.htm
[2] http://www.english-grammar-revolution.com/list-of-conjunctions.html
[3] https://www.youtube.com/user/StudyMusicProject
[4] https://www.youtube.com/user/WorkoutMusicService

Figure 1: Cloud from *Relax* track titles

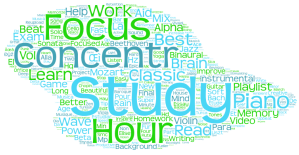

Figure 2: Cloud from *Study* track titles

Figure 3: Cloud from *Workout* track titles

Figure 4: Cloud from *Other* track titles

Listening to the examples of *Trance* music in *Relax* and *Study*, one can immediately identify a difference in texture where a *Study* music example has a slightly higher density than a *Relax* music example. Another difference is the presence of drop-like events [19] in a *Study* music example and a complete absence of such events in the *Relax* example. We refer to drop-like events as those that follow similar acoustic and rhythm patterns as drop events that are typically associated with electronic dance music (EDM). Drop events generally occur as combinations of two different events, viz. *drop* and *build*, defined as follows:

(1) Drop: A point in the EDM track, where the full bassline is re-introduced and generally follows a recognizable build section.
(2) Build: A section in the EDM track, where the intensity continuously increases and generally climaxes towards a drop.

These events are associated with the increasing intensity of the music and reaching a climax before the beat returns. We investigate these events because they represent the building up of intensity and changing of rhythm, which could be important for music for study (to not let the listener zone out) or workout (to push the listener to intensify the workout). Finally, the *Study* music examples were completely devoid of vocals, unlike the *Relax* music examples that have vocals at certain points in the track. This analysis revealed that even within one genre (*Trance*), some musical properties could make one track of that genre suitable for the *Relax* and the other track for the *Study* category.

Music for *Relax* is bound to be very different from the music for *Workout* as they are activities at the extremes of physical exertion with *Relax* requiring least physical activity and *Workout* requiring high level of physical activity. Even though we expect the two categories to be related to completely different music, they still have some genres in common, like *dubstep* and *hip hop*. Listening to an example for both the *Relax* and *Workout* category in the *dubstep* genre, one can clearly observe a difference in terms of tempo and texture. As expected, the tempo is higher and the music is more dense for the *Workout* example as compared to the *Relax* example. Listening to the *Workout* example, one can observe the prominence of the bassline, which is at times "naked" without melodic layers. In contrast, in the *Relax* example the bass is much less prominent. As indicated earlier, *Relax* music does not contain drops but there are many drops in the *Workout* example. Our conclusion here is therefore the same as above. Though the tracks belong to the same genre, there are significant variations that make it challenging to rely on genre information alone to distinguish between music for *Relax* and *Workout*.

Finally, we look at two examples from the same genre (*progressive house*), but from different activity categories: *Study* and *Workout*, and notice the presence of vocals in the *Workout* track. Another key difference is the presence of many drop-like events in the *Workout* track and limited number of such events in the *Study* track.

Based on the analysis reported above, we can conclude that genre- and instrument-related information alone is not sufficient to predict suitability of a music track for an activity. Observing the individual examples, we see the main reason for this is the local properties of a music track, i.e., localized variations in low-level features. In the subsequent section, we therefore propose a method which segments the tracks into windows of different time resolutions in order to investigate how to optimally capture these local variations for the posed classification task. Furthermore, the insights presented above motivate our decision to consider the presence/absence of drop-like events as one of the features in the design of activity-related music classification framework, as stated in Section 1.

5 HOW TO IDENTIFY MUSIC FOR ACTIVITY CATEGORIES

In this section, we describe our approach to developing an automatic classification method for activity-to-music mapping. Since the information on genre or instruments is not helpful in detecting music for a given activity, classification based on other and more relevant information needs to be developed.

We start off by noting that recent advances in deep learning, such as [13], may enable unsupervised extraction of relevant features. However, we would like the features that we identify as contributing towards identifying music for activities to be explainable, and we would also like to carry out an assessment of the temporal resolution that is appropriate for feature extraction. Explainability of deep learning pipelines for music currently still is in a pioneering phase [16]. For these reasons, we choose to investigate features and models that are already well understood and reflect different musical characteristics. More specifically, we take as input a basic set of low- and mid-level (rhythm and tonality related) features known from the MIR field. These features and their corresponding dimensionality (in parenthesis) are listed in Table 2.

In addition to the problem of understanding what features are suitable for the task in the first place, the main open issue related to how the features are extracted is the selection of the time window t to optimally aggregate the feature values in order to capture the above mentioned informative local signal variations in the best possible way. In order to discover the best value for t, we devise an algorithm that we run on our training data set and that uses repeated random sub-sampling validation [7] to test different values

Type of music	Genres	Instruments
Relax	classical, binaural, jazz, ambient, house, trance, dubstep, chillstep, hip hop, trap, rock, country, folk	piano, guitar, flute, saxophone, violin, drums
Study	classical, binaural, jazz, ambient, trance, Drum & bass, electronic, deep/electro/progressive house	piano, violin, guitar, viola
Workout	electro, dubstep, progressive house, rock, rap, EDM, hip hop, electronic, techno	piano, drum
Others	jazz, classical, binaural, house, ambient, rock, folk, dubstep, EDM, electronic, electro, trance	piano, saxophone, guitar, drums

Table 1: Genres and musical instruments present in each activity category.

Type of features	Features
Low-level features	Avg. loudness (1), dynamic complexity (1), pitch salience (2), spectral centroid (2), spectral complexity (2), spectral decrease (2), spectral energy (2), spectral entropy (2), spectral flux (2), spectral kurtosis (2), spectral rms (2), spectral rolloff (2), spectral skewness (2), spectral spread (2), zero crossing rate (2), MFCC (13), Image moments (162) [19]
Rhythm features	Number of beats (1), tempo (1), danceability (1), onset rate (1), statistical spectrum descriptor (168) [14], rhythm histogram (60) [14]
Tonal features	chromagram (12), key strength (1), pitch class profile (pcp) (36)
High-level features	Number of events (1) [19], Affect (3)

Table 2: Low-level/mid-level/high-level features we use for distinguishing between music for different activities.

of t. In the same algorithm, we also embed the search for the best value of another parameter d, which stands for the number of most discriminative features used for classification. We use a simple k-nearest neighbour (k-NN) classifier and repeat the algorithm to identify the combination of d and t that gives the best classification performance. The proposed algorithm is defined as follows:

(1) Consider the range of $t = 0.5, 1, 5, 10, 15, 20, 25, 30$ seconds
(2) For each value of t, follow these steps:
  (a) Extract features for each segment and combine them into a single feature vector.
  (b) Randomly divide the training data into X_{train} (90%) and X_{val} (10%).
  (c) Select a value of d from the set 10, 11, 12 ... to 50 features.
  (d) Use X_{train} for feature selection and pick the top-d most discriminating features. Before feature selection, we normalize each feature.
  (e) Use X_{train} with selected features to build a training model.
  (f) Use this model to predict labels in X_{val}.
  (g) Aggregate the segment-level labels using a majority vote to obtain a single label for a track and then compute the f-score.
  (h) Repeat steps 2 (d) – (g) for the whole range of d.
  (i) Repeat the whole process ten times for different X_{train} and X_{val} each time to obtain average validation performance.
(3) Choose the t with the best average validation performance.

In this paper, we aim to understand the phenomena underlying the activity-related music classification and *not* to optimize the classification itself. This is the reason for which we chose a simple and standard k-NN classifier, which has minimal number of parameters to be tuned. Regarding the range we considered for t, we also investigated the window sizes beyond 30 seconds (up to 60 sec) and found that the performance does not improve. For feature selection, we use a method that deploys mutual information and that is available in the feature selection toolbox [4]. Once we identify the best values of d and t, we evaluate the performance on the test set to predict the links between the music tracks and the activity categories.

The other features we introduce are based on intuition, informed by the analysis in Section 4. Here, we consider three affect dimensions, namely arousal, valence and tension, and assess their impact to activity-related music classification experimentally, using conventional scores [12]. We do the same with the feature encoding the number of drop-like events [19] found in a music track: while the consideration of this feature initially was also based on intuition, the potential of this feature has been strengthened by the analysis reported in the previous section. The affect scores are extracted over non-overlapping segments of duration t seconds (result of the algorithm described above) and for the events feature, we count the number of drop-like events in the entire music track.

6 EXPERIMENTAL EVALUATION

In this section, we evaluate our proposed method and also compare its performance to a number of existing methods and approaches that we found related to the problem addressed in this paper. Additionally, we report which features are the most discriminative for our classification task. Finally, we summarize the insights we gained from a failure analysis, based on which we propose topics for future research in this direction.

Window size(sec.)	F-score(relax)	F-score(study)	F-score(workout)	Overall F-score	No. of features
0.5	0.41	0.23	0.56	0.4	46
1	0.43	0.24	0.55	0.41	45
5	0.41	0.28	0.57	0.41	48
10	0.59	0.58	0.79	0.65	40
15	0.59	0.63	0.86	0.69	38
20	0.58	0.61	0.83	0.67	26
25	**0.60**	**0.69**	**0.89**	**0.73**	**25**
30	0.60	0.59	0.89	0.69	35

Table 3: F-scores for different classes and number of discriminative features across different window sizes.

Feature	F-score(relax)	F-score(study)	F-score(workout)	Overall F-score
Events (E)	0.79	0.64	0.72	0.72
Affect (A)	0.63	0.51	0.73	0.62
A + E	0.65	0.51	0.71	0.62

Table 4: F-scores using high-level features at a windows size of 25 sec.

6.1 Experimental design and results

For the experiment, we have a training set, containing 300 relax tracks, 250 study tracks and 250 workout tracks. For the test set, we use 50 each of relax, study and workout tracks. We focus on tracks that are specific for a single activity category, as reflected in our labels. Music suitable for multiple categories is an interesting topic for future work, but we do not look at it here.

For finding t and d according to the algorithm described in the previous section, we set k = 10 for the k-NN classifier after evaluating different values of k on a smaller development set (not included in the train and test set). For extracting low-level, tonal features and most of the rhythm related features, we use the Essentia framework [2]. For statistical spectrum descriptors and rhythm histogram, we use the source code provided by Lidy et al. [14]. For extracting the low-level features, we use a non-overlapping frame size of 100 ms. Regarding the high-level features, we use the method proposed in [19] to detect the drop-like events in a given music track. We rely on the dataset released by Yadati et al. [11] to train models and predict the presence of events in our dataset. Finally, for computing the affect scores, we use the MIRtoolbox [12] that gives us a 3-dimensional feature vector with one score per dimension.

Table 3 shows the f-scores per activity category obtained while executing the algorithm for optimizing the values of t and d, as introduced in the previous section. We note again that the classification here is performed using the low- and mid-level features only. It can be observed that the best classification performance was obtained at a window size of 25 seconds. Examining the f-scores obtained at this window size, we can say that the simple classifier performs reasonably well in distinguishing between music for the three different categories.

As indicated in the last column, for the window size of 25 seconds, the best number d of discriminative features to use is 25. Here, we list the features (and their dimensionality) that are found to be most discriminative in this case: tempo (1), dynamic complexity (1), danceability (1), onset rate (1), spectral centroid (1), spectral flux (1), image moments (6), PCP (4), rhythm pattern (4), rhythm histogram (3) and MFCC (2). This is a mix of rhythm features, low-level features and tonal features, with a majority of them being rhythm-related and with PCP being the only representative of the tonal features. A key observation here is that most of the selected features (tempo, danceability, rhythm pattern etc.) generally need longer time segments to be computed. We therefore believe that the flexibility we allowed in the selection of the time window t was critical for pushing these features forward as being most informative for classification and therefore also critical for getting the most out of the signal and achieving the best possible classification performance.

We also performed the classification based on the high-level features, first separately and then integrated with low- and mid-level features. We computed the affect features in the time interval corresponding to the optimal value of t, namely 25 seconds. Experiments using other window sizes showed, however, that this parameter is not critical, resulting in relatively constant classification performance. Computation of the event feature is not dependent on the time window as this is solely the number of drop-like events found in a music track. The classification results for different features and their combinations are presented in Table 4. We observe that the high-level features generally perform worse than low- and mid-level features. An interesting exception is the result obtained for the events feature and *Relax* category. The detector of the drop-like events that we adopted from [19] is namely known for its high precision and low recall. This is beneficial for the *Relax* music tracks having no drops and less beneficial for the tracks from other two categories where drop-like events are present, but because of the detector deployed, not well detectable. This result shows the potential of this feature to improve the overall classification performance upon the one obtained by using low- and mid-level features, however, under the condition that the detector of drop-like events performs well. We discuss this further in the next section.

So far, we looked at the performance of our method in isolation. We now compare our method with existing methods which classify music tracks into activities. Specifically, we compare it with the

method proposed by Wang et al. [18] and also with two other methods that we devised as being representative of common approaches deployed in standard MIR classification tasks. The four methods entering a comparative analysis are:

(1) *Our method*: As a representative of our proposed approach we choose the method variant deploying low- and mid-level features with the best performance in Table 3, namely for the time window of 25 seconds and 25 features.
(2) *Full track*: We aggregate the low-level features, extracted from 50ms frames, over the entire track by computing the mean and variance. We then combine these features with other rhythm and tonal features extracted from the whole track. We perform feature selection and select the most discriminative features (51 in this case). Using a k-nearest neighbour classifier, we predict the labels of the music tracks in the test set and compute the f-scores for the three categories. Such a method is inspired from the field of static emotion recognition [8], which aggregates the features over the entire music track in order to give an affect score for a track.
(3) *One segment*: We select one 25-second segment from each track in the training data and extract the features as before. We then perform a feature selection and obtain the most discriminative features (49 in this case). We divide each music track in the test set into 25-second segments and select these 49 most discriminative features. Using a k-nearest neighbour classifier, we predict the labels of each 25-second segment in the test set and use a majority vote to get a single label for a track. We then compute the f-scores for the three categories. This method is inspired by existing MIR approaches, especially genre recognition [8], where a short segment taken from the track is used for feature extraction and classification.
(4) *Wang et al.*: Wang et al. extract features from a 512-sample frame every 30 seconds and compute the mean and variance of these features over the entire track. Then, they use an adaboost classifier to predict the labels of music tracks in the test set. We follow this procedure on our dataset and compute the f-scores for the individual categories.

Figure 5 summarizes the results of all four methods for the three target activity categories. We can clearly see that our method outperforms other methods. We further observe that the *Full track* method that aggregates features over the entire track performs better than the *One segment* method and the *Wang et al.* method at least in two categories: *Relax* and *Study*. From Table 3 and Figure 5, we can conclude that aggregating over a longer window size helps in classifying the music track into one of the three activity categories and, based on the experiments on our dataset, the best window size is 25 seconds.

6.2 Failure analysis and outlook

Through different experiments, we have shown that we can distinguish between music for different activities and that our method performs better than the related existing methods. However, even the best results are not perfect. In this section, we focus at these

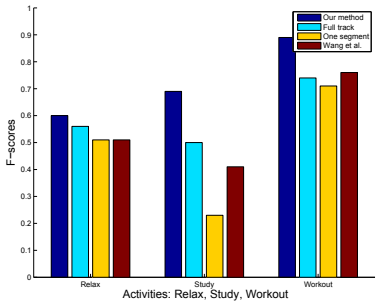

Figure 5: Comparison of performance for three activity categories across different methods.

imperfections by analyzing the failure cases where music tracks belonging to a certain activity category are assigned a wrong label.

Figure 6 illustrates the confusion matrix for the predicted labels on the test set and the numbers in the boxes indicate the number of correctly/wrongly classified samples. The first observation we make is that there is considerable confusion between the *Relax* and *Study* categories. Here, we take a look at individual examples and try to find patterns that lead to this confusion between *Relax* and *Study*. Using a majority vote to aggregate the labels seems to be the reason behind some of the misclassifications. One of the tracks in the relax category, which had 72 segments in total, is misclassified as a *Study* track because 33 segments are classified as *Relax*, 34 segments are classified as *Study* and the remaining segments are classified as *Workout*. A majority vote clearly finds that the track is *Study*, but the competition between *Relax* and *Study* categories was close. We also found many other examples where the difference between the number of segments classified as *Relax* and those classified as *Study* is low. There was even an example in the *Study* category that contained equal number of segments classified as *Relax* and *Study*, but the max operator chose the category of the track as *Relax*. In order to investigate this phenomenon further, we measured the mean and standard deviation of the difference between the number of segments in the top two categories for each example. For correctly classified examples, the mean is 48.6 while the standard deviation is 16.2. For incorrectly classified examples, the mean is 26 while the standard deviation is 6.1. We clearly see lower values for incorrectly classified examples, indicating that there is a closer competition between categories for incorrectly classified examples. Clearly, deploying majority vote has drawbacks as it does not reflect how strong the majority is. This calls for investigating different aggregation strategies that can combine the labels of individual segments into a single label for the track in a more robust fashion. Alternatively, we would also like to explore whether we could choose the segments in a smart way (analogous to feature selection) that are most discriminative for an activity category, which removes the need for an aggregation step. Another possible direction could be to consider the labels for the segments as a sequence instead of considering them as a bag of segments (current method). The temporal ordering could provide additional information that could reduce the misclassfication rate.

We initially hypothesized that drop-like events are completely absent in *Relax* music, while they are present in the other two

categories. This is confirmed by the classification results reported in Table 4, in particular by a high f-score for *Relax* music when events feature is used. As explained in the previous section, this effect is additionally emphasized due to a strong bias of the event detector used towards high precision. However, there is more to it. Some of the *Study* music tracks also do not contain drop-like events and this resulted in a confusion between *Relax* and *Study* categories. Furthermore, there are similar numbers of drop-like events in some *Study* music and *Workout* music tracks, which results in lower f-scores for these two categories. Another reason for failure is that there are more subtle drop-like events in *Study* music while *Workout* music has more pronounced events and the drop detector missed detecting some of the subtle events.

Mapping between low-level features and affect is a difficult proposition and we have used an off-the-shelf toolbox to compute the affect scores for the music tracks. Observing the results reported in Table 4, the affect based classifier performs reasonably well, but there is a scope for improvement. We could look at different strategies to compute affect scores in the future and investigate its impact on the classification performance.

An aspect of activity-based music that needs further attention is the presence of distractors, which are musical characteristics that might distract the user from his/her activity. For instance, one of the observations in Section 4 was that *Study* music did not have any vocals while the other two types of music could contain vocals. In the future, one could investigate to which extent the presence/absence of vocals is informative as a feature for this classification task. In general, one could search for additional sources of information, e.g., user comments, that can help identify the distractors for different activities. Here are some examples of user comments that can be used to identify if the track is really useful for an activity:

- Comment on a relax music track: "There is a jarring piano sound in the middle!"
- Comment on a study music track: "This track contains vocals and distracting while working"

The biggest challenge we see when relying on user comments is to spot the comment with the relevant information among plenty of (largely noisy) comments posted by the users.

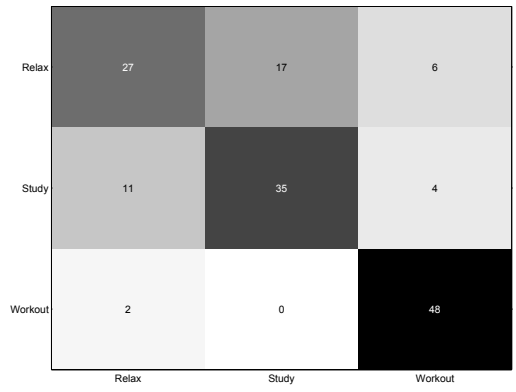

Figure 6: Confusion matrix for the best performing case of t = 25 seconds

7 CONCLUSION AND OUTLOOK

In this paper, we have addressed the challenge of identifying appropriate music for common daily activities. In this way, we made a critical step towards developing a music recommender system that takes into consideration both the aspects of what music is and what it can do for a listener. We have focused on the three activity categories that we found to be be common via a study of textual metadata on YouTube: *Relax*, *Study* and *Workout*.

One of our key findings is that standard music information retrieval (MIR) approaches, in particular, those focused on genre or instrument information, are not well-suited for addressing the problem of activity-based music classification. Another important finding is that this task requires more timeline information (25 seconds) for feature extraction from an audio track, i.e., the window size must be longer than what is currently conventional in the MIR literature.

Based on these findings we have developed a method that identifies the time resolution at which the low-level features should be aggregated and also the best number of discriminative features to be used. Using the features extracted at the identified temporal resolution, our classifier could successfully distinguish between music for the three different activity categories and also outperform existing methods.

This paper opens interesting perspectives for future work. From the musical content perspective, we plan to investigate additional information to improve the identification of music for activities. Here, we have taken a bag-of-segments approach. Moving forward, however, we anticipated that incorporation of the temporal order of the segments could, as mentioned above, provide further insight. Further, also as mentioned above, users post comments on YouTube for different music tracks. Some of these touch on the suitability of a music track is for a specific activity. These comments are a promising source of information. Additionally, high-level features, e.g., presence/absence of vocals, could also improve classification.

Our work here is based on the insight that there are general characteristics of music which have a similar reception across a broad population. In pursuit of these general characteristics, we focus on information about music tracks provided by uploaders. We adopt an assumption used recently in work on video uploader intent [10]: the fact that uploaders are publishing on a public platform, accessible to millions of users, makes it likely that they are taking the musical reception of the general population into account. The fact that we focus on here on broad consensus on which music is appropriate for which purposes, should not preclude future study of the role played by individual preferences in users' choices of music for different activities. Individual preferences should also be understood as preferences of groups of users who pattern together, such as introverts and extroverts, as studied by [17]. Moving forward, understanding where universal music preferences fall short of being useful will allow us to gain further insight into the performance of the classifier. Specifically, we would like to investigate the relatively large confusion between the music for the *Relax* and *Study* categories from a user's perspective. Such a user study would allow us to determine whether the classifier should be further improved, or whether the category labels must be refined to make it possible to cater for finer-grained preferences within the population.

REFERENCES

[1] Anna Aljanaki, Frans Wiering, and Remco C Veltkamp. 2015. Emotion based segmentation of musical audio. In *Proceedings of the 17th Conference of the International Society for Music Information Retrieval (ISMIR 2015)*.

[2] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, O. Mayor, Gerard Roma, Justin Salamon, J. R. Zapata, and Xavier Serra. 2013. ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2013)*.

[3] Sara Bottiroli, Alessia Rosi, Riccardo Russo, Tomaso Vecchi, and Elena Cavallini. 2014. The cognitive effects of listening to background music on older adults: processing speed improves with upbeat music, while memory seems to benefit from both upbeat and downbeat music. *Frontiers in Aging Neuroscience* 6 (2014), 284.

[4] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* 13 (Jan. 2012), 27–66.

[5] Andrew Demetriou, Martha Larson, and Cynthis C. S. Liem. 2016. Go With the Flow: When Listeners use Music as Technology. In *Proceedings of the 18th Conference of the International Society for Music Information Retrieval (ISMIR 2016)*.

[6] Ricardo Dias, Manuel J Fonseca, and Ricardo Cunha. 2014. A User-centered Music Recommendation Approach for Daily Activities.. In *Proceedings of the 1st Workshop on New Trends in Content-based Recommender Systems (CBRecSys)*.

[7] Werner Dubitzky. 2009. *Fundamentals of Data Mining in Genomics and Proteomics*. Springer-Verlag, Berlin, Heidelberg.

[8] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. 2011. A Survey of Audio-Based Music Classification and Annotation. *IEEE Transactions on Multimedia* 13, 2 (April 2011), 303–319.

[9] Alan Hanjalic, Christoph Kofler, and Martha Larson. 2012. Intent and Its Discontents: The User at the Wheel of the Online Video Search Engine. In *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*.

[10] C. Kofler, S. Bhattacharya, M. Larson, T. Chen, A. Hanjalic, and S. F. Chang. 2015. Uploader Intent for Online Video: Typology, Inference, and Applications. *IEEE Transactions on Multimedia* 17, 8 (Aug 2015), 1200–1212.

[11] Martha Larson, Karthik Yadati, Mohammad Soleymani, and Pavala Shakthi Nathan Chandrasekaran Ayyanathan. 2014. MediaEval 2014 Crowdsourcing Task: Crowdsorting multimedia comments. (Oct 2014). http://osf.io/h92g8

[12] Olivier Lartillot and Petri Toiviainen. 2007. MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio. In *Proceedings of the 9th International Society for Music Information Retrieval Conference (ISMIR 2007)*.

[13] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y. Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems (NIPS 2009)*.

[14] Thomas Lidy and Andreas Rauber. 2005. Evaluation of Feature Extractors and Psycho-acoustic Transformations for Music Genre Classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*.

[15] Lie Lu, D. Liu, and Hong-Jiang Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1 (Jan 2006), 5–18.

[16] J. Pons, T. Lidy, and X. Serra. 2016. Experimenting with musically motivated convolutional neural networks. In *Proceedings of the 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*.

[17] Geen Russell. 1984. Preferred stimulation levels in introverts and extroverts: Effects on arousal and performance. *Journal of Personality and Social Psychology* 46, 6 (June 1984), 1303–1312.

[18] Xinxi Wang, David Rosenblum, and Ye Wang. 2012. Context-aware Mobile Music Recommendation for Daily Activities. In *Proceedings of the 20th ACM International Conference on Multimedia (MM 2012)*.

[19] Karthik Yadati, Martha Larson, Cynthia C. S. Liem, and Alan Hanjalic. 2014. Detecting Drops in Electronic Dance Music: Content based approaches to a socially significant music event. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*.