ESTIMATING DOUBLE THUMBNAILS FOR MUSIC RECORDINGS

Nanzhu Jiang and Meinard Müller

International Audio Laboratories Erlangen

nanzhu.jiang@audiolabs-erlangen.de, meinard.mueller@audiolabs-erlangen.de

ABSTRACT

Audio thumbnailing, which aims at finding the most representative audio segment of a music recording, is an important task in music information retrieval. In general, the notion of a thumbnail is not well-defined and several musical parts may be good thumbnail candidates. For example, for popular music, both a verse and a refrain section may serve as suitable thumbnail candidates. Instead of considering only one thumbnail, we consider in this paper the problem of finding the two most representative segments that correspond to different musical parts. We denote these two segments as double thumbnails. As our main technical contributions, we propose two approaches for computing double thumbnails, both extending a previously introduced repetition-based thumbnailing procedure. In the first approach, which is straightforward, we simply apply the original thumbnailing procedure two times in an iterative fashion. In the second approach, we introduce a novel method for jointly estimating the two thumbnails within one optimization procedure. Finally, we report on experimental results demonstrating the performances of the two double thumbnailing procedures and indicate directions towards full music structure analysis.

Index Terms— Music, Thumbnailing, Repetition, Structure, Segmentation

1. INTRODUCTION

Automatic music structure analysis constitutes a central research topic in the field of music information retrieval [1]. A prominent subproblem, which is closely related to music structure analysis, is the *audio thumbnailing* task, where the objective is to automatically determine the most representative segment of a given music recording [2, 3, 4, 5, 6, 7]. Typically, such a segment has many (approximate) repetitions throughout the piece, which makes listeners remember this segment as a representative for the piece of music. Therefore, most automated thumbnailing procedures aim to identify a structural section with many repetitions as the thumbnail segment [1, 8, 9, 10, 11, 12, 13].

In general, the notion of a thumbnail is not clearly defined and several musical parts may be good thumbnail candidates. Different listeners may have different preference on choosing a segment as the thumbnail for a certain piece of music. For example, in popular music, both a verse or a refrain section may serve as a suitable thumbnail candidate. Therefore, instead of considering only one thumbnail, in this paper we consider the problem of finding the two most representative segments that correspond to different musical parts. We denote these two segments as *double thumbnails*. Figure 1 shows an example for such double thumbnails (horizontal axis).

Another motivation of estimating double thumbnails is that such estimation can help for performing a full structure analysis of a piece of music. A thumbnail and its related repetitions often correspond to an important structure part of the music. Therefore, by estimating multiple thumbnails, distinct categories of structural parts may be derived. Figure 1 indicates two such structural parts and their repetitions (vertical axis). In particular, the structure of popular music often consists of only two main repeating sections which are the verse and the refrain, whereas other sections such as the intro and the bridge sections are not repetitive. For such pieces, by estimating double thumbnails and their repetitions for such pieces, we can already identify the entire music structure.

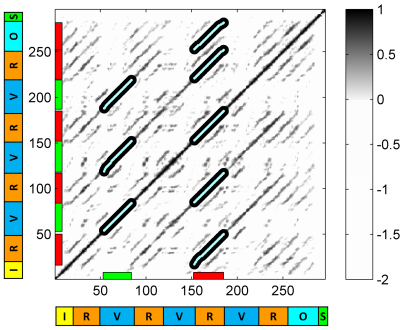

Fig. 1. Illustration of the double thumbnails computed for the Beatles song "Devil In Her Heart" using the joint approach. The double thumbnails (horizontal axis) exactly corresponds to a V (verse) and a R (refrain) section in the ground truth segmentation (indicated by the colored rectangles). We also present the computed optimal joint path family (cyan paths), and its induced segments (vertical axis) which correspond to the two main repetitive verse and refrain parts in the ground truth. Note that the O (outro) part annotated in the ground truth is actually a fading version of the refrain part.

In our previous work [7], a thumbnailing procedure was introduced in order to find the most repetitive segment for a music recording. In this paper, we build upon this procedure and with the goal to find double thumbnail segments. As main contributions of this paper, we propose two approaches for computing double thumbnails (related to the two most repetitive sections) for a music recording. In the first approach, we simply apply the original thumbnailing procedure two times in an iterative fashion. In the second approach, we compute the two thumbnails within one optimization procedure that tries to jointly maximize the score and coverage of two different disjoint segments. Furthermore, by extracting the repetitions of thumbnail segments, we can also identify large portions of the repetitive structures of the music recording.

The remainder of this paper is organized as follows. First, we briefly summarize our previous work in Section 2. Then, we introduce the iterative approach in Section 3 and the joint approach in Section 4. Finally, we report on our systematic experiments and conclude in Section 5.

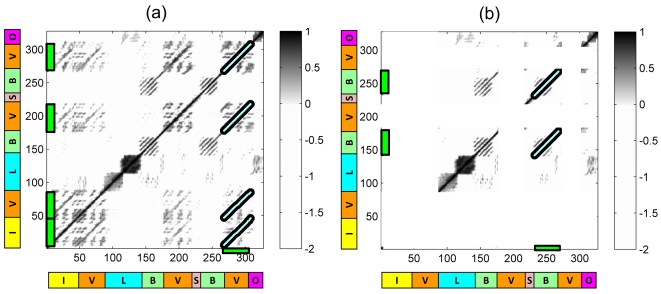

Fig. 2. Illustration of the iterative approach for computing double thumbnails for the Beatles song "Birthday". We present the enhanced self-similarity matrix (SSM), the thumbnail segment (horizontal axis), its optimal path family (cyan colored), and the induced segments (vertical axis). The colored rectangles indicate the ground truth structure annotation. **(a)** The first round computation (the original thumbnailing procedure). **(b)** The second round computation. Note that some regions which correspond to the induced segments of the first round are deleted in the SSM.

2. THUMBNAILING PROCEDURE

Before we introduce the estimation of double thumbnails, we first briefly describe the original thumbnailing procedure proposed in [7]. The main idea is to compute a fitness measure that captures *repetitiveness* as well as *coverage* for each possible segment of a given audio recording [7]. In the computation of the fitness measure, first an enhanced *self-similarity matrix* (SSM) is computed on the basis of chroma features [14] extracted from the music recording. To deal with local tempo differences and local key changes between the repetitions, we enhanced the SSM to achieve a higher degree of transposition invariance and tempo invariance [15]. Next, for each segment, an *optimal path family* that simultaneously reveals the relations between the segment and all other similar segments is computed. By projecting such an optimal path family to the vertical axis, one obtains an induced segment family, where each element in this family is similar to the given segment. Note that by our imposed constraints, these induced segments can not overlap with each other. The fitness measure of a segment is associated with some kind of score and coverage of the optimal path family. After that, we compute fitness values for all possible segments of an audio recording and select the segment with the maximum fitness as the thumbnail. As an illustration, Figure 2a shows the thumbnail segment, the optimal path family of the thumbnail, and the induced segments. By comparing the thumbnail to the ground truth annotation, we can see that it corresponds to the third V (verse) part in the annotation. Furthermore, three of the four induced segments correspond to the three annotated V parts. The first induced segment (the bottom one on the vertical axis) corresponds to the I (Intro) part, however, this intro is actually an instrumental version of the verse and can be considered as a special "verse".

3. THE ITERATIVE APPROACH

In the example shown in Figure 2, according to the ground truth annotation, the most repetitive part is the V sections. Furthermore, the B (bridge) part is also repeated. To detect the B part sections, one idea is to first exclude the V part sections from further considerations and then to again apply the thumbnailing procedure on the remainder of the recording. By excluding the induced segments of the previous iteration and re-estimate the thumbnail segment, we can derive double thumbnails and their repetitive sections.

To illustrate the behavior of this iterative approach, we revert to the example in Figure 2. The first round of the approach works exactly as the original thumbnailing procedure (Figure 2a). In the second round, the approach deletes those regions which correspond to the induced segments of the first round from the SSM and applies the thumbnailing procedure again. This yields the result shown in Figure 2b. By comparing this result to the ground truth, we can see that the thumbnail in Figure 2b exactly correspond to the B part, and its induced segments go along with all repetitive B sections. In this way, in two rounds of computation, we successfully identified all repetitive parts of this audio recording.

One problem of the iterative approach is that the thumbnail estimated in the second round is dependent on the result of the first round. Therefore, if the estimation of the thumbnail and its induced segments is erroneous, the identification of other repeating sections in the next round becomes problematic. Instead of this "greedy" iterative approach, we consider an alternative approach in the next section.

4. THE JOINT APPROACH

We now introduce a second approach, named "joint approach", which optimizes the estimation of two thumbnail segments simultaneously. Based on our definition of the fitness measure that captures repetitiveness of one segment [7], we now extend it and propose a joint fitness measure to capture the repetitiveness of a pair of segments. We closely follow the notations which originally introduced in [7] and extend some definitions.

4.1. Joint Path Family

Let $X = (x_1, x_2, \ldots, x_N)$ be a feature sequence and $\mathcal{S} \in \mathbb{R}^{N \times N}$ an enhanced self-similarity matrix. We denote two disjoint segments as: $\alpha = [s_1 : t_1] \subseteq [1 : N]$ and $\beta = [s_2 : t_2] \subseteq [1 : N]$ where $s_1 \leq t_1 < s_2 \leq t_2$. Let $|\alpha| := t_1 - s_1 + 1$ and $|\beta| := t_2 - s_2 + 1$ denote their lengths, respectively. A *path* over α having path length L is a sequence $p^\alpha = ((n_1, m_1), \ldots, (n_L, m_L))$ of cells $(n_\ell, m_\ell) \in [1 : N]^2$, $\ell \in [1 : L]$, satisfying $m_1 = s_1$ and $m_L = t_1$ (boundary condition) and $(n_{\ell+1}, m_{\ell+1}) - (n_\ell, m_\ell) \in \Omega$ (step size condition). We use

$$\Omega = \{(1, 2), (2, 1), (1, 1)\}, \tag{1}$$

which constrains the slope of a path within the bounds of $1/2$ and 2 (see [16]). The *score* of p^α is defined as $\sigma(p^\alpha) = \sum_{\ell=1}^{L} \mathcal{S}(n_\ell, m_\ell)$. For a path p, we associate two segments defined by the vertical projection $\pi_1(p) := [n_1 : n_L]$ and horizontal projection $\pi_2(p) := [m_1 : m_L]$. By definition we have $\pi_2(p^\alpha) = \alpha$. The projection of a path onto the vertical axis, $\pi_1(p^\alpha)$, is referred as an *induced segment* of a path p^α. Similarly, we introduce the notion of a path p^βover the segment β.

Extending the notion of a path family [7], we introduce a *joint path family* over α and β, which is a set

$$\mathcal{P}^{\alpha\beta} := \{p_1^\alpha, \ldots, p_U^\alpha, p_1^\beta, \ldots, p_V^\beta\} \tag{2}$$

of size $U + V$, consisting of paths p_u^α over α and paths p_v^β over β, where $u \in [1 : U]$ and $v \in [1 : V]$. Recall from Section 2 that the induced segments of a path family cannot overlap with each other. We also impose this constraint to the induced segments of a joint path family. In other words, we require that the set $\{\pi_1(p_1^\alpha), \ldots, \pi_1(p_U^\alpha), \pi_1(p_1^\beta), \ldots, \pi_1(p_V^\beta)\}$ consists of pairwise

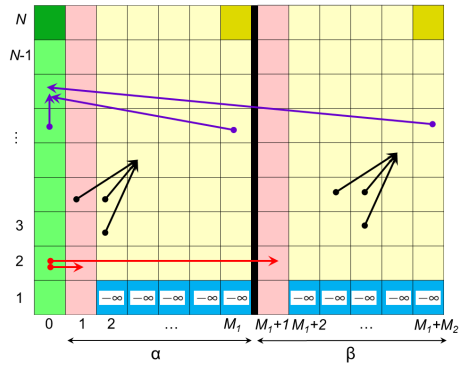

Fig. 3. Illustration of the optimization scheme in computing the accumulated score matrix D. The differently colored regions and arrows indicate the various step conditions as explained in the text.

disjoint segments. Next, extending the definition for the score of a path, we define the score for a joint path family as

$$\sigma(\mathcal{P}^{\alpha\beta}) := \sum_{u=1}^{U} \sigma(p_u^\beta) + \sum_{v=1}^{V} \sigma(p_v^\beta). \tag{3}$$

There are in general many possible joint path families over α and β. Among these path families, there exists an optimal path family of maximal score, defined as

$$\mathcal{P}^*_{\alpha\beta} := \underset{\mathcal{P}^{\alpha\beta}}{\operatorname{argmax}}\ \sigma(\mathcal{P}^{\alpha\beta}). \tag{4}$$

4.2. Optimization Scheme

Based on the optimization scheme introduced in [7], we now describe a modified algorithm that can efficiently compute an optimal joint path family. Let $X = (x_1, x_2, \ldots, x_N)$ be the feature sequence of the entire audio recording, $Y := (x_{s_1}, \ldots, x_{t_1})$ and $Z := (x_{s_2}, \ldots, x_{t_2})$ the feature sequences corresponding to α and β, respectively. Based on DTW (Dynamic Time Warping, see, e. g., [17, 16]), we use a modified version to simultaneously align paths between Y (or Z) and some sub-sequences of X, with the constraint that no overlaps between these sub-sequences of X are allowed. The goal is to determine the optimal alignment that defines the optimal joint path family of maximal score. Note that we impose the entire segments of α and β to be aligned with sub-sequences of X. Furthermore, to skip some sub-sequences of X which are neither similar to α nor to β, certain sections of X can be left completely unconsidered in the alignment.

To account for these constraints, we introduce some new steps that allow us to skip certain sections of X and to jump from the end to the beginning of the given segment α (or β). First, we define an $N \times (M_1 + M_2)$ submatrix $\mathcal{S}^{\alpha\beta}$ by taking the columns s_1 to t_1 and s_2 to t_2 of $\mathcal{S}$. Next, we introduce an accumulated score matrix D. By setting different step conditions for different regions in D, we realize the above mentioned constraints. To this end, we define $D \in \mathbb{R}^{N,(1+M_1+M_2)}$, (with rows indexed by $[1:N]$ and columns indexed by $[0:(M_1+M_2)]$), by the following recursion:

$$D(n,m) = \mathcal{S}^{\alpha\beta}(n,m) + \max\{D(i,j) \mid (i,j) \in \Phi(n,m)\} \tag{5}$$

for $n \in [2:N]$ and $m \in [2:M_1] \cup [(M_1+2):(M_1+M_2)]$ (the yellow regions in Figure 3), where

$$\begin{aligned}\Phi(n,m) = \{(n-i, m-j) \mid (i,j) \in \Omega\} \quad \cap \\ \{[1:N] \times ([1:(M_1-1)] \cup [(M_1+1):(M_1+M_2-1)])\}\end{aligned} \tag{6}$$

denotes the set of possible predecessors (see the black arrows in Figure 3). So far, these definitions are used for computing the accumulated score during path alignments.

Then, we need to allow for the possible skipping of sections in X. Similar as in [7], the first column of D indexed by $m = 0$ plays a special role, and it is recursively defined as:

$$D(n,0) = \max\{D(n-1,0), D(n-1,M_1), D(n-1,M_1+M_2)\} \tag{7}$$

for $n \in [2:N]$ and initialized by $D(1,0) = 0$ (see the green region and the purple arrows in Figure 3). The term $D(n-1,0)$ enables the algorithm to move upwards without accumulating any (possibly negative) score, thus allows for skipping some sections of X without penalty (negative score). Note that the term $D(n-1,M_1)$ closes up a path over α, and the term $D(n-1,M_1+M_2)$ closes up a path over β. The later two terms ensure that the entire segment α or β is aligned to the sub-sequence of X, and the next possible sub-sequence of X to be aligned does not overlap with the previous aligned sub-sequence.

After introducing how we align a path and close a path, now we present how we start a new path. This is realized by controlling the column for $m = 1$ and $m = M_1 + 1$ in D which correspond to the beginning of α and β,respectively. We define the new constraints as:

$$D(n,1) = D(n,0) + \mathcal{S}^{\alpha\beta}(n,1) \tag{8}$$

$$D(n,M_1+1) = D(n,0) + \mathcal{S}^{\alpha\beta}(n,M_1+1) \tag{9}$$

for $n \in [1:N]$ (see the pink regions and the red arrows in Figure 3).

Finally, to initialize the D matrix, we set $D(1,m) = -\infty$ for $m \in [2:M_1] \cup [(M_1+2):(M_1+M_2)]$ (see the blue region in Figure 3), which forces the first path to start either with the first element of α or the first element of β. Based on these definitions, the score of an optimal joint path family is then given by

$$\sigma(\mathcal{P}^*_{\alpha\beta}) = \max\{D(N,0), D(N,M_1), D(N,(M_1+M_2))\} \tag{10}$$

(see the shadowed cells in the top row in Figure 3). The first term $D(N,0)$ reflects the situation that the optimal path family may skip the alignment with final section of X, and the later two terms $D(N,M_1)$ and $D(N,(M_1+M_2))$ ensure that for the other cases, the last path is either aligned with the entire segment α or with the entire segment of β. The associated optimal joint path family $\mathcal{P}^*_{\alpha\beta}$ can be derived from D by using a back-tracking algorithm as in classical DTW (see [16, Chapter 2]).

4.3. Joint Fitness Measure

We now define the new joint fitness measure. Similar as in [7], we associate the joint fitness measure for a pair of segments with their optimal joint path family. We consider two properties of the joint path family, which are the score and the coverage. In addition, the contribution of a segment itself to the score and the coverage need to be excluded, otherwise the segment representing the entire audio file will get the maximum score and coverage, which we do not want. First, we consider the score measurement. Let $\mathcal{P}^*_{\alpha\beta} = \{p_1^\alpha, \ldots, p_U^\alpha, p_1^\beta, \ldots, p_V^\beta\}$ be an optimal path family for a pair of segments α and β. Then, the *normalized score* $\bar{\sigma}(\alpha,\beta)$ is defined as:

$$\bar{\sigma}(\alpha,\beta) := \frac{\sigma(\mathcal{P}^*_{\alpha\beta}) - |\alpha| - |\beta|}{\sum_{u=1}^{U} L_u^\alpha + \sum_{v=1}^{V} L_v^\beta} \tag{11}$$

where L_u^α and L_v^β are the lengths of the respective paths p_u^α and p_v^β from the optimal joint path family. Second, we consider some kind

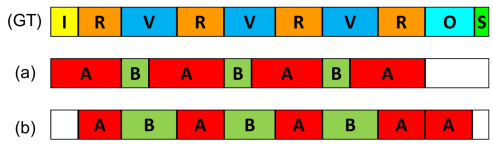

Fig. 4. Illustration of the structure segmentation result derived from the two proposed approaches for Beatles song "Devil In Her Heart". **(GT)** Ground truth structure segmentation. **(a)** Estimation result by the iterative approach. **(b)** Estimation result by the joint approach.

of coverage measure. Let $\mathcal{A}^*_{\alpha\beta} := \{\pi_1(p^\alpha_1), \ldots, \pi_1(p^\alpha_U), \pi_1(p^\beta_1), \ldots, \pi_1(p^\beta_V)\}$ be the segment family induced by $\mathcal{P}^*_{\alpha\beta}$, and let $\gamma(\mathcal{A}^*_{\alpha\beta})$ be the coverage of this induced segment family, which is defined as:

$$\gamma(\mathcal{A}^*_{\alpha\beta}) = \sum_{u=1}^{U} |\pi_1(p^\alpha_u)| + \sum_{v=1}^{V} |\pi_1(p^\beta_v)|. \tag{12}$$

Then, the *normalized coverage* $\bar{\gamma}(\alpha, \beta)$ is defined as:

$$\bar{\gamma}(\alpha, \beta) := \frac{\gamma(\mathcal{A}^*_{\alpha\beta}) - |\alpha| - |\beta|}{N}. \tag{13}$$

Finally, combining the normalized score and the normalized coverage, we define the *joint fitness measure* for α and β to be their harmonic mean:

$$\varphi(\alpha, \beta) := 2 \cdot \frac{\bar{\sigma}(\alpha, \beta) \cdot \bar{\gamma}(\alpha, \beta)}{\bar{\gamma}(\alpha, \beta) + \bar{\sigma}(\alpha, \beta)}. \tag{14}$$

Similar as in [7], among all possible pairs of segments of an audio recording, the double thumbnails are defined to be the pair of segments of maximal joint fitness:

$$(\alpha, \beta)^* := \underset{\alpha,\beta}{\operatorname{argmax}} \ \varphi(\alpha, \beta). \tag{15}$$

As an illustration of our joint approach, Figure 1 shows the estimated result of double thumbnails for Beatles song "Devil in Her Heart" as well as the optimal joint path family and its induced segment family. These induced segments are then transferred into the structure segmentation as can be seen in Figure 4b. Here, we see that the induced segments (denoted by the A and B sections) successfully reveal the V and R parts in the ground truth, respectively. As a comparison, we also shows the result of the iterative approach computed for this song in Figure 4a. We can see that the iterative approach estimated a thumbnail which is too long in the first round, thus result in the problematic estimation of the thumbnail in the second round.

5. EVALUATION

We now describe our systematic evaluation. So far we have not seen a public standard evaluation for double thumbnails. Therefore, in order to be comparable with other algorithms, we use one standard MIREX evaluation measure for music structure segmentation [19], which is the pairwise frame clustering evaluation presented in F-measure (F), Precision (P) and Recall (R) [20]. Note that in this paper we use the full structure evaluation as a measure for illustrating the performance of our double thumbnailing technique. As an example scenario, we use the Beatles dataset which contains 180 recordings of "The Beatles" and the ground truth structure annotations [21]. For each recording, we apply the iterative approach and the joint approach [1], obtaining two kinds of double thumbnails as well as their induced segments. For both approaches, we treat the resulting induced segments as a kind of music structure segmentation and evaluate them using the above mentioned measure. In our experiments, we use a feature resolution of 2 Hz.

Approach	F	P	R
Serra [22]	0.71	0.68	0.79
Iterative	0.69	0.71	0.70
Joint	0.68	0.77	0.64
Max	0.74	0.79	0.73
UpperLimit	0.97	0.97	0.97

Table 1. Structure evaluation results using the pairwise frame clustering P/R/F values averaged on the Beatles dataset.

Table 1 shows the evaluation result for various approaches. For comparison, the first row shows the results of an state-of-the-art algorithm for full structure analysis suggested by Serra et al. [22], where he gets an F-measure of 0.71. Using our proposed methods, the iterative approach yields an F-measure of 0.69, and the joint approach gets an F-measure of 0.68. By individual inspection, we found that the joint approach outperforms the iterative approach far better for some of the songs, but works worse for some other songs, which is mainly due to over-segmentation. After that, in order to see what we can best achieve, we select for each song the better result of the two approaches, and average over all songs to generate the max possible result we can get. This yields an F-measure of 0.74 shown in the fourth row, which is roughly 0.05 higher compared to either of the two approaches. Such difference indicates that the two approaches actually behave differently on various songs. Also, it shows that if we select the appropriate approach for various songs, we can further improve in structure segmentation that outperform other algorithms. Finally, in the last row we present the theoretical upper limit of the result that using double thumbnails estimation. To this end, we extract sections of the two most repetitive parts from the ground truth as our "computed" result, and further evaluate them with the original ground truth which consists of all sections. This yield an F-measure of 0.97. Comparing to this upper limit, the results above indicate that all those approaches somehow hit a practical ceiling. In addition, this very high value also supports our assumption that for Beatles dataset (representative of popular music), most songs have only two main repeating sections. Therefore, our double thumbnail estimation fits well in popular music scenario.

In conclusion, we introduced two novel different approaches for estimating double thumbnails. By deriving repetitive sections of the double thumbnails, we also contribute to the full structure analysis especially for music pieces with two repetitive parts as main structures.

Acknowledgments: This work has been supported by the German Research Foundation (DFG MU 2686/5-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Fraunhofer-Institut für Integrierte Schaltungen IIS.

[1] In practice, it is too expensive to compute joint fitness values for all possible pairs of segments of an audio recording. Therefore, we use a sampling and refinement acceleration strategy as introduced in [18], to compute joint fitness values only for a limited number of segment pairs.

6. REFERENCES

[1] Roger B. Dannenberg and Masataka Goto, "Music structure analysis from acoustic signals," in *Handbook of Signal Processing in Acoustics*, David Havelock, Sonoko Kuwano, and Michael Vorländer, Eds., vol. 1, pp. 305–331. Springer, New York, NY, USA, 2008.

[2] Mark A. Bartsch and Gregory H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[3] Wei Chai and Barry Vercoe, "Music thumbnailing via structural analysis," in *Proceedings of the ACM International Conference on Multimedia*, Berkeley, CA, USA, 2003, pp. 223–226.

[4] Matthew Cooper and Jonathan Foote, "Automatic music summarization via similarity analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002, pp. 81–85.

[5] Masataka Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1783–1794, 2006.

[6] Mark Levy, Mark Sandler, and Michael A. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006, pp. 13–16.

[7] Meinard Müller, Nanzhu Jiang, and Peter Grosche, "A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, no. 3, pp. 531–543, 2013.

[8] Matthew Cooper and Jonathan Foote, "Summarizing popular music via structural similarity analysis," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2003, pp. 127–130.

[9] Roger B. Dannenberg and Ning Hu, "Pattern discovery techniques for music audio," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002, pp. 63–70.

[10] Namunu C. Maddage, "Automatic structure detection for popular music," *IEEE Multimedia*, vol. 13, no. 1, pp. 65–77, 2006.

[11] Geoffrey Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 35–40.

[12] Christophe Rhodes and Michael A. Casey, "Algorithms for determining and labelling approximate hierarchical self-similarity," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 41–46.

[13] Jouni Paulus, Meinard Müller, and Anssi P. Klapuri, "Audio-based music structure analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 625–636.

[14] Meinard Müller and Sebastian Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, USA, 2011, pp. 215–220.

[15] Meinard Müller, Nanzhu Jiang, and Harald Grohganz, "SM Toolbox: MATLAB implementations for computing and enhancing similiarty matrices," in *Proceedings of the AES Conference on Semantic Audio*, London, GB, 2014.

[16] Meinard Müller, *Information Retrieval for Music and Motion*, Springer Verlag, 2007.

[17] Roger B. Dannenberg and Ning Hu, "Polyphonic audio matching for score following and intelligent audio editors," in *Proceedings of the International Computer Music Conference (ICMC)*, San Francisco, USA, 2003, pp. 27–34.

[18] Nanzhu Jiang and Meinard Müller, "Towards efficient audio thumbnailing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 5192–5196.

[19] "Music information retrieval evaluation exchange (mirex), task structure segmentation," `http://www.music-ir.org/mirex/wiki/2014:Structural_Segmentation`, Accessed: 2010-10-01.

[20] Mark Levy and Mark Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 318–326, 2008.

[21] Matthias Mauch, Chris Cannam, Matthew E.P. Davies, Simon Dixon, Christopher Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler, "OMRAS2 metadata project 2009," in *Late Breaking Demo of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, 2009.

[22] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014.