# SEARCHING FOR DOMINANT HIGH-LEVEL FEATURES FOR MUSIC INFORMATION RETRIEVAL

*Massimiliano Zanoni* [1], *Daniele Ciminieri* [2], *Augusto Sarti* [1], *Stefano Tubaro* [1]

Dipartimento di Elettronica e Informazione - Politecnico di Milano - Italy
[1] {zanoni,sarti,tubaro}@elet.polimi.it
[2] daniele.ciminieri@mail.polimi.it

## ABSTRACT

Music Information Retrieval systems are often based on the analysis of a large number of low-level audio features. When dealing with problems of musical genre description and visualization, however, it would be desirable to work with a very limited number of highly informative and discriminant macro-descriptors. In this paper we focus on a specific class of training-based descriptors, which are obtained as the log-likelihood of a Gaussian Mixture Model trained with short musical excerpts that selectively exhibit a certain semantic homogeneity. As these descriptors are critically dependent on the training sets, we approach the problem of how to automatically generate suitable training sets and optimize the associated macro-features in terms of discriminant power and informative impact. We then show the application of a set of three identified macro-features to genre visualization, tracking and classification.

***Index Terms***— High-level descriptors, Music genre classification, Music Information Retrieval

## 1. INTRODUCTION

In the past few years the exponential growth of networked musical contents has created much demand for the development of applications for cataloging, exploration and fruition of a large musical database. The study of effective solutions has been a main task within the Music Information Retrieval (MIR) community [1]. In particular, one of the main issues that remains open concerns the choice of an effective and expressive paradigm for music description. In general, representation is still performed using traditional modalities based on meta-tags (a context-based approach), generally defined by humans, or based on a set of low-level features (a content-based approach). Unfortunately, in particular for heterogeneous streams, the two approaches are not sufficient to describe the audio content: the user may be interested in knowing what currently happens on a particular stream, exploiting a simple semantic description of the related audio characteristics. The context-based approach is able to produce meaningful descriptors, but are generally intended to be "global": they tend to describe the whole excerpt, or big segments of it. On the other hand, the content-based approach based on low-level features produces time-variant descriptors, but is semantically poor. One of the main open issues in the music description area is the definition of a set of content-based time-variant and highly descriptive high-level features that are semantically meaningful (macro-descriptors).

The literature is rich with studies that focus on compact descriptors that are characterized by a high level of abstraction. Most of them rely on a model-based approach to extract information on harmony, melody, rhythm, etc. [2][3][4]. More recently, a new class of unstructured (training-based, as opposed to model-based) macro-descriptors have been introduced [5], which are evolutionary (time-varying) in nature. Each descriptor is defined as the log-likelihood of a Gaussian Mixture Model (GMM) trained with a set of short musical excerpts that exhibit a certain semantic homogeneity. In order to define such descriptors, musical excerpts need to be carefully collected and organized. The training set of each macro-descriptor has to be populated by excerpts that should be short enough to guarantee the desired exclusive semantic homogeneity (only one high-level characteristic shared by all the excerpts in the same training set). This approach, however, greatly suffers from the difficulty of choosing the correct high-level features, and generating the related datasets accordingly.

In this study we try to reverse this paradigm by investigating the possibility of identifying the macro-features and the related training datasets in an automated fashion, through an optimization process. The former starts from the assumption that musical excerpts that share common high-level features tend to naturally cluster in the feature space, and that each of the resulting clusters can be used for defining a different high-level descriptor. With this goal in mind, we define two different functionals to be optimized and we evaluate which approach is the most suitable for the problem at hand. The first method that we propose, therefore, determines the minimal set of low-level features that maximize a measure of the clustering quality. The second optimization method starts from the assumption that what matters is the discriminating

ability of the resulting high-level features. Based on this assumption we propose a method for identifying a minimal set of low-level features that generate maximally discriminant high-level descriptors. In this paper we implement and compare the two approaches and show that the second one greatly outperforms the first.

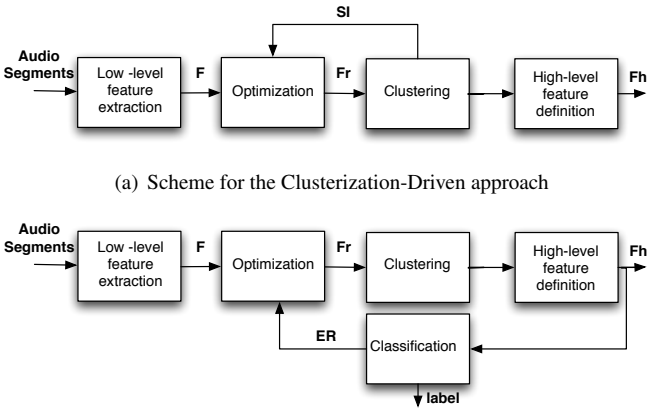
Most of the applications on Music Information Retrieval(MIR) are greatly influenced by the choice of the data description paradigm. Music genre-classification and genre-visualization are a specific example. In this paper we also describe an application of the method to the visualization and classification problem and we will show some advantages in using the defined macro-descriptors. The standard approaches tend to perform classification on music genre through the determination of an appropriate label. Content-base tagging systems are usually based on statistical pattern recognition classifiers [6][7][8]. Tags, however, are "global" descriptors, in the sense that they apply to the whole musical excerpt, therefore are unable to describe and track the genre evolution over the song extension. Moreover, due to the subjectiveness of the genre taxonomy, it would be desirable to be able to describe and track genre transitions, cross-genre excerpts, and new trends in genre cross population. The use of a meaningful time-variant representation generated by a very limited number of highly informative and highly discriminant descriptors (macro-descriptors) can be suitable in dealing with problems of genre description, visualization and tracking. In this paper we also describe an application of the method to the visualization and classification problem.

## 2. APPROACH

The overall block diagrams that describe the two approaches are shown in Figs. 1(a) and 1(b). Such diagrams exhibit the same four main blocks: *Low-level feature extraction*, *Optimization* (low-level feature reduction), *Clustering* and *High-level feature generation*, although the optimization method in either case is driven by different criteria, as shown by the different feedback controls. Techniques adopted in this study represent a part of the standard in MIR areas and permit to cover a wide range of cases.

### 2.1. Reference Scheme

Given the training dataset, a large collection of low-level features $F$ is extracted. The collection is chosen large enough to best characterize audio segments, we considered a set composed by 44 features, and to best capture Timbral, Rhythmic and Tonal (Harmonic) characteristics of the musical excerpts. The resulting feature space is then narrowed down in dimensionality through an iterative optimization process aimed at ultimately determining the best macro-descriptors. The optimization process will be shown in detail in Sections 2.2 and 2.3. The resulting reduced set of feature $F_r$ is used in clustering process. The clustering method used in the study is the *K-means* algorithm [9]. K-means aims at partitioning $n$ observations into $N_C$ clusters $C_0$, ..., $C_{N_c}$ ($1 < N_c < n$) in which each observation belongs to the cluster with the nearest mean, so as to minimize the within-cluster sum of squares (WCSS):




(a) Scheme for the Clusterization-Driven approach

(b) Scheme for the Classification-Driven approach

**Fig. 1**: Overall Schemes. $F$ is the starting set of features, $Fr$ the reduced set of feature, $Fh$ the set of marco-descriptors, SI the Silhouette Index, ER the Classification Error

$$\arg\min_{C} \sum_{i=1}^{N_C} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 \quad (1)$$

where $x_j$ is an observation and $\mu_i$ is the mean of points in $C_i$. Once clusters are defined, the set of marco-descriptors are modeled on the log-likelihood of a properly trained Figueredo-Jain Gaussian Mixture Models (FJ-GMM) [10] for each cluster. A GMM provides a statistical model for data point distribution by using a mixture of Gaussian components and it is defined as follows.

$$p(x|\lambda) = \sum_{m=1}^{N_C} c_m b_m(x) \quad (2)$$

where $x$ is an observation, $c_m$ is the weight associated with the component and $b_m(x)$ is a Gaussian density function, parameterized by a mean vector $\mu_m$ and the covariance matrix $\Sigma_m$.

### 2.2. The Clustering-Driven Optimization Process

In this work the optimization process is performed using a Genetic Algorithm (GA) [11]. GAs are stochastic-optimization methods that encode each point in a solution space into a string called a "chromosome". The role of the GA is to choose the subset of feature that produce the optimal solution. For that reason, each gene in the chromosome is associated to a specific feature in a binary fashion. The binary digit, in fact, encodes the presence of that feature in the proposed solution. If $F$ is the set of features and $N_F$ its cardinality, each chromosome will have a length of $N_F$. The considered cardinality

of the population in this work is 50 chromosomes. The initial population of chromosomes is randomly chosen and *parent selection*, *crossover* and *mutation operators* are used in order to model its evolution. The adopted selection scheme is *statistically uniform*, while parents are chosen depending on a fitness value. The chosen crossover method is the *scattered crossover*, where a child is generated starting from two parents. Based on a random binary vector of length $N_{Fr}$, a gene is selected from the first parent in the case of 0 in the corresponding gene in the random vector, otherwise is selected from the second parent. At last, the adopted mutation scheme is based on a probability-based mutation rate.

The quality of the solution is depicted by the *fitness value*, which relates to the objective function of the optimization problem. In the *Clustering-Driven* approach the Silhouette index ($SI$) [9] is chosen as *fitness values*. The Silhouette Index is a measure of intra-cluster compactness and the inter-cluster distance and it is defined as follow: given an observation $x$ belonging to cluster $C_i$, $SI(x) = 0$ if $x$ is the only point has been in $C_i$, otherwise:

$$S(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \tag{3}$$

$a(x)$ being the average distance between $x$ and all points in its cluster $C_i$, and $b(x)$ being the distance between $x$ and its nearest cluster $C_j$ with $C_i \neq C_j$. The Silhouette index of the whole cluster is defined as the average index over all observations.

### 2.3. The Classification-Driven Optimization Process

As shown in 1(b), the *Classification-Driven* optimization process differs from the clustering driven method in the index used for the convergence of the GA. In this approach, in fact, the optimization process is controlled by the discriminant ability of the macro-descriptors. As a consequence, the overall rate in classification accuracy is used as *fitness value*. In order to do so, a Genre Classification step is needed. Genre classification is based on a battery of trained Support Vector Machines (SVMs) [12]. Given the set $G$ of all genres, a SVM for each possible pair of genres is produced, and a one-against-one majority-voting classification paradigm is adopted. Given the nonlinearity of the problem, nonlinear SVMs is here used, based on radial basis functions. Parameter estimation is performed using a grid-search approach and the cross-validation method is described in [12]. Each SVM is trained and tested using high-level features extracted from the audio segments that are in the labeled training set, as well as in the unlabeled test dataset.

The adopted error rate is the Correct Classification Rate (CCR) which is directly obtained from the confusion matrix resulting from the SVMs label prediction and defined as:

$$R_{CC} = \frac{\sum_{i=1}^{D} M_f(i,i)}{N} \tag{4}$$

where $M_f$ is the confusion matrix, dependent on a chosen set of features, $N$ is the total number of elements in $M_f$ and $D$ is the number of classes. The error rate can now be defined as:

$$E_R = 1 - R_{CC} \tag{5}$$

This rate ranges from 0 (perfect classification) to 1 (totally wrong classification).

## 3. EXPERIMENT SETUP AND IMPLEMENTATION

As one of the main goals of the macro-descriptors is to capture the temporal evolution of the captured features, the training for defining the macro-descriptors was based on 3s-long audio segments, which were extracted from longer excerpts through a texture analysis process based on spectral peak detection over the spectrum's *novelty* function [13]. This allowed us to determine the temporal locations of relevant texture changes. The *novelty* curve has been obtained by the convolution of the similarity matrix, resulting from computing the correlation between all pairs of frames of the spectrogram of the signal, along with the main diagonal, as described in [13]. The training datasets $D_{GMM}$ and $D_{SVM}$ were made of 3000 segments each, equally distributed over the various musical genres. For each segment we extracted the set $F$ of low-level features. Most features were averaged over widows of 0.021 seconds with a 50% overlap. The shape of the window was dependent on the feature. Rhythmic features, on the other hand, needed longer frames to capture meaningful information; therefore they were computed over the entire segment. For clustering we chose a K-means algorithm, which needs an a-priori definition of the number $N_C$ of clusters to be discovered. In our case $N_C$ did not go beyond 4, due to the limited amount of data available for GMM training.

As explained above, the *Clustering-Driven* optimization process, performed by the GA, is controlled by the silhouette index. As we can see in Table 1, the minimum number of considered clusters is limited to 2. Moreover, in order to keep the computational cost to a manageable level, a predefined set of possible values of $N_{Fr}$ was considered: $\{9, 12, 15, 18\}$. The lower boundary guarantees that the features are never limited to a single group of descriptors (e.g. chroma features). In the *Classification-Driven* approach the GA is controlled by the classification accuracy while no clustering quality rate is considered. As a consequence, we do not need to predefine the number of features to use. In this approach the macro-descriptors defined in each step are used for training the rake of SVMs. The resulting models $S = s_1, .., s_{N_s}$ are then used in a test phase over a still unused database: $T_{svm}$ and the classification accuracy is finally retrieved.

## 4. TESTING AND EVALUATION

Our reference database for training and testing is public available one used in MIREX 2004 [1]

### 4.1. Semantic Consistency

In order to test the semantic consistency of audio clusters and determine suitable labels for the corresponding macro-descriptors, we had a set of 24 individuals conduct a perceptual evaluation and fill out a questionnaire. Testers where invited to listen to a 1-minute audio stream for each cluster, each made of a sequence of 3-second segments corresponding to the points that laid the closest to the centroid of their cluster. The testers were then asked to select a label (out of a set of 7) that best described the stream and differentiated it from the others. The *Clustering-Driven approach*, obviously, favors the case with the highest silhouette index, which corresponds to 2 clusters and 12 features, as shown in Table 1. The list of features resulting by the optimization process is: *zero crossing rate, spectrum spread, roughness, irregularity, spectral flux, MFCC coefficient 1, pulse clarity, Chroma (C,C#,D,F,A)*[14]. The results of the perceptual test are shown in Fig. 2. As we can see, there seems to be no clear opinion over which label applies to each macro-descriptor, which confirms that the clustering quality is not a suitable criterion for grouping musical segments with similar high-level characteristics. More conclusions can be drawn by looking at Table 1, where the accuracy of the classification step is shown as well. We notice, in fact, that the macro-descriptors obtained with the clustering-based criterion appear to have a weak discrimination ability. Much better results come from the classification-driven approach. In this case the configuration that produced the highest classification accuracy was made of 3 clusters and the list of features resulting by the optimization process is: *zero crossing rate, estimated bpm, RMS, spectral centroid, brightness, spectral entropy, spectral flatness, roughness, irregularity, inharmonicity, spectral flux, event density, MFCC (coefficients 1,4,5,9), harmonic flux, pulse clarity, Chroma (D,F,G#,A#,B)* [14]. As shown in Fig. 3, the testers clearly identified the labels that they judged as the most suitable for describing the determined high-level features and related macro-descriptors: **Grooviness**, **Classicality** and **Roughness**.

| $N_C$ | $N_{Fr}$ | **silhouette** | **accuracy** |
|---|---|---|---|
| 2 | 9 | 0.647 | 0.368 |
| | 12 | 0.784 | 0.388 |
| | 15 | 0.691 | 0.382 |
| | 18 | 0.529 | 0.394 |
| 3 | 9 | 0.437 | 0.416 |
| | 12 | 0.542 | 0.392 |
| | 15 | 0.469 | 0.370 |
| | 18 | 0.527 | 0.429 |

**Table 1**: Clustering quality and Classification accuracy. $N_C$ is the number of clusters and $N_{Fr}$ the number of features

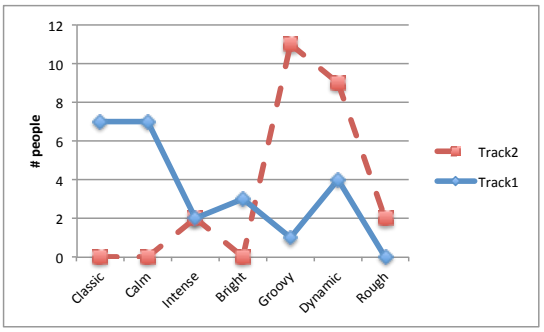



**Fig. 2**: Perception labeling in the Clustering-Driven approach. Tracks are the streams propose to the testers

[1] http://www.music-ir.org/

| | classical | electronic | jazz blues | metal punk | pop rock |
|---|---|---|---|---|---|
| classical | **63.5** | 1 | 29.5 | 0 | 6 |
| electronic | 36 | **4** | 31 | 11 | 18 |
| jazz blues | 22.5 | 3 | **42.5** | 7.5 | 24.5 |
| metal punk | 9 | 2.5 | 5 | **66** | 17.5 |
| pop rock | 27 | 3 | 25 | 29 | **16** |

**Table 2**: Classification using the Clustering-Driven approach

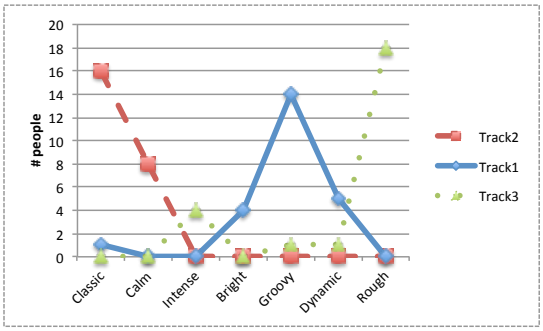



**Fig. 3**: Perception labeling in the Classification-Driven approach. The tracks are the streams proposed to testers

### 4.2. Discrimination Capability

By using the trained GMM models $B = b_1, .., b_{N_C}$ and SVM $S = s_1, .., s_{N_S}$ we tested the discrimination ability of macro-descriptors, for both approaches, over a previously "unseen" database $T'_{SVM}$, made of 200 homogeneous 3s segments belonging to *classic*, *jazz-blues*, *electronic*, *metal-punk* and *pop-rock* genres. The results are presented using the confusion matrix. As expected, macro-descriptors defined using the *Clustering-Driven* approach turned out to have a limited discriminant ability. The results are shown in table 2. Goods results, on the other hand, were obtained with the *Classification-Driven* approach, as shown in Table 3. The

| | classical | electronic | jazz blues | metal punk | pop rock |
|---|---|---|---|---|---|
| classical | **88.5** | 2.5 | 9 | 0 | 0 |
| electronic | 12 | **57** | 19 | 3.5 | 8.5 |
| jazz blues | 20.5 | 9 | **67.5** | 2.5 | 0.5 |
| metal punk | 2 | 9 | 12 | **66** | 11 |
| pop rock | 12 | 28 | 26.5 | 17.5 | **16** |

**Table 3**: Classification using Classification-driven approach

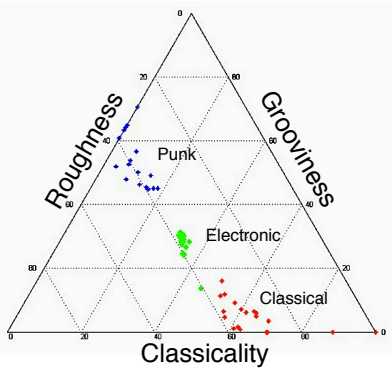



**Fig. 4**: Resulting triangular plot for a stream composed by segments belonging to three different genres (punk, classic, electronic)

poor accuracy obtained in the pop/rock genre is mainly due to the inherent inhomogeneity of the dataset class.
The results that we obtained are comparable with those of other recent works, even if we consider high-level features on a very short portion of the whole song [7][8]. This confirms that the macro-descriptors are meaningful as well as discriminant.

### 4.3. Visualization

As an example of the application of the determined macro-descriptors, we developed a visualization system that re-maps them onto a 3D diagram, to track their temporal evolution. Each descriptor was mapped onto a different axis of the triangular graph and on the RGB color space. Segments that are similar in genre would therefore tend to cluster on the diagram, while cross-genre transitions would be correctly visualized. The quality of the tracking was assessed through a questionnaire submitted to 19 individuals. A 3-min. long stream made of a sequence of 3s segments (not previously "seen") belonging to three different genres (punk, classic, electronic) were proposed to a set of testers, who were asked to rate the quality of the tracking using a ranking from 1 to 5. The visualization system was well received, as 4 people rated it 3 out of 5, 14 people rated 4 out of 5, and 1 person gave it a full score. In Fig. 4 we can see how genres tend cluster in the visualization diagram: the right-bottom cluster represents classical segments, the one in the center represents the electronic genre and the left-top one represents the segment belonging to punk music.

## 5. CONCLUSION AND FUTURE DEVELOPMENTS

In this work we approached the problem of how to automatically identify unstructured (training-based) high-level macro-features for music visualization and description. We defined and tested two optimization processes based on different objective functions: one that favors clustering quality, and the other favoring discrimination properties. We concluded that, while the former function does not lead to sufficiently meaningful descriptors, the latter generates macro-features that carry a semantic interpretation while retaining a relevant discriminant power. In order to further confirm the validity of the second method we developed a visualization system based on the macro-features that the system found, which proved effective to track genre transitions and cross-genre musical excerpts. As for the discriminant power of the resulting macro-features, we showed that a classification system based on them has roughly the same performance as state-of-the-art classifiers based on low-level features.

## 6. REFERENCES


[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," in *Proc. IEEE*, 2008, vol. 96, pp. 668–696.

[2] C. Perez-Sancho, D. Rizo, J.M. Inesta, P.J.P. de Leon, S. Kersten, and R. Ramirez, "Genre classification of music by tonal harmony," *Journal of Information Science and Engineering*, vol. 14, pp. 21–22, 2010.

[3] R. Mayer, R. Neumayer, and A. Rauber, "Rhythm and style features for musical genre classification by song lyrics," in *proc. of the 9th International Conference of Music Information Retrieval (ISMIR)*, 2009.

[4] D. Bogdanov, J. Serra, N. Wack, P. Herrera, and X. Serra, "Unifying low-level and high-level music similarity measures," *IEEE Transaction on Multimetida*, vol. 13, pp. 687–701, 2011.

[5] G. Prandi, A. Sarti, and S. Tubaro, "Music genre visualization and classification exploiting a small set of high-level semantic features," in *proc. of the 12th International Conference on Digital Audio Effects (DAFx09)*, Como, Italy, 2009.

[6] J. G. A. Barbedo and A. Lopes, "Automatic genre classification of musical signals," *EURASIP Journal on Advances in Signal Processing*, pp. 1–13, 2007.

[7] Y. Panagakis and C. Kotropoulos, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, pp. 576–588, 2010.

[8] A. Holzapfel and Y. Stylianou, "Musical genre classification using non-negative matrix factorization-based features," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, pp. 424–434, 2008.

[9] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 2005.

[10] Mario A. T. Figueiredo and Anil K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, pp. 381–396, 2000.

[11] F. Pachet and P. RoyRauber, "Exploring billions of audio features," in *proc. of International Workshop on Content-Based Multimedia Indexing (CBMI 07)*, 2007, pp. 227–235.

[12] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery Journal*, vol. 2, 1998.

[13] J. Foote and M. Cooper, "Media segmentation using self-similarity decomposition," in *proc. of Storage and Retrieval for Multimedia Databases (SPIE' 03)*, 2003, pp. 67–75.

[14] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora, *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*, John Wiley & Sons, 2005.