Musical Structure Analysis using Image Segmentation Networks

Christopher Uzokwe {cnu25@drexel.edu}, Dr. Youngmoo Kim {ykim@drexel.edu}

DREXEL UNIVERSITY
ExCITe Center
Expressive and Creative Interaction Technologies
METlab
Music . Entertainment . Technology

Overview

We explore the musical structure analysis task using convolutional neural networks developed for object detection and classification. Self similarity matrices generated from audio features are stacked to form RGB layers of an "image." For initial exploration, an off the shelf implementation of the Faster R-CNN model is applied through Facebook AI Research's Detectron2 framework.

Motivation

This exploration is motivated by recent advances in deep learning networks that are applied to music information retrieval tasks, as well as image recognition tasks.

Region based detection on an SSM is not new, but previous works have focused on measures of novelty. What else can we detect?

Can robust, open source image detection systems help to facilitate reproducible research in the musical structure analysis field?

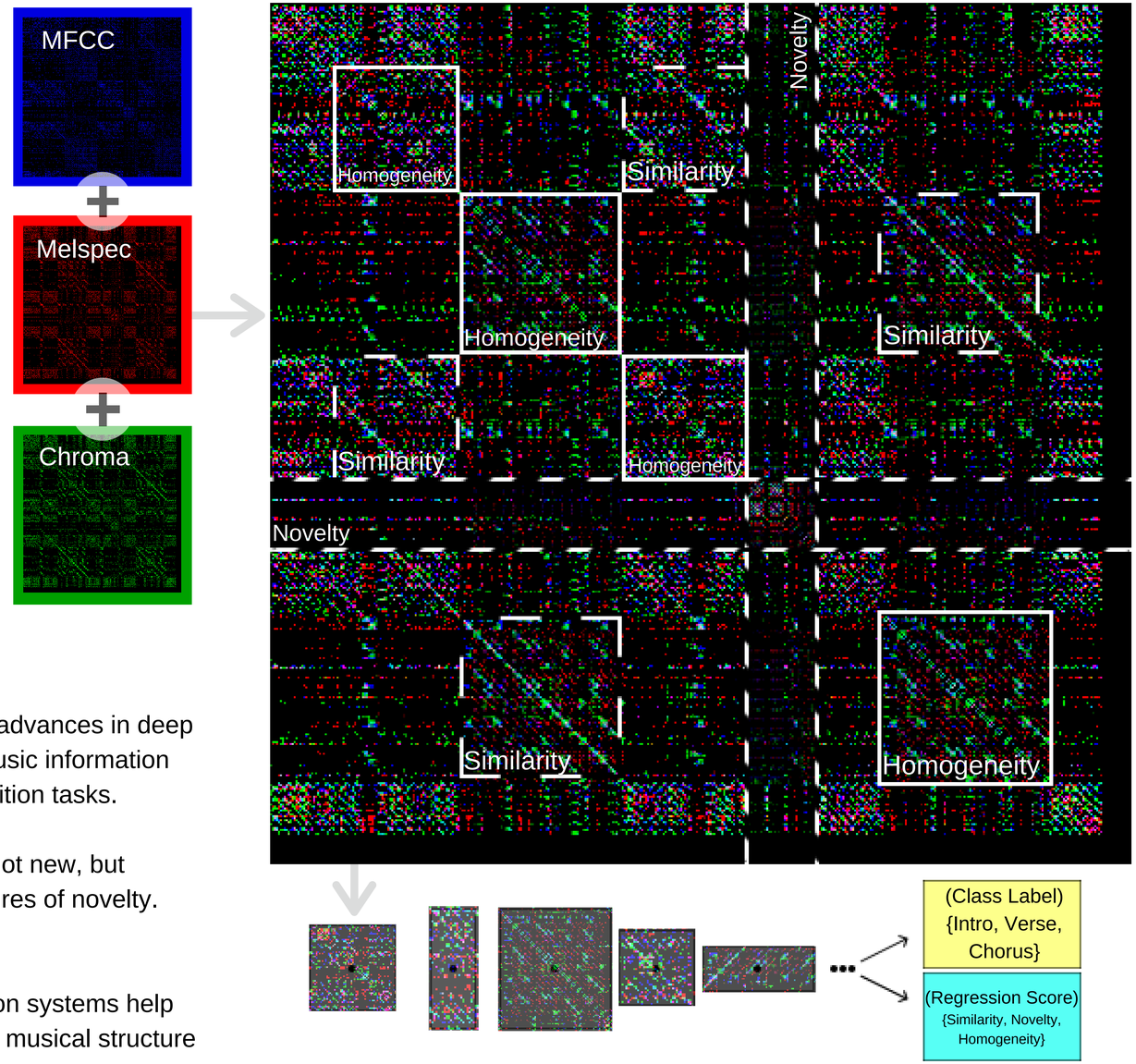

Figure 1: Evaluation pipeline of "I Should Have Known Better," by The Beatles. "Image" constructed from audio features has regions proposed through RPN, identified with CNN.

Approach

The Faster R-CNN is fed self similarity matrices with ground truth bounding boxes constructed from the boundary annotations of the audio. For simplicity, we only use boundaries labeled as intro, verse, or chorus. Boundary annotations come from a subset of 445 songs in the SALAMI dataset, which contains structural analyses of popular music from varying genres.

Observations

- Intros, verses, and choruses are different in different songs - training objective needs refinement
- Predicted bounding box normalized area overlap with ground truth: 0.859 (Std Dev 0.01). The network bounds homogenous regions, but doesn't incorporate novel inflection points
- Region proposals are currently made at any aspect ratio at any point of the image. We could constrain these variables for better or more targeted predictions

Next Steps

We are continuing this work by reapplying the Faster R-CNN from Detectron2 with a training method that better replicates previous works to generate comparable results, and assess reproducibility. We will also reconstruct our region proposal network to better suit the task of identifying characteristics in different regions.