Musical Structure Analysis using Image Segmentation Networks

Christopher Uzokwe
Drexel University
cnu25@drexel.edu

Dr. Youngmoo Kim
Drexel University
ykim@drexel.edu

ABSTRACT

To date, most research in automatic musical structure analysis and segmentation from audio has been conducted using established audio features with boundary detection and clustering methods [1]. In other Music IR tasks, we have seen significant advances in performance using deep learning, specifically convolutional neural networks -- networks pioneered in image understanding problems. Several experiments have examined the “images” of musical self-similarity [2,3], computed from acoustic features (e.g., STFT, mel-spectrogram, and mel-frequency cepstrum coefficients), which can provide a compelling visual representation of overall musical structure, but remain difficult to automatically interpret. In a recent trial, we directly trained and applied a robust visual object detection network (Faster R-CNN) on layered self-similarity matrices (SSMs) to identify musical segments (intro, verse, chorus) of a song, and highlight them using bounding boxes as done in visual object detection. We performed an initial assessment of this task using an “off-the-shelf” implementation, Facebook AI Research’s Detectron2 [4], of the Faster R-CNN [5] recognition method on SSMs generated using audio features as separate “color channels”. Preliminary results of the system and general knowledge of this task provide some indications that visual object detection methods examining the entire SSM may allow us to characterize musical sections and reveal indicative features identifying different segments.

We are continuing this work by modifying our initial implementation built using Detectron2’s framework to more accurately replicate experiments conducted in [2,3]. This is motivated by the desire achieve results that can be compared to other deep learning approaches, so we could validate our system’s performance and approach. The Detectron2 framework is also an open source project that facilitates reproducible research that could be extended to musical structure analysis from a deep learning perspective, a similar motivation to that in [1], and a shortcoming to the results comparison in [2]. Additionally, we will construct a Faster R-CNN system with more specific constraints aligning with our audio and musical objectives.

Homogeneity, Novelty, and Similarity in the Self Similarity Matrix: The self-similarity matrix provides visual cues about what sections in a song share similar features at different times (similarity), what sections are comprised of similar features (homogeneity), and points of great inflection in the song (novelty). A Region Proposal Network (described below) gives us the ability to cross reference proposed features on the diagonal of the matrix (homogeneity, novelty), with those on the off diagonal (similarity). Each feature can work to establish a structure profile for a song, artist, or genre, rather than a baked in, yet ever changing verse, chorus structure.

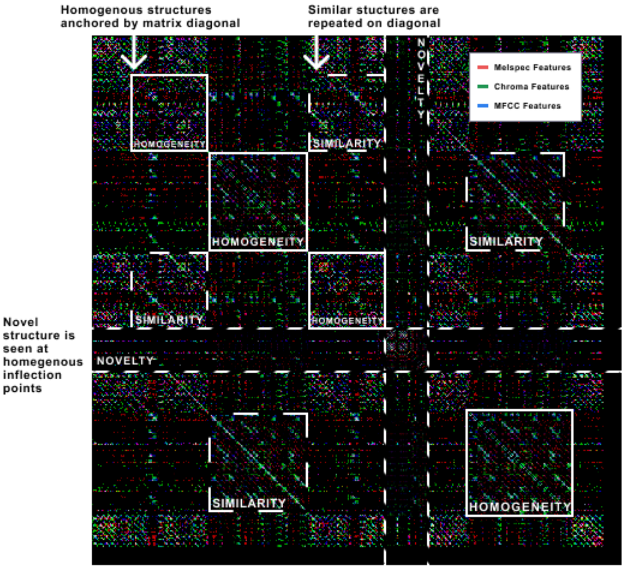

Figure 1. Self-similarity matrix created from the MFCC, Chroma, and Mel-Spectrogram features of “I should have known better” by the Beatles. Homogenous sections have similarity replicas which exist in the off diagonal sections, and general novelty is seen across the length of the SSM.

Faster R-CNN with RPN: The Faster R-CNN is a Region-based Convolutional Neural Network (R-CNN) which uses a Region Proposal Network (RPN) as an attention mechanism to make computationally nearly cost free region proposals (object bounds, objectness scores), which are then shared with the R-CNN for classification or regression. Currently, the objects proposed can be of any aspect ratio at multiple different scales. Existing CNN-based approaches for automatic structure analysis are essentially fed region proposals that only exist along the diagonal (short-time similarity) of the SSM at fixed sizes [2,3]. Although this method is useful in detecting novelty in between frames, it is unable to incorporate overall musical structure, which includes the homogeneity in longer segments, as well as the indicators of similarity in the off diagonal regions (similar music occurring at different times). Simultaneous detection of different elements could aid in making statistical decisions such as in [6,7].

Current Progress: We have used a subset of 445 annotations from the Structural Analysis of Large Amounts of Music Information (SALAMI) dataset, which contains annotations of musical boundaries from recordings from different genres. After pre-computing audio features to generate each self-similarity matrix, we turn our musical section boundaries into bounding boxes using a section's start time (multiplied by our spectrogram's fps) as x1, y1, and the duration to calculate x2, y2. The preliminary trial used only segments marked as intro, verse, and chorus, and these features were trained jointly across all songs in the SALAMI dataset in a 5-fold cross validation run. Bounding boxes formed had an average normalized area overlap (intersection over union) of 0.859 (Std Dev 0.01). Although these results were achieved using an off-the-shelf object detection system, in addition to a fundamentally flawed training method (intros, verses, and choruses do not look or sound the same across different songs), the network is still able to detect indicators of homogeneity. Previous works have detected novelty in sections [2,3], so the Faster R-CNN may need further formation to include indications of similarity and create relevant encodings.

Next Steps: We have presented a brief initial survey of how more robust visual object detection systems may be a valid approach to understanding the segmentation of musical structure using SSMs. Further exploration of this task will include testing variations of the Detectron2 framework to replicate deep learning approaches done in the past, as well as highlight the nature of the reproducibility. Additionally, we are developing a Faster R-CNN implementation with constraints more suitable for identifying regions of homogeneity, novelty, and similarity.

REFERENCES

[1] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research.," in *ISMIR*, 2016, pp. 547–553.

[2] A. Cohen-Hadria and G. Peeters, "Music structure boundaries estimation using multiple self-similarity matrices as input depth of convolutional neural networks," *AES International Conference Semantic Audio* 2017, Jun 2017 Erlangen, Germany. hal-01534850.

[3] K. Ullrich, J. Schlüter, and T. Grill, "Boundary Detection in Music Structure Analysis using Convolutional Neural Networks.," in *ISMIR*, 2014, pp. 417–422.

[4] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, *Detectron2*. https://github.com/facebookresearch/detectron2, 2019.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] A. Maezawa, "Music Boundary Detection Based on a Hybrid Deep Model of Novelty, Homogeneity, Repetition and Duration," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 206–210, doi: 10.1109/ICASSP.2019.8683249.

[7] G. Shibata, R. Nishikimi, E. Nakamura, and K. Yoshii, "Statistical Music Structure Analysis Based on a Homogeneity-, Repetitiveness-, and Regularity-Aware Hierarchical Hidden Semi-Markov Model.," in *ISMIR*, 2019, pp. 268–275.