Distant Viewing Toolkit: A Python Package for the Analysis of Visual Culture

Taylor Arnold[1] **and Lauren Tilton**[2]

1 University of Richmond, Department of Mathematics and Computer Science **2** University of Richmond, Department of Rhetoric and Communication Studies

Links

- Repository
- Documenation
- Project

Compiled: 04 August 2019

Summary

Moving images have served as a dominant form of cultural expression in the U.S. since the beginning of the 20th century. However, with currently available tools, the formal analysis of moving images has been restricted to close analyses of a relatively small set of works. The project "Distant Viewing Toolkit (DVT) for the Cultural Analysis of Moving Images" allows scholars to work with large-scale collections by building an open source software library to facilitate the algorithmic production of metadata summarizing the content (e.g., people/actors, dialogue, scenes, objects) and style (e.g., shot angle, shot length, lighting, framing, sound) of time-based media. The software allows scholars to explore media in many forms, including films, news broadcasts, and television, revealing how moving images shape cultural norms. To illustrate DVT's ability to address humanities questions, the project will conduct case studies and extensive testing in cooperation with a group of scholars.

The DVT software library addresses the challenges of working with moving images by summarizing media objects through the automated detection of stylistic and content-driven metadata. It algorithmically approximates the ways in which humans process moving images by identifying and tracking objects, people, sound, and dialogue. As a result of advances over the past two years in deep learning and computer vision, it is now possible to build models capable of automatically performing these annotation tasks with human-like accuracy (He 2016; Szegedy 2017). These advances, combined with the recent increased interest in and access to large corpora of moving images, make this the perfect time for building an automated annotation tool specifically designed for application in the humanities.

The DVT software library will work by allowing users to input raw media files in a variety of formats. The input files will then be analyzed to detect the following features: (1) the dominant colors and lighting over each shot; (2) time codes for shot and scene breaks; (3) bounding boxes for faces and other common objects; (4) consistent identifiers and descriptors for scenes, faces, and objects over time; (5) time codes and descriptions of diegetic and non-diegetic sound; and (6) a transcript of the spoken dialogue. These features serve as building blocks for the analysis of moving images in the same way words are the foundation for text analysis. From these extracted elements, higher-level features such as camera movement, framing, blocking, and narrative style can be derived and analyzed.

DVT offers two output formats. The first provides data stored as a self-contained interactive website. The page can be opened locally in any web browser and requires no technical programming expertise from the user. This format provides tools to explore and visualize the extracted information. The second output format consists of a collection of plain-text JSON files. These files are optimal for technical users looking to integrate the output into larger

analytic pipelines such as building a visual search interface within a public-facing archive. In either output format, the generated metadata is significantly smaller in size compared to the raw media files, making it easy to share and publish the resulting data files.

In order to extract the metadata elements, DVT will make direct use of several specific deep learning frameworks and models. The toolkit utilizes the architecture of three open source programming libraries: dlib (King 2009), ffmpeg (Tomar 2006), and TensorFlow (Abadi et al. 2016). Within these frameworks, novel computer vision and sound processing algorithms extract the required features. Specifically, the project draws from OpenFace (Amos et al. 2016) for face detection; YOLO9000 for object detection (Redmon 2017); the Places-CNN (Zhou et al. 2016) for scene detection; Colorization (Zhang et al. 2016) for working with black and white images; GOTURN for object tracking (Held 2016); and CMUSphinx (Lamere 2003) for converting sound to text. These specific algorithms were chosen due to their open-source licenses, use of the most up-to-date techniques, and the institutional support behind the algorithms at CMU, MIT, and Berkeley. Our work in building DVT consists in modifying and stitching together these six models for our specific humanities-centric needs. For example, only one of these models works directly with moving images, taking only still images as inputs, and only one is able to process black and white images. Our toolkit will extract individual frames, colorize if necessary, apply each algorithm to the frames, and then intelligently combine the results into a single cohesive structure.

The six models we are building in the DVT software library will allow for the use of new, domain specific data to improve the performance of a generically trained algorithm. This process, known as transfer learning, is one of the key reasons for the popularity of deep learning in machine learning. A major part of building the DVT library will be applying transfer learning to tweak the open-source computer vision algorithms to better function on moving images. This will be done by first hand-labeling a training set of 10,000 still frames with information about common objects and characters found in each frame (tasks 3 & 4). Likewise, sound and shot break information will be hand-recorded from several hours of raw material (tasks 2 & 5). Time coded closed captioning is available for some of our source materials and can be used to create training data for speech recognition (task 6). These hand labeled datasets can be used to apply transfer learning to the base model, updating them for our specific humanities-focused application tasks. Putting the models together, along with extensive documentation, yields the DVT software library, capable of creating summary metadata directly from raw media files.

See (Arnold & Tilton, 2019; Arnold et al., 2019).

Acknowledgements

The Distant Viewing Toolkit is supported through a Digital Humanities Advancement Grant from the National Endowment for the Humanities (HAA-261239-18).

References

Arnold, T. B., & Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. *Digital Scholarship in the Humanities*. doi:10.1093/digitalsh/fqz013

Arnold, T. B., Tilton, L., & Berke, A. (2019). Visual style in two network era sitcoms. *Cultural Analytics*. doi:10.22148/16.043