# The *Musical Avatar* - A visualization of musical preferences by means of audio content description

Martín Haro[1] Anna Xambó[1,2] Ferdinand Fuhrmann[1] Dmitry Bogdanov[1]
Emilia Gómez[1] Perfecto Herrera[1]
[1]Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[2]Music Computing Lab, The Open University, Milton Keynes, UK
{martin.haro, ferdinand.fuhrmann, dmitry.bogdanov,
emilia.gomez, perfecto.herrera}@upf.edu, a.xambo@open.ac.uk

## ABSTRACT

The music we like (i.e. our musical preferences) encodes and communicates key information about ourselves. Depicting such preferences in a condensed and easily understandable way is very appealing, especially considering the current trends in social network communication. In this paper we propose a method to automatically generate, given a provided set of preferred music tracks, an iconic representation of a user's musical preferences – the *Musical Avatar*. Starting from the raw audio signal we first compute over 60 low-level audio features. Then, by applying pattern recognition methods, we infer a set of semantic descriptors for each track in the collection. Next, we summarize these track-level semantic descriptors, obtaining a user profile. Finally, we map this collection-wise description to the visual domain by creating a humanoid cartoony character that represents the user's musical preferences. We performed a proof-of-concept evaluation of the proposed method on 11 subjects with promising results. The analysis of the users' evaluations shows a clear preference for avatars generated by the proposed semantic descriptors over avatars derived from neutral or randomly generated values. We also found a general agreement on the representativeness of the users' musical preferences via the proposed visualization strategy.

## Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Audio input/output*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Methodologies and techniques, Modeling, Systems*

## General Terms

Design, Algorithms, Human Factors

*AM'10*, September 15–17, 2010, Piteå, Sweden.

## Keywords

Music information research (MIR), user modeling, audio content analysis, semantic retrieval, music visualization

## 1. INTRODUCTION

Music always played an exceptional role in peoples' life. It allows us to directly influence our physical and mental condition, relate and connect memories, and express our emotions in a creative way. In addition, we also use it to communicate information about ourselves: there is evidence that the music we like (i.e. our musical preferences) is linked to our personalities and serves as a representative identity in social networks [23]. Similarly, North and Hargreaves indicate that music can operate as a "badge" which guides peoples' social cognition [19]. In a later study, the authors also reveal correlations between musical and general lifestyle preferences, e.g. interpersonal relationships or living arrangements [20]. Thereby, humans utilize music to establish friendships and form communities [15]. Moreover, musical preferences are found to be the most common topic in conversational situations, where people get to know each other [24]. All this suggests that our musical preferences reflect both individual emotions and attitudes towards society, politics or – even more abstract – life in general.

In the course of the evolution of the World Wide Web, new kinds of social networks became popular. People use online communities to share opinions, impressions and experience about general life (e.g. Facebook) or more specific topics (e.g. Last.fm in the case of music). Here, communication is done via text fragments, pictures, short videos or audio/music. As highly informational features like physical appearance, nonverbal behavior, facial features or clothing styles are not available in the usual interaction situation, the individual representation inside the community (i.e. the user's profile) replaces the aforementioned cues for identity claims. Therefore, users try to compress self-contained, meaningful, funny, and personally representative information into this compact representation to serve as first impression of the user in the network (e.g. consider the importance of the user's photo).

In its interdisciplinary character, music information research (MIR) combines, among others, music theory, music cognition, signal processing, computer science, and mathematics to understand and explain the connections between humans and music. Formally, it models these connections, and exploits the derived models for enhancing the interaction of people with music. Recent research has put more effort into obtaining semantically meaningful descriptions of the musical content given that the low-level information extracted from the raw audio signal was found to be insufficient for describing the concepts humans usually associate with music [5, 2]. Additionally, current technology allows to aim for complete systems, which can provide better means to assess usability and create social impact to enforce further advancements in research [7].

User modeling for music applications has generated some research in MIR in the past years [6, 8, 26], but translating these models into visual counterparts has remained, to our knowledge, unexplored. In this work we present a method which maps the musical preferences of a given user to the visual domain. Given a collection of music tracks, provided by the user as a prototypical example of his/her musical preferences, we extract descriptive information from the audio of each of these files, and design a mapping algorithm that combines these data into a graphical representation – the *Musical Avatar*. More precisely, the proposed method is using audio content analysis to generate high-level semantic descriptions for a single music track, which can be regarded as a point in a multi-dimensional semantic descriptor[1] space. From the collection of all points – each corresponding to one music track – several statistical measures are derived to obtain a summarization of the user's musical preferences within this space. This summarized description is then mapped to the visual attributes of the *Musical Avatar*. We finally ask the user to evaluate the accuracy and usefulness of the extracted descriptors and the selected mapping strategies in order to draw conclusions about the proposed method and detect future research directions.

A *Musical Avatar* is thought to be used within online communities or social networks (Facebook, Last.fm, Messenger, Skype, Twitter – to name just a few) as an individual representation based on the user's musical preferences. In a similar way, avatars derived from the proposed method could also label different playlists from the same user, i.e. several *Musical Avatars* can be assigned to different and musically dissimilar subsets within a single music collection, thus indicating their distinct musical facets as an easily recognizable icon. Finally, as the applied music content processing system outputs a set of points within a multi-dimensional semantic space, it is possible to retrieve similar users and/or relevant musical recommendations based on the user's musical preferences.

What follows is a short overview of the content of this article: In Section 2 we present the proposed methodology. Within this section we describe the acquisition of the user data used in this research. Furthermore, we report the audio content analysis system and its specific features, applied to obtain semantic descriptors, and the resulting compact representation of a set of music tracks. We then explain the mapping from the acoustic-semantic domain to the visual domain, including the proposed graphical design. In Section 3, we describe the user evaluation of the generated avatars and discuss the obtained results. Finally, in Section 4 we conclude this article and highlight several future research directions.

## 2. METHODOLOGY

In this section we will explain the steps followed to generate a *Musical Avatar* for a single user. Figure 1 shows the block diagram of the proposed methodology.

### 2.1 User Data Gathering

As a first step, we ask the user to gather the minimal set of music tracks sufficient to grasp or convey her/his music preferences. It is important to note that these are not artist names but single music tracks which are informative by themselves (without any additional context). The user then provides either a folder with the selected tracks in audio format (e.g. mp3) or the needed information to unambiguously identify and retrieve each track (i.e. artist, name of the piece, edition, etc.). We also ask the user to provide some additional information, including personal data (gender, age, interest for music, musical background), a description of the strategy followed to select the music pieces, and the way they would describe their musical preferences. This information will help us for further analysis and to evaluate the provided avatars.

### 2.2 Descriptor Extraction

We now describe the procedure of obtaining a semantic representation of the user's preferences within the used audio content analysis system. For each music track, we calculate a low-level feature representation using an in-house audio analysis tool[2]. In total it provides over 60 commonly used low-level audio features, characterizing global properties of the given tracks. They include inharmonicity, odd-to-even harmonic energy ratio, tristimuli, spectral centroid, spread, skewness, kurtosis, decrease, flatness, crest, and roll-off factors, Mel frequency cepstral coefficients (MFCCs), spectral energy bands, zero-crossing rate [21], spectral and tonal complexities [27], transposed and untransposed harmonic pitch class profiles, key strength, tuning, chords [10], beats per minute (BPM) and onsets [4]. Most of these features are extracted on a frame-by-frame basis and then summarized by their means and variances across all frames. In the case of multidimensional features (e.g. with MFCCs), covariances between components are also considered.

Having computed low-level features for each track, we then follow the procedure presented in [3] to infer semantic descriptors. We perform a regression by suitably trained classifiers producing different semantic dimensions such as genre, culture and moods. We use standard multi-class support vector machines (SVMs) [28], which have been shown to be an effective tool for various classification tasks in MIR [14, 17, 30]. More concretely, we employ 14 ground truth music collections (including full tracks and excerpts) and execute 14 classification tasks corresponding to these data. For some descriptors we use already existing collections in the MIR field, while for others we use manually labeled in-house datasets (for more detailed information regarding the used collections see [3] and references therein). The regression results form a high-level descriptor space, which contains the probability estimates for each class for each SVM classifier. We use the libSVM implementation with the C-SVC method and a radial basis function kernel with default parameters[3] after a correlation-based feature selection (CFS) [11] over all

[1] In the remainder of the paper we will pragmatically refer to a high-level semantic descriptor with the generic term "descriptor".

[2] http://mtg.upf.edu/technologies/essentia

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

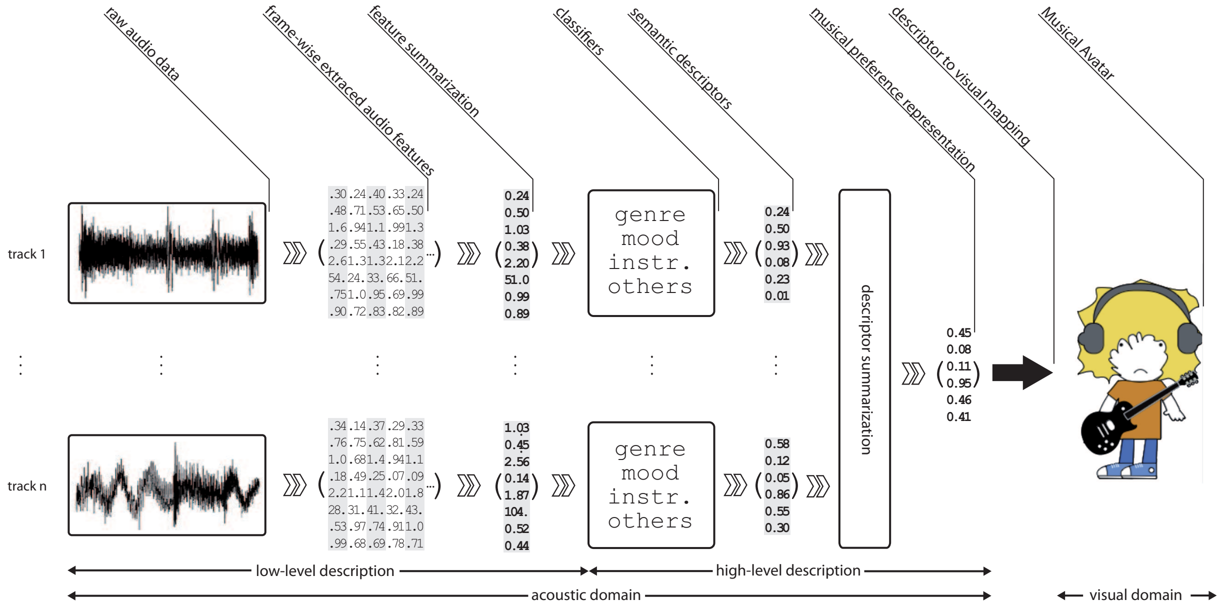

**Figure 1: Block-diagram of the proposed methodology. The user's music tracks are analyzed and represented with low-level audio features which are later transformed, by means of classifiers, into semantic descriptors. The individual track descriptors are furthermore summarized into the user profile which is finally mapped to a series of graphical features of the *Musical Avatar*.**

**Table 1: Selected descriptors per semantic dimensions (i.e. genre, moods & instrumentation and others).**

| Genre | Moods & Inst. | Others |
|---|---|---|
| Classical | Happy | Party |
| Jazz | Sad | Vocal |
| Metal | Aggressive | Tonal |
| Dance | Relaxed | Bright |
| Rock | Electronic | Danceable |
| Electronic | Acoustic | |

$[0, 1]$-normalized low-level features.

With the described procedure we obtain 77 descriptors, including categories of genre, culture, moods, instruments, rhythm and tempo. Although such a rich semantic description can provide valuable cues for visualization, in this prototypical study we opt for simplicity, and thus reduce the amount of the semantic descriptors. Therefore, we perform an additional filtering considering classifiers' accuracy as a criterion while preserving the representativeness of the semantic space. To this extent, we ask a subset of users to manually annotate their own music collections, filling in the same descriptors as those inferred by the classifiers. We then compare the manual annotations with the classifiers' outputs using Pearson correlation, and opt for the best performing descriptors. The resulting 17 descriptors, which are used for constructing the *Musical Avatar*, are presented in Table 1.

## 2.3 Descriptor Summarization

Having computed the selected descriptors for all the tracks in the user's collection we apply different strategies to obtain a compact representation which can be mapped to the visual domain. To remove global scaling and spread, we first standardize each descriptor (i.e. subtracting the global mean and dividing by the global standard deviation). We estimate the reference means ($\mu_{R,i}$) and standard deviations ($\sigma_{R,i}$) for each decriptor i from a representative music collection containing more than 100,000 tracks including all common Western musical genres.

Moreover, as the visualization process (see Section 2.4) takes normalized values as inputs (i.e. values between 0 and 1) we range-normalize the aforementioned standardized descriptor values according to the following equation:

$$N_i = \frac{d_i - min}{max - min}, \tag{1}$$

where $d_i$ is the standardized value of descriptor $i$, and since $d_i$ has zero mean and unit variance, we set the respective $min$ and $max$ values to $-3$ and 3, as according to Chebyshev's inequality at least 89 % of the data lies within 3 standard deviations from its mean [9]. All resulting values smaller than 0 or greater than 1 are clipped. The resulting scale can be seen as a measure of preference for a given category.

We then summarize the descriptor values by computing the mean for every normalized descriptor ($\mu_{N,i}$). At this point, we decide to represent the resulting mean values by quantizing the respective preference using three strategies, namely: *continuous* (i.e. all possible values), *ternary* and *binary*. Each quantization level conveys a different degree of data variability. For the *binary* quantization we force the descriptors to be either 1 or 0, representing only two levels of preference (i.e. 100% or 0%). In the *ternary* case a third value is introduced representing a neutral degree of preference (i.e. 50%). The *continuous* quantization includes all possible degrees of preference. These three types of quantization are computed as follows: In the *continuous* case we maintain the computed $\mu_{N,i}$ values without further changes. In the *binary* case we quantize all $\mu_{N,i}$ values below 0.5 to zero and all values above (or equal) 0.5 to one. In the *ternary* case we perform the quantization directly from the original descriptor values, that is, we calculate the mean for every descriptor ($\mu_i$) and quantize the mean values according to the following criteria:

$$Ternary_i = \begin{cases} 1 & \text{if } \mu_i > (\mu_{R,i} + th_i), \\ 0.5 & \text{if } (\mu_{R,i} - th_i) \leq \mu_i \leq (\mu_{R,i} + th_i), \\ 0 & \text{if } \mu_i < (\mu_{R,i} - th_i), \end{cases} \tag{2}$$

where $th_i = \sigma_{R,i}/3$.

## 2.4 Visualization

The final procedure consists of converting the summarized descriptors to a set of visual features to generate the *Musical Avatar*. We apply the cultural approach of representing urban tribes as described in [16], since in these subcultures music plays a relevant role in both the personal and cultural identities. The subcultures are often identified by specific symbolisms, which can be recognized visually. Thus, our approach is mapping the semantic descriptors into a basic collection of cultural symbols. As a proof-of-concept we select an iconic cartoony style for the following reasons: firstly, it is a less time-consuming technique compared to other approaches more focused on realistic features [1, 22, 25]; secondly, it is a graphical medium which, by eliminating superfluous features, amplifies the remaining features of a personality [18]; and thirdly, there already exist popular avatar collections of this kind such as Meegos[4] or Yahoo Avatars[5].

Though colors have been successfully associated with musical genres [12] or moods [29], cultural differences have been reported when trying to define a global mapping. Instead, in our study the relevant role is played by the graphical symbols, which are filled with arbitrary colors related to them.

In order to decide which and how the urban tribes are realized, we focus on the information provided by the selected descriptors and the design requirements of modularity and autonomy. Starting from a neutral character[6], the body is divided into different parts. For each of these parts we provide a set of graphic symbols. Each of these graphic symbols always refers to the same descriptors (e.g the mouth is always defined by the descriptors of “Moods” and “Others” but never by “Genre” descriptors, see Table 2). Besides, we introduce a label to identify the gender of the avatar, each providing a unique set of graphic symbols. Apart from the body elements, a set of possible backgrounds is also added to the graphic collection in order to support some descriptors of the “Others” category such as *Party*, *Tonal* or *Danceable*. In addition, the *Bright* descriptor value is mapped to a grey background color that ranges from RGB(100,100,100) to RGB(200,200,200). All graphics are done in vector format because of their rescalability. Table 2 shows the relation between graphic groups and semantic dimensions used by the mapping algorithm.

**Table 2: Mapping of the descriptor dimensions to the graphic groups.**

| | Dimensions | | |
|---|---|---|---|
| Graphic Group | Genre | Mood | Others |
| Background | | | • |
| Head | • | • | • |
| Eyes | | • | • |
| Mouth | | • | • |
| Complement | • | • | • |
| Suit | • | | • |
| Hair | • | | |
| Hat | • | • | |
| Complement2 | | | • |
| Instrument | • | • | |

To obtain the user’s musical preferences in terms of graphic elements we construct a vector space model and define the Euclidean distance as a measure of dissimilarity therein. For each graphic symbol we choose the best among the set of all available candidates which is closest to the corresponding subset of the user’s vector model. This subset is defined according to the mapping criteria depicted in Table 2. As a result, a particular *Musical Avatar* is generated for a concrete user’s musical preferences. The prototype implementation is done using Processing[7].

Taking into account the different summarization strategies described in Section 2.3, mapping is done in either a discrete or continuous space resulting in different data interpretations and outputs. These differences imply that in some cases the graphic symbols have to be defined differently: for instance, the *Vocal* descriptor set to 0.5 in the *continuous* scenario case means ”she likes both instrumental and vocal music”, whilst this neutrality is not present in the *binary* scenario. Furthermore, in the *continuous* scenario, properties such as size or chromatic gamma can be exploited while this is not possible within the discrete vector spaces. Figure 2 shows a graphical example of our visualization strategy where, given a user model, the best (i.e. the closest in Euclidean distance) graphic symbol for each graphic group is chosen. Besides, Figure 3 shows a sample of *Musical Avatars* from the three summarization strategies and Figure 4 shows a random sample of different *Musical Avatars.*

# 3. EXPERIMENTS AND RESULTS

## 3.1 User Data Analysis

In order to evaluate the proposed method, we worked with a group of 11 users (8 male and 3 female). They were aged between 25 and 45 years old (average $\mu = 33$ and standard deviation $\sigma = 5.35$) and showed a very high interest in music (rating around $\mu = 9.64$, with $\sigma = 0.67$, where 0 means no interest in music and 10 means passionate about music). Ten of the eleven users play at least one musical instrument, including violin, piano, guitar, singing, synthesizers and ukulele.

[4] `http://meegos.com`

[5] `http://avatars.yahoo.com`

[6] A neutral character corresponds to an empty avatar. It should be noted that the same representation can be achieved if all descriptor values are set to 0.5.

[7] `http://processing.org`

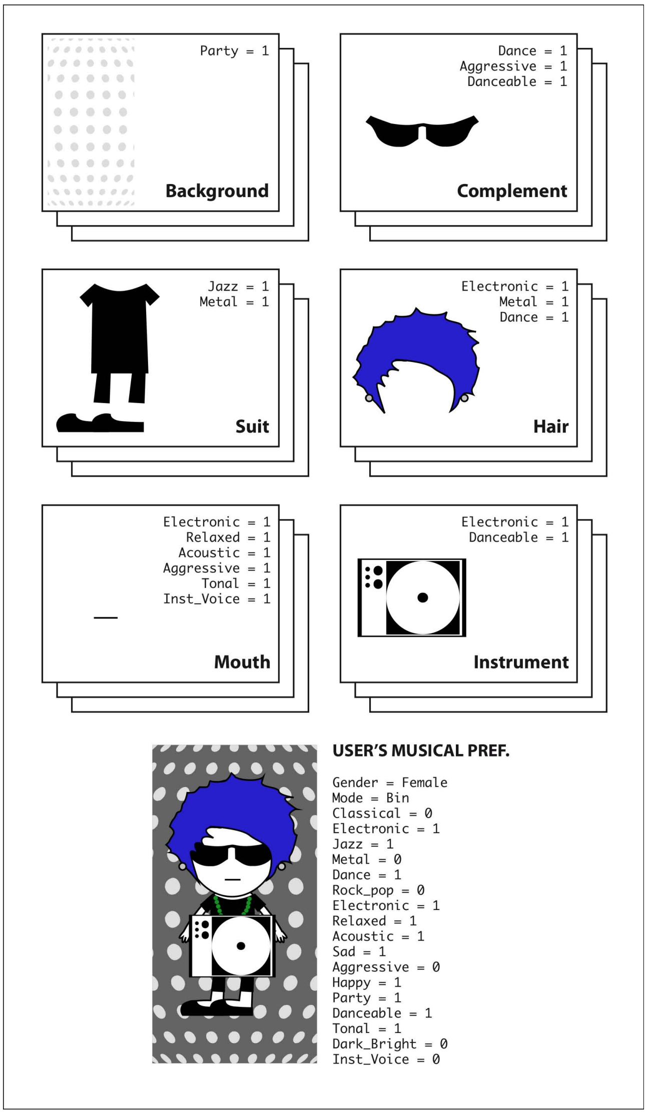

**Figure 2: Sample of the visualization strategy. It can be seen how the descriptor values influence the selection of the different graphic elements used to construct the avatar. The values inside the graphic element boxes represent all possible descriptor values that can generate the presented element.**

The number of tracks selected by the users to convey their musical preferences was very varied, ranging from 19 to 178 music pieces ($\mu = 74.09$, $\sigma = 48.23$). The time spent for this task also differed a lot, ranging from half an hour to 180 hours ($\mu = 30.41$, $\sigma = 54.19$).

It is interesting to analyze the provided verbal descriptions about the strategy followed to select the music tracks. Some of the users were selecting one song per artist, while some others did not apply this restriction. They also covered various uses of music such as listening, playing, singing or dancing. Other users mentioned musical genre, mood, expressivity, musical parameters and chronological order as driving parameters for selecting the tracks. Furthermore, some users implemented an iterative strategy by gathering a very large amount of music pieces from their music collection and performing a further refinement to obtain the final selection.

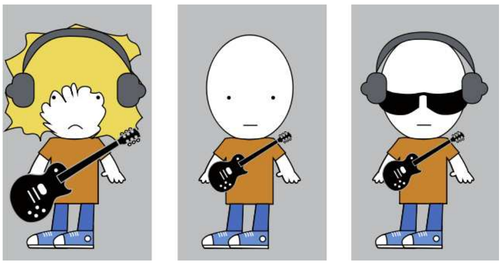

**Figure 3: Sample *Musical Avatars* from the three summarization strategies (i.e. *binary*, *ternary* and *continuous*).**

Finally, each user provided a set of labels to define their musical preferences. Most of them were related to genre, mood and instrumentation, some of them to rhythm and few to melody, harmony or expressivity. Other suggested labels were attached to lyrics, year and duration of the piece. The users' preferences covered a wide range of musical styles (from classical to country, jazz, rock, pop, electronic, folk) and musical properties (e.g. acoustic vs. synthetic, calm vs. danceable, tonal and dissonant).

## 3.2 Avatar Evaluation

After having generated the three *Musical Avatars* for each user (one per summarization strategy, as described in Section 2.3), we asked the users to answer a brief evaluation questionnaire. The evaluation consisted in performing two tasks. In the first task, we asked the user to manually assign the 17 semantic descriptors used to summarize her/his music collection (see Table 1). For this assignment we requested a real number between 0 and 1 to rate the degree of preference for each descriptor (e.g. 0 meaning "I don't like classical music at all" up to 1 meaning "I like classical music a lot"). Next, to introduce the user to the visual nature of the *Musical Avatars*, we showed 20 randomly generated avatars. For the second task, we presented the user six avatars: namely, the three images generated from his/her own collection, two randomly generated avatars and one neutral avatar. We asked the user to rank these images assigning the image that best express her/his musical preferences to the first position in the rank (i.e. rank = 1). Finally, we asked for written feedback regarding the images, the evaluation procedure, or any other comment[8].

From the obtained data we first analyzed the provided rankings to estimate the accuracy of the different mapping strategies. Given the users' preferences for the various *Musical Avatars* presented to them, we computed the mean rank for each of the different algorithm variants examined in the questionnaire. Resulting means and standard deviations can be seen in Table 3. A within-subjects ANOVA tested the effect of the summarization method on the ratings obtained from the subjects. A pre-requisite for this type of experimental design is to check the sphericity assumption (i.e. all the variances of the differences in the sampled population are equal) using the Mauchly's test, which indicated that

[8] A screenshot of the evaluation and more *Musical Avatars* are available online `http://mtg.upf.edu/project/musicalavatar`.

**Table 3: Mean ranks and standard deviations for the different summarization strategies obtained by the user evaluation. The random column corresponds to the average values of the individual random results (see text for details).**

| | Continuous | Binary | Ternary | Random | Neutral |
|---|---|---|---|---|---|
| $\mu$ | 1.73 | 2.27 | 2.91 | 4.28 | 5.18 |
| $\sigma$ | 0.79 | 1.49 | 1.45 | 1.16 | 0.98 |

the assumption was assumable [13]. Then, the effect of the summarization was found to be significant (Wilks Lambda = 0.032, $F(4, 7) = 52, 794$, $p < 0.001$). Pairwise comparisons (a least significant differences t-test with Bonferroni correction, which conservatively adjusts the observed significance level based on the fact that multiple comparisons are made) revealed significant differences between two groups of avatars: on one side the random and the neutral avatars (each one getting ratings that cannot be considered different from the other one), and on the other side the *binary*, *ternary* and *continuous* avatars (which get ratings that are statistically different from the random and the neutral ones, but without any significant difference between the three). The differences between those two clusters of avatars are clearly significant ($p < 0.005$) except for the differences between random and *ternary*, and between *binary* and neutral, which are only marginally significant ($p <= 0.01$).

We then introduced a dissimilarity measure to assess the significance of the generated description of musical preferences. In particular, we estimated how the computed representation performs against a randomly generated baseline. Therefore, we first computed the Euclidean distance between the obtained descriptor vector representing the user profile (standardized and range-normalized) and the vector containing the users' self-assessments provided in the first task of the evaluation. We then generated a baseline by averaging the Euclidean distances between the self-assessments and 10 randomly generated vectors. Finally, a t-test between the algorithm's output ($\mu = 0.99, \sigma = 0.32$) and the baseline ($\mu = 1.59, \sigma = 0.25$) showed a significant difference in the sample's means, $t(11) = -5.11, p < 0.001$. Additionally, Figure 5 shows box plots of the obtained dissimilarities.

## 3.3 Discussion

From the results presented above, we first observe that the generated description based on audio content analysis shows significant differences when compared to a random assignment. Moreover, as we can see in Figure 5, the mean distance to the user-provided values is remarkable smaller for the generated data than for the random baseline, i.e. the provided representations better resembles the users' self-assessments in terms of similarity.

Furthermore, Table 3 clearly shows a user preference for all three proposed quantization strategies over the randomly generated *Musical Avatars* and the neutral one. In particular, the *continuous* summarization strategy has been found top-ranked, followed by the *binary* and *ternary* quantization strategy. This ranking, given the ANOVA results, should be taken just as approximative until a larger sample of user evaluations is available. On the other hand, we can also see that the neutral avatar is less preferred than the random avatars. This suggests that the users prefer images that carry some information (even if it does not match the users' preferences) rather than avatars lacking of visual features. This poses the problem of visualizing users with varied musical preferences (i.e. mean values of a majority of the descriptors close to 0.5), especially in the case of the ternary quantization. Unfortunately, the difference between the random and neutral avatars showed not significance, probably due to the small number of participants in the study.

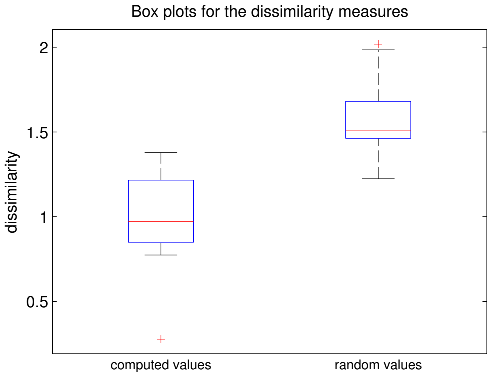

**Figure 5: Box plots of the dissimilarity estimation. The Euclidean distance obtained by the computed descriptors shows a significantly lower mean than the one obtained by the 10 randomly generated descriptor vectors.**

Evaluation of the users' comments can be summarized as follows. First, a general tendency towards an agreement on the representativeness of the *Musical Avatar* can be observed. As expected, some users reported missing categories to fully describe their musical preferences (e.g. country music, musical instruments). This suggests that the provided semantic dimensions seem to grasp the essence of the user's musical preference, but fail to describe subtle nuances in detail. Indeed, by providing better semantic descriptions of the musical content under consideration (i.e. better classifier/descriptors), the algorithm's accuracy in describing these aspects would benefit to a great extent. In consequence, since we are working with state-of-the-art algorithms [3], the available tools are only able to solve the problem on a very coarse level.

Finally, the meaningfulness of some visual features could not be decoded by some users (e.g. glasses, head shape). Therefore we confirm our intuition that the mapping of the descriptors to the visual domain is not a trivial task and will need further research efforts. Moreover, one key aspect of the proposed approach is the necessity of an explicit preference statement in form of a sufficient set of music tracks by the user. Alternatively one may use listening behavior information, exploiting services such as Last.fm, to infer such sets automatically.

# 4. CONCLUSIONS

We have presented a method which maps the musical preferences of a given user to the visual domain. To this extent, given a collection of music tracks (provided by the user as a prototypical example of his/her preferences) the proposed algorithm extracts high-level semantic information by audio content analysis of the tracks. Then, it summarizes the obtained descriptions into a simple representation of the user as a point in a semantic space, and finally maps it into a graphical representation - the *Musical Avatar*. We considered 3 summarization strategies, which generate a simplified user representation in continuous and discrete ways. This user profile is then mapped to the visual features of the avatar. We carried out subjective evaluations of the considered strategies together with a random and neutral avatars as alternative baselines. According to the obtained results, we found participants' preference for all three proposed approaches over both baselines. In general, we conclude that the *Musical Avatar* provides a reliable, although coarse, visual representation of the user's music preferences.

As future work, we plan to focus our research on performance improvements, enriching the current method with more semantic dimensions (e.g. instrument information). Furthermore, we want to increase the number and quality of the visual attributes as well as improve the mapping strategy to obtain a more intuitive representation. Finally, we plan to conduct a large scale web-based user evaluation in order to better assess the representativeness of the obtained avatars and to further refine the proposed method.

## 5. REFERENCES

[1] N. Ahmed, E. de Aguiar, C. Theobalt, M. Magnor, and H.-P. Seidel. Automatic generation of personalized human avatars from multi-view video. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 257–260, 2005.

[2] J. Aucouturier. Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In J. Minett and W. Wang, editors, *Language, Evolution and the Brain*, pages 35–64, 2009.

[3] D. Bogdanov, J. Serrà, N. Wack, and P. Herrera. From low-level to high-level: Comparative study of music similarity measures. In *International Workshop on Advances in Music Information Research*, 2009.

[4] P. M. Brossier. *Automatic Annotation of Musical Audio for Interactive Applications*. PhD thesis, QMUL, London, UK, 2007.

[5] O. Celma and X. Serra. Foafing the music: Bridging the semantic gap in music recommendation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):250–256, 2008.

[6] W. Chai and B. Vercoe. Using user models inmusic information retrieval systems. In *International Symposium on Music Information Retrieval*, 2000.

[7] S. Downie, D. Byrd, and T. Crawford. Ten years of ISMIR: Reflections on challenges and opportunities. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, pages 13–18, 2009.

[8] M. Grimaldi and P. Cunningham. Experimenting with music taste prediction by user profiling. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 173–180, 2004.

[9] G. Grimmett. *Probability and random processes*. Oxford University Press, 3rd edition, 2001.

[10] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, UPF, Barcelona, Spain, 2006.

[11] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the International Conference on Machine Learning*, pages 359–366, 2000.

[12] J. Holm, A. Aaltonen, and H. Siirtola. Associating colours with musical genres. *Journal of New Music Research*, 38(1):87 – 100, March 2009.

[13] C. Huberty. *Applied MANOVA and discriminant analysis*. Wiley-Interscience, Hoboken N.J., 2nd edition, 2006.

[14] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra. Indexing music by mood: Design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 2009.

[15] D. Levitin. *The world in six songs: how the musical brain created human nature*. Plume, New York, 2009.

[16] M. Maffesoli. *The Time of the Tribes: The Decline of Individualism in Mass Society*. Sage, 1995.

[17] M. Mandel and D. Ellis. Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Symposium on Music Information Retrieval*, pages 594–599, 2005.

[18] S. McCloud. *Understanding Comics: The Invisible Art*. Kitchen Sink Press, 1993.

[19] A. North and D. Hargreaves. Music and adolescent identity. *Music Education Research*, 1(1):75–92, 1999.

[20] A. C. North and D. J. Hargreaves. Lifestyle correlates of musical preference. *Psychology of Music*, 35(1):58–87, 2007.

[21] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Project Report*, 2004.

[22] E. Petajan. *Real-Time Vision for Human-Computer Interaction*, chapter MPEG-4 Face and Body Animation Coding Applied to HCI. Springer US, 2005.

[23] P. Rentfrow and S. Gosling. The Do Re Mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

[24] P. Rentfrow and S. Gosling. Message in a ballad. *Psychological Science*, 17(3):236–242, 2006.

[25] D. Sauer and Y.-H. Yang. Music-driven character animation. *ACM Transactions Multimedia Computing, Communications and Applications*, 5(4):1–16, 2009.

[26] S. Stober and A. Nürnberger. User-adaptive music information retrieval. *Künstliche Intelligenz*, 23(2):54–57, 2009.

[27] S. Streich. *Music complexity: a multi-faceted description of audio content*. PhD thesis, UPF, Barcelona, Spain, 2007.

[28] V. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 2nd edition, 1999.

[29] M. Voong and R. Beale. Music organisation using colour synaesthesia. In *CHI '07: extended abstracts on Human factors in computing systems*, pages 1869–1874, 2007.

[30] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian. Musical genre classification using support vector machines. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 429–432, 2003.

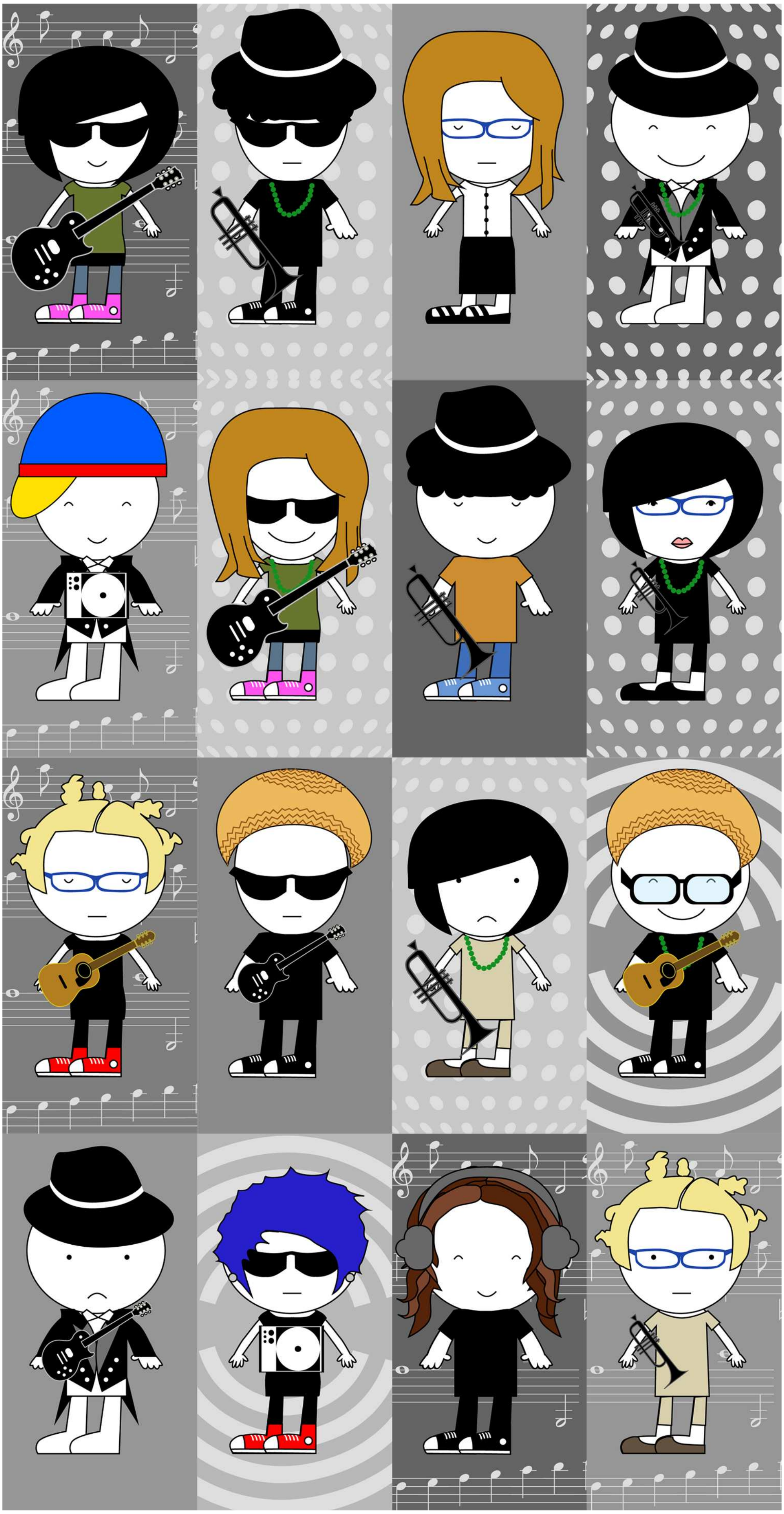

Figure 4: A random sample of *Musical Avatars*.