Learning audio-visual correspondences for music-video recommendation

Thomas Langlois

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2019-14
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-14.html

May 1, 2019

MuVi Score Dataset: Modeling Human Music-Video Pairing Preferences

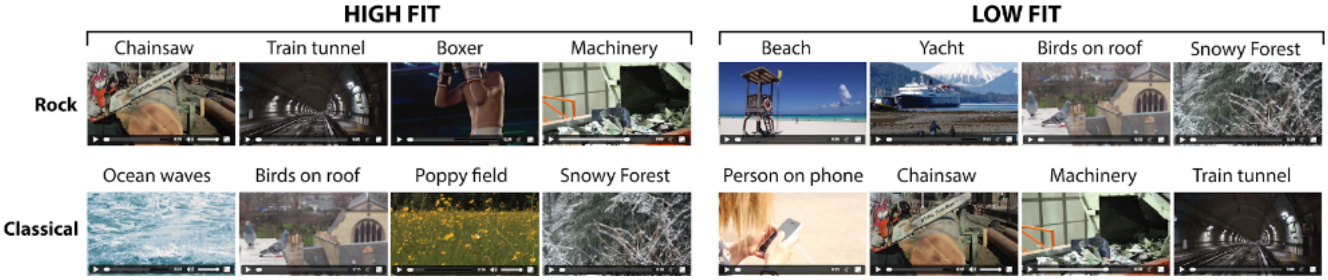

Figure 1: Videos of natural scenes are better set to classical music, while urban scenes are preferred paired with rock music. We generated a novel dataset of videos paired with a variety of musical genres for the purposes of learning human audio-visual correspondences in the domain of music and video.

1. Introduction

Even before the advent of sound film, when advances in sound technology enabled the reliable synchronization of recorded sound with motion pictures, it was not uncommon for live orchestras to play accompanying music for silent films. In fact, music and other performance arts share a long history spanning many cultures, and appear to have been joined at the hip for most of recorded history. Today, music is an integral part of nearly all multimedia art, ranging from movies, music-videos, to video games and dance. Despite the importance that music has played in these media, it remains a mystery why humans prefer one music-video pairing over another nor have these preferences been accurately modeled computationally.

In this paper, we address this problem by creating a dataset of short video segments combined with different music tracks and obtained human ratings for the pairing. We also describe a model — a three-stream audio-visual convolutional network — that predicts these human judgments. Our primary contribution is a novel dataset of videos paired with a variety of music samples, for which we obtained human aesthetic judgments (ratings of the degree of "fit" between the music and video).

While multiple lines of work have developed methods to pair music to video based on features that were conceived a priori, such as semantic labels, emotional labels [12], spatial-temporal dynamics [9], and some low-level visual and acoustic features [11], [8], none have attempted a data-driven approach using actual human judgments to learn the most useful representations for such a task automatically using contemporary machine learning methods. In addition, while some of this work relies on heuristics with some basis in known multi-modal perceptual processes from psychology and cognitive-neuroscience, they assume that human audio-visual correspondences rely solely on shared spatial temporal dynamics of video and audio content. Our hope is that this novel dataset can serve as a springboard for a new vein of research into human audio-visual correspondences in the context of music and video, where no assumptions are made from the outset about which audio-visual features are implicated in human cross-modal correspondences. We also sketch out some approaches to learning these correspondences directly from the data in an end-to-end manner using contemporary machine learning methods, and present some preliminary results.

2. Dataset

Generating a suitable dataset to study human audio-visual representations in the context of music and video poses several challenges. Mainstream movies are well-known and biased from the outset. Also, they seldom contain negative examples, as the video content and the music were engineered to match as well as possible. Alternatives, such as music-videos are also problematic, since they typically show a performer synchronized to the audio, singing on a stage. Our challenge was to generate a novel dataset that could ostensibly have been obtained from an actual movie studio, but without the problems described above.

2.1. Videos

We collected two kinds of datasets. The first asked users to rate videos on a scale from 1-7, providing an absolute ranking of music-video pairs. This dataset consisted of 1,061 high production-quality stock videos. The content of the videos ranged from nature scenes (mountain ranges, animals in the wild, oceans, beaches, snowy forests) to urban scenes (amusement parks, city streets, construction sites, trains, airports, people playing sports). A few representative thumbnails are shown in Figure 3. We randomly sampled 5 second clips from the videos and generated all possible pairwise combinations of these videos with 5 second samples of 75 songs (see below for details), yielding 79,575 unique music-film combinations.

The second dataset asked users to select which of two au-

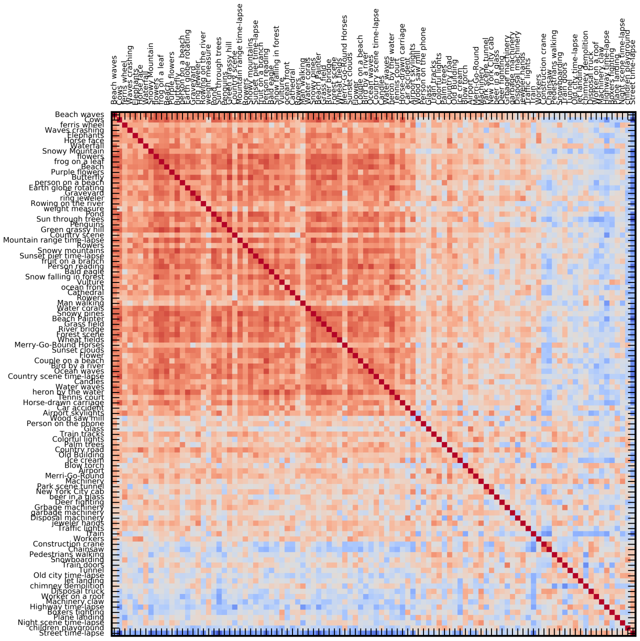

Figure 2: Hierarchical correlation clustering of a representative sample of videos and their average fit judgments to a set of different musical genres. Clustering reveals similar distributions of musical "fit" judgments for natural scenes (e.g., beach scenes, forest scenes, pastures, sunsets, and mountain ranges, cows grazing), in the top left quadrant, that differ from distributions of fit judgments for videos corresponding to urban scenes (e.g., machinery, demolition, cityscapes, boxers fighting). Natural scenes are preferred with classical music, country, ambient soundtracks and smooth jazz, while urban scenes are preferred with rock, metal, and house techno, and bebop jazz genres.

dio tracks went best with a given video. We scraped 21,159 videos scraped from Flickr, filtering for HQ videos. Each of these videos was then paired with a random sample of two unique music tracks from the full set of 4,756 musical samples we obtained from the Million Song Dataset [3].

2.2. Human judgments

Since the objective of generating a large set of short videos paired with music is ultimately to discover what multi-modal features humans use to make good audio-visual pairings, we obtained human annotations regarding the degree of fit between the video and musical content of each of the 79,757 music-video combinations. To do this, we created an experiment in Amazon Mechanical Turk in which workers were instructed to: "Rate how well the music and the video fit. Workers could make a rating on a 7-point scale, with 1 meaning Extremely poor fit, and 7 meaning Extremely good fit. We obtained 9 fit ratings for each of the unique combinations. Figure 2 shows a hierarchical clustering of a representative sample of the videos used in the ratings experiment, which reveals clusters of musical fit profiles for different kinds of video content. it reveals that nature scenes (such as Beach scenes, forest scenes, pastures, sunsets, and mountain ranges, cows grazing) are typically preferred paired with audio soundtracks of classical music (classical piano, violin cello), country, ambient

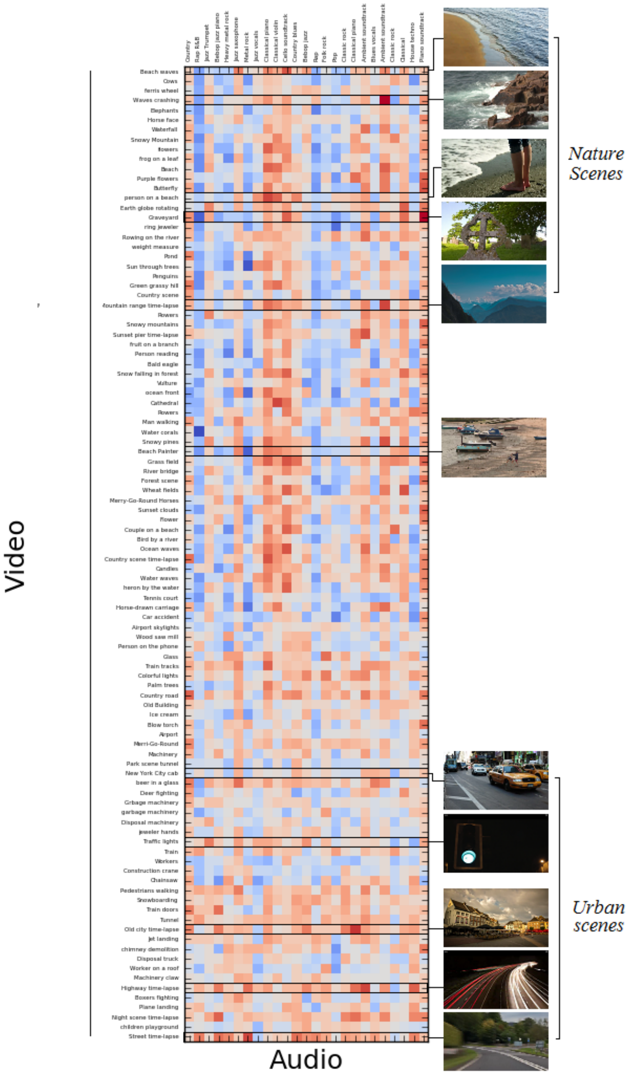

Figure 3: Average ratings of fit judgments for 25 genres of music combined with a representative sample of 100 videos. A few qualitative examples are shown: A beach scene, and crashing waves on a cliff are preferred paired with ambient soundtracks and classical music, a graveyard scene is best paired with classical piano music, and a time-lapse video of a car speeding through a street is well matched with heavy metal music. Rows are hierarchically clustered, so that videos (rows) with similar music preferences are closer together

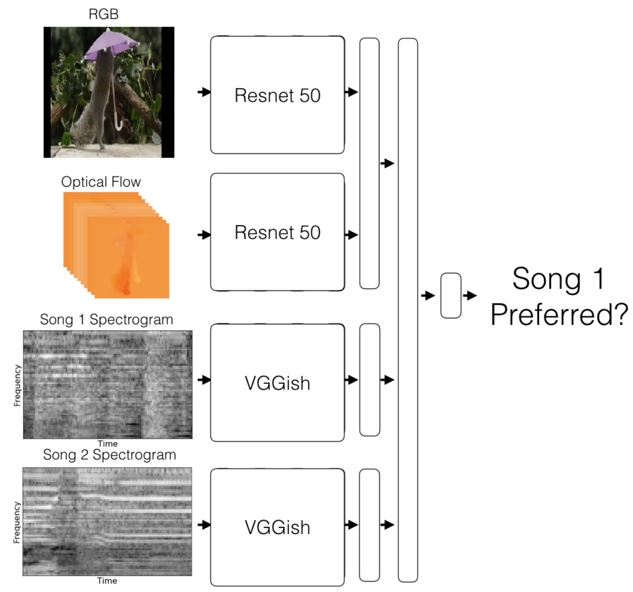

Figure 4: The model used in the experiments inspired by [1] [7] Both resnet 50 models are pre-trained on ImageNet and the VGG-ish network is pre-trained on Acoustic Event Detection task from the AudioSet dataset [5].

soundtracks, and smooth jazz, while urban scenes (machinery, demolition, cityscapes, boxers fighting), are typically preferred paired with alternative rock, house techno, heavy metal, and bebop jazz. Figure 3 reveals a handful of qualitative examples: A sombre video of a graveyard is preferred with a slow piano music track, while a sped-up time-lapse video of a car careening through a street is well matched with heavy metal rock.

While these "absolute" judgments of match quality are useful for analysis, they are not well-suited to training computational models, due to the lack of calibration between subjects (e.g. subjects may differ in their average rating). Therefore, we also collected relative comparisons between pairings using a two-alternative-forced choice experiment (2AFC). To do this, we generated a set of music videos by combining 21,159 videos (sampled from Flickr) with a random pair of musical samples from our music dataset. We then asked humans to say which of the two music tracks better fit the video.

3. Experiments

To test the usefulness of our data, we trained a multimodal neural network to reproduce the relative rankings. We took inspiration from the the audio-visual embedding network proposed in [2]. However, we make three modifications. First, we replace the VGG visual network [10] with an equivalent ResNet architecture [6] also pre-trained on ImageNet. Next, we add a second stream to the vision embedding, also initialized from a pre-trained ResNet architecture, to process optical flow. We modify the first convolution to take in a 20-dimensional volume rather than the typical 3-D RGB volume by replicating the filter weights. Then we replace the audio portion of the network with VGG-ish [7] — a VGG-style model pre-trained on the Acoustic Event Detection task from [7] using the AudioSet dataset [5]. Finally, we pass both music tracks we'd like to compare through the same audio network, concatenate those vectors with the visual embedding of the network and pass the concatenated representation through a hidden fully-connected layer to the output which predicts whether one sound is more preferred than the other when passed through the network. Figure 4 diagrams the model.

We also asked whether the results would improve with additional temporal context, rather than a static frame. To address this, we replaced the visual ResNet model with a 3D convolutional network: I3D [4]. For this, we used network weights pretrained on Kinetics and ImageNet datasets, and we used only the flow stream to reduce the model size. We found that this model significantly outperformed the ResNet-based model, suggesting that temporal analysis is useful for this task.

3.1. Results

Table 1: Audio-visual CNN Results

Model	Test Accuracy
Random Guessing	50%
Two-Stream ResNet50	53.2% ± 1.62
I3D	55.94% ± 1.61

Our models perform slightly better than chance and the temporal context of I3D seems to furhter improve the results. We anticipate that these numbers will improve with the inclusion of more human ratings.

References

[1] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE, 2017. 3

[2] R. Arandjelović and A. Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017. 3

[3] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Ismir*, volume 2, page 10, 2011. 2

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017. 3

[5] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 3

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. 3

[8] Z. Liao, Y. Yu, B. Gong, and L. Cheng. Audeosynth: music-driven video montage. *ACM Transactions on Graphics (TOG)*, 34(4):68, 2015. 1

[9] R. Macrae, X. Anguera, and N. Oliver. Muvisync: Realtime music video alignment. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 534–539. IEEE, 2010. 1

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[11] J. Wang, E. Chng, C. Xu, H. Lu, and Q. Tian. Generation of personalized music sports video using multimodal cues. *IEEE Transactions on Multimedia*, 9(3):576–588, 2007. 1

[12] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40, 2012. 1