CA Journal of Cultural Analytics

Visual Style in Two Network Era Sitcoms

Taylor Arnold, Lauren Tilton, and Annie Berke

07.19.19

Peer-Reviewed By: Richard Rogers

Clusters: Data

Article DOI: 10.22148/16.043

Dataverse DOI: https://doi.org/10.7910/DVN/S84TSX

Journal ISSN: 2371-4549

Cite: Taylor Arnold, Lauren Tilton, and Annie Berke, "Visual Style in Two Network Era Sitcoms," Journal of Cultural Analytics. July 19, 2019.

Extensive scholarship in media studies has established how formal elements of moving images—such as camera angles, sound, and framing—reflect, establish, and challenge cultural norms. Prior computational analyses attempting to analyze these elements have primarily relied on summarizing relatively low-level features. Beginning with the early work of Barry Salt, one particularly prominent metric is the distribution of median shot lengths.[1] Works on shot length include Yuri Tsivian's Cinemetrics project,[2] Arclight,[3] and Jeremy Butler's ShotLogger.[4] Other computational analyses examine the aggregation of language and average

[1] Barry Salt, "Statistical Style Analysis of Motion Pictures," *Film Quarterly* 28, no. 1 (1974): 13.

[2] Yuri Tsivian and Gunars Civjans, *Cinemetrics: Movie Measurement and Study Tool Database*, 2005.

[3] Charles R Acland and Eric Hoyt, *The Arclight guidebook to media history and the digital humanities* (Reframe Books, 2016).

[4] Jeremy Butler, "Statistical Analysis of Television Style: What Can Numbers Tell Us About TV Editing?" *Cinema Journal* 54, no. 1 (2014): 25-44.

shot color,[5] image compositions,[6] and analysis of film scripts.[7] These projects have demonstrated the feasibility of distributing extracted metadata from copyrighted materials and the power of computational techniques in accessing useful information over a large collection of moving images. However, there is much to be learned from other extractable metadata beyond shot detection and much to be studied in audiovisual media that can only be discovered through computer vision.

This essay shows how face detection and recognition algorithms, applied to frames extracted from a corpus of moving images, can capture formal elements present in media beyond shot length and average color measurements. Locating and identifying faces makes it possible to algorithmically extract time-coded labels that directly correspond to concepts and taxonomies established within film theory. For example, knowing the size of detected faces, for example, provides a direct link to the concept of shot framing.[8] The blocking of a scene can similarly be deduced by knowing the relative positions of identified characters within a specific cut. Once produced on a large scale, these extracted formal elements can be aggregated to explore visual style across a collection of materials. It is then possible to understand how visual style is used within the internal construction of narrative and as a way to engage broadly with external cultural forces. The method is an example of an approach to large scale image analysis that Arnold and Tilton have termed *distant viewing*.[9]

Distant viewing is a methodological and theoretical framework for studying large image collections. Because of the way that images make meaning and how computers process these forms, Arnold and Tilton argue that we must teach the computer how to "view," which requires constructing a representation of elements contained within the visual material, often in the form of a metadata schema and algorithm. Using computational methods, the methods call for the automatic extraction of semantic elements such as facial recognition and shot breaks, which are important to this study of visual style.

The following analysis is split into two sections. In the first section, we focus

[5] Manuel Burghardt, Michael Kao, and Christian Wolff, "Beyond shot lengths using language data and color information as additional parameters for quantitative movie analysis," 2016.

[6] Kevin L Ferguson, "Digital Surrealism: Visualizing Walt Disney Animation Studios," *DH and Media Studies Crossovers* 11 (1 2017).

[7] Joel Burges, Nora Dimmock, and Joshua Romphf, "Collective Reading: Shot Analysis and Data Visualization in the Digital Humanities," *DH and Media Studies Crossovers* 3 (3 2016).

[8] Common framing types include 'wide-shots' and 'close ups'.

[9] Taylor Arnold and Lauren Tilton, "Distant viewing: analyzing large visual corpora," *Digital Scholarship in the Humanities*, (2019). For more, also visit the Distant Viewing Lab at the University of Richmond at distantviewing.org.

on the general methodology for applying computer vision techniques to capture visual elements within moving images. Understanding the context of the algorithms is important because most of our work would have been impossible only a few years ago. In order to establish that the results are valid, predictive metrics from our own hand-coded datasets are presented. While much of the computer science literature focuses on only a single focused task, our methodology establishes how elements extracted from existing tools can be combined to study a corpus of moving images as a whole object.

The second section applies our technique for extracting formal elements to a corpus of moving images containing every broadcast episode of two Network Era sitcoms: *Bewitched* (1964-1972) and *I Dream of Jeannie* (1965-1970). This corpus includes 393 episodes and over 150 hours of material. We present techniques for aggregating and visualizing the computed time-based metadata. We then show how these analyses establish the way that visual style constructs character centrality and the formation of narrative structure. Specifically, we illustrate how formal elements serve to differentiate the role of the female characters within each series. Across every metric analyzed, Samantha is distinctively positioned as the leading character on *Bewitched*. Jeannie, however, is consistently shown to be visually and narratively dominated by Tony. The differing nature of Samantha and Jennie within their respective series challenge existing feminist and queer readings that often equate the two sit-coms with one another.[10] Our analysis provides a path for understanding suburban feminism in 1960s America through two differing perspectives.

Method

Computer vision

Computer vision research has evolved considerably over the past decade. Advances have come in part by utilizing custom hardware in the form of graphical processing units (GPUs) and the use of convolutional neural networks. While full replication of the entire human visual system remains an ambitious open task, image processing algorithms can now perform as well as or better than manual expert annotations on narrowly focused tasks. A recent medical study, for example,

[10] Morris Meyer, "I Dream of Jeannie: Transsexual Striptease as Scientific Display", *TDR* 25, no. 1 (1991). Karen M. Stoddard, "Bewitched And Bewildered," *Journal of Popular Film and Television* 8, no. 4 (1981).

produced an algorithm for classifying skin lesions as malignant or benign that was shown to be at least as good as predictions of 21 board-certified dermatologists.[11] Similar results include the accurate prediction of diabetic retinopathy,[12] lymph node metastases in breast cancer patients,[13] and pneumonia diagnoses from chest x-rays.[14]

Much of the work in computer vision has been driven by concrete industry applications. Examples include facial recognition algorithms for security applications and object localization methods for use in self-driving cars. Guided by these applications, academic research groups have built open source software libraries for efficiently applying deep learning models to image corpora.[15] Within these frameworks, computer vision researchers have also published models that can be applied off-the-shelf to common image processing tasks.[16] These public models are frequently chained together, adapted to specific needs through the process of transfer learning, and packaged as user-facing applications. Consumers constantly interact with such computer vision algorithms; facial recognition is central to both the iPhone FaceID feature (unlocking a mobile device) and the internal Facebook algorithms for the semi-automatic tagging of friends in uploaded photo albums.

Despite recent advances and available software, there remain substantial barriers in the application of state of the art computer vision methods to the analysis of cultural materials. There are limited end-to-end tools that take images as inputs and directly extract useful semantic information as outputs, and a significant amount of code is required to glue together all of the constituent elements. Analyzing moving images demands additional work; most computer vision software frameworks and models can be applied only to still images. Code is needed to extract frames from digitized video files, apply models to each frame, and sensibly aggregate the output.[17] Additionally, computer vision algorithms are typically

[11] Andre Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* 542, no. 7639 (2017): 115.

[12] Varun Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Journal of the American Medical Association* 316, no. 22 (2016): 2402-2410.

[13] Babak Ehteshami Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Journal of the American Medical Association* 318, no. 22 (2017): 2199-2210.

[14] Pranav Rajpurkar et al., "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv*:1711.05225, (2017).

[15] Such as the popular *TensorFlow* library: Martin Abadi et al., "TensorFlow: a system for large-scale machine learning.," in *OSDI*, vol. 16 (2016), 265-283.

[16] The Caffe "model zoo" hosted on GitHub is one example of a source for many of the most commonly used pre-train models.

[17] Extracting frames from video files is relatively easy with existing libraries but figuring out how

trained on modern, color, high-resolution datasets. There is no guarantee that they will perform well on historic materials (black and white images; grainy film stock; material digitized at a low resolution).[18] They must also be interrogated for the social and cultural assumptions that are built in with special attention to the problematic assumptions that can be built into data and algorithms.[19] Extensive data creation as well as testing and tuning of computer vision algorithms is often required to achieve acceptable results.[20]

The method presented in this article extracts structured information from moving images in order to analyze production and editing style and is a type of distant viewing. Our focus is on the application of facial recognition algorithms to locate and identify characters with a shot. Details regarding framing, shot blocking, and narrative structure can be inferred directly from information about the characters. We refer to the description of who and where characters are in a shot as *shot semantics*. Our goal is to build algorithms that are able to extract these semantics automatically from a corpus. Shot semantics provide features that are of a higher complexity relative to shot breaks and color analysis. By incorporating a more nuanced view of the moving images, our analysis not only provides a deeper understanding of the formal decisions made by actors, writers, camera operators, directors, and editors but lays bare those meanings that are not necessarily intended but are still articulated through form and style.

Establishing a semantics for shots within a television episode involves several steps. We first split an input video file into individual shots, then locate faces within every frame, identify the faces in a frame, and finally put all of the information together to cluster shots based on camera angles and distances. Each of these steps have their own challenges; we address these in the following subsections.

to sensibly combine frame-level semantic information across shots and scenes is a difficult question that has received relatively little attention within the computer vision community.

[18]

[19] Examples of such work includes Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the* World (The MIT Press, 2019); dana boyd and M.C. Elish, "Situating Methods in the Magic of Big Data and Artificial Intelligence", Communication Monographs, (2017); Sofia Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism,* (New York University Press, 2018); Jessica Marie Johnson, "Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads", *Social Texts* 137 (2018): 57-79.

[20] As anecdotal evidence of this, we experienced trouble applying what were at the time (2014) state of the art face detection algorithms to photographs from the 1930's. The algorithms were not able to detect the majority of people wearing hats because hats were far less common in the modern training data used to build the model.

Shot boundary detection

One of the first tasks in understanding a video file is to break up the sequence of frames into distinct units known as *shots*—or uninterrupted footage between cuts—in a process known as shot boundary detection.[21] Because the problem of computationally detecting shot boundaries is challenging, there are several competing algorithms and software for detecting shot breaks for varying applications and available computing power.

In general, the problem of computationally detecting shot boundaries is quite challenging. Fortunately, the vast majority of cuts in most film and television episodes consist of abrupt transitions between alternative camera shots. These are both easy to define and relatively easy to detect. The algorithm we use in our analysis uses a combination of histogram differences and down sampled pixel intensities. Our method functions by measuring the extent to which two successive shots are different from one another in terms of the general color palette and the distribution of brightness over the frame. If these differences fall above a specified threshold, we indicate the presence of a cut.[22] We also added special logic to avoid erroneous cuts during fade-in and fade-out events. Performance is very good on our two series, with a precision of 98.7% and recall of 97.2% (F1 score: 0.979), when tested on a hand-coded set of 10 episodes from each series. Most of the detected errors were the result of difficult crossfades (false negatives) or bright strobe lights used in special effects (false positives).

Face and people detection

A substantial understanding of the visual style implicit in the production, directing, and editing of television can be identified by knowing where specific characters are in each frame, as the sitcom narrative is one that consistently advances through character dialogue and action. The task of identifying the location of characters in an image can be split into two subsequent tasks. The first, face detection, attempts to find the location of all faces present in a given image. Once faces are located, the process of face recognition predicts which person is associated with a given face.

[21] Owing to the centrality of this task in video processing, there are a number of competing terms used in the literature, including "shot transition detection," "shot detection," and "cut detection".

[22] More specifically, we down sampled the image to a 32-by-32 grid, converted to hue, saturation, and value space, and computed the 40th percentile of the absolute deviation between successive frames. We also computed 32-bins per channel histogram counts on the original size image in RGB-space. A cut was coded whenever both of these values passed beyond a manually tuned hard threshold. Detected cuts within 12 frames of one another were merged together.

Relatively fast and accurate face detection algorithms have existed to detect well-framed faces for over two decades. These include the original Haar-wavelet method for general purpose object detection,[23] the Viola-Jones object detection framework,[24] and the histogram of oriented gradients (HOG) detector.[25] All of these methods identify faces by, approximately, finding portions of an image that are shaped like a face. This makes the algorithms robust to lighting, skin tone, and adaptable to multiple shot widths. These methods, particular the HOG detector, continue to be popular and are often provided in modern image processing libraries.[26] The downside of shape-based estimators is that they are not able to easily extend to faces in profile. On television shows, wide shots frequently display multiple characters turned inwards as they engage in conversation. This is particularly true of sitcoms, which are primarily driven by dialogue, as opposed to action sequences or special effects.

As with many other areas of computer vision, the use of neural networks has enabled a great improvement in the performance of face detection models. Two prominent examples are Faster R-CNN and FAREC-CNN.[27] Figure 1 shows an example of the faces detected by a neural network model compared to the shape based models. Notice that all of the faces in profile that are missed by the HOG detector are found with the neural network model.

The neural network-based face detection algorithms performed very well on our television corpus. We applied the FAREC-CNN to every ten frames in six episodes of each of our two series. We then labelled where the algorithm had mistakenly identified objects as positive faces or had failed to detect faces when present. For the latter, we only considered a face as being "missed" if at least one eye of a face was present in the frame and excluded any "extras" who were in the far background and removed from the main action.[28] Overall, we ended with

[23] Constantine P Papageorgiou, Michael Oren, and Tomaso Poggio, "A general framework for object detection," in: *Sixth international conference on Computer Vision* (IEEE, 1998), 555-562.

[24] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1* (IEEE, 2001).

[25] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005.* (IEEE, 2005), 886-893.

[26] For example, a trainable HOG detector algorithm is provided by the popular OpenCV library, dlib library, and the scikit-image Python package.

[27] Xudong Sun, Pengcheng Wu, and Steven C.H. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomputing* 299 (2018): 42-50; S Sharma, Karthikeyan Shanmugasundaram, and Sathees Kumar Ramasamy, "FAREC: CNN based efficient face recognition technique using Dlib," in *Advanced Communication Control and Computing Technologies* (ICACCCT), 2016,, 192-195.

[28] The latter condition was rarely used in our labelling; these two sitcoms are set primarily in a personal home or office with a minimum of scenes containing characters in the background.

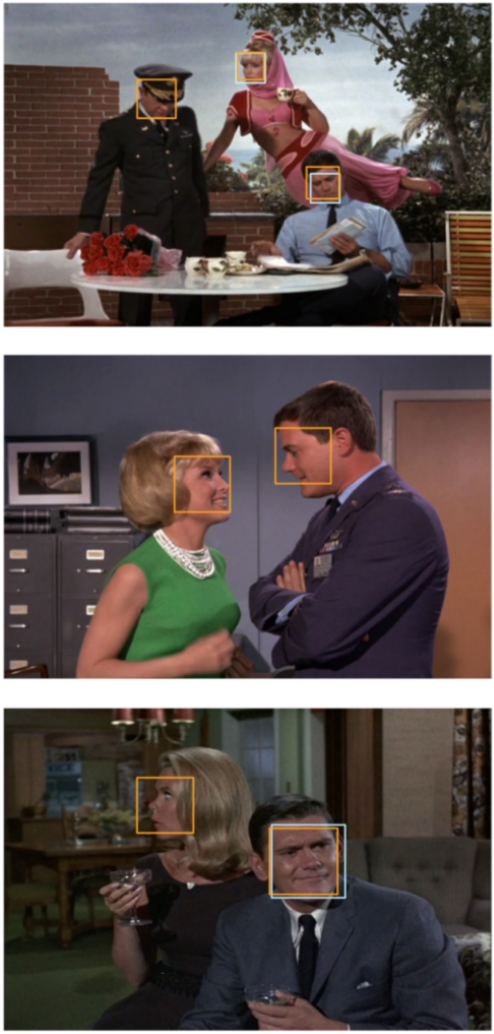

Figure 1: Faces detected using a HOG detector (blue) and a neural network (orange) from screen shots of I Dream of Jeannie and Bewitched.

a precision of over 98.3% and a recall of 95.1% (F1 Score: 0.967). Almost all of the false negatives came from wide shots where a character's face was partially obscured. The majority of false positives were edge-cases where the algorithm detected the face in a painting or statue. For comparison, the HOG detector had an overall recall of only 55.2%. Given the centrality of the face detection results to our analysis, this is strong evidence of the importance that recent advances in computer vision have had to our work.

Face recognition

After detecting where faces exist in individual frames, the next step is to identify which characters are associated with each face. The widespread use of neural network in face recognition preceded its applications in detection by several years, largely due to the presence of large training datasets and the relatively small size of the images involved. Open source libraries with neural network based recognition algorithms include OpenFace and OpenBR.[29] Both libraries are reasonably popular and their respective methods are widely cited. These algorithms, as well as most other face recognition methods, require a preprocessing step whereby a detected face is *aligned* such that key reference points (such as both eyes, the nose, and mouth) and standardized. This approach works well for the high-quality faces detected by algorithms optimized for finding front-oriented faces. Trying to align the faces in profile, such as the middle panel in Figure 1, is impossible given that only part of the face is visible.

In our analysis, we have made use of the VGGFace2 face recognition model.[30] Unlike other recognition algorithms, this approach does not require the input images to be aligned and was particularly built to handle faces that can only be seen in a low resolution or in profile. We hand labelled faces from the top four characters in five episodes of each of our two series and compared these results to the VGGFace2 predictions.[31] Defining precision as the proportion of assigned faces that were correctly identified and recall as the proportion of main characters correctly identified, Figure 2 gives a precision and recall curve for varying cut-off

[29] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan, OpenFace: A general-purpose face recognition library with mobile applications, technical report (CMU-CS-16-118, CMU School of Computer Science, 2016); J. Klontz et al., "Open Source Biometric Recognition," *Biometrics: Theory, Applications and Systems*, 2013, 42- 50.

[30] "VGGFace2: A dataset for recognising faces across pose and age," in *IEEE Conference on Automatic Face and Gesture Recognition* (IEEE, 2018).

[31] The algorithm can be applied to faces that are not in the training set by supplying a single *reference* image for every character of interest. The algorithm returns which faces in the dataset appear to be the same as one of the reference images.

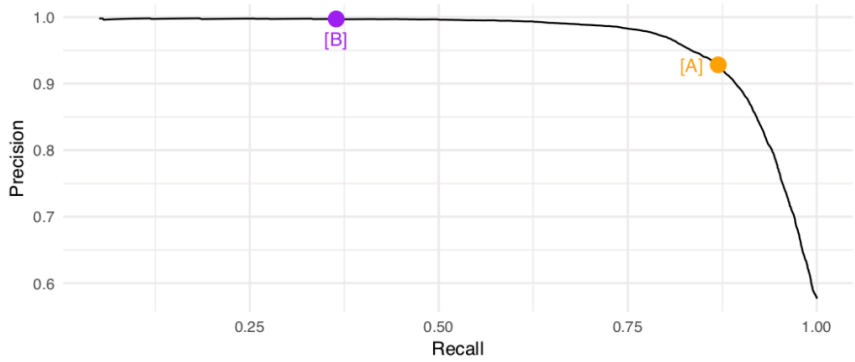

Figure 2: Face recognition testing results for accurate detection of 4 primary characters in *Bewitched* using varying cut-off scores. Precision is measured as proportion of correctly labelled faces as a ratio of all faces determined to be one of the main characters. Recall is the proportion of main characters that are correctly identified. The orange dot [A] shows the optimal F1 score; the purple dot [B] shows a model with 99% precision.

scores in the algorithm. Using the cut-off value with the maximal F1 Score (0.898) yields an overall precision of 92.8% and recall of 87.0%.[32] Similarly, a precision of 99.0% can be achieved while maintaining a recall of only 67%.

Our current use of the VGGFace2 model for face detection works very well for the subsequent analysis of *Bewitched* and *I Dream of Jeannie*. Importantly, the key results are unchanged regardless of whether we use the high-precision model cut-off value or attempt to achieve the optimal F1 Score. One important hole in the current application is the difficulty of VGGFace2 to successfully identify children. Our initial intention was to include the character Tabitha from *Bewitched*; however, the precision of this task across all available face recognition algorithms made this unworkable.[33]

[32] For comparison, our original attempt using the ResNet-32 model achieved on F1 Score of only 0.516. See the following for the basis of this original model: Kaiming He et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770-778.

[33] When applying this algorithm to other corpora it would be prudent to conduct some manual testing in order to find other potential holes in the dataset. For example, other face recognition software has frequently failed when applied to datasets with non-white actors. See this article for a synopsis of racial biases in face recognition algorithms: Joy Buolamwini and Gebru, Timnit, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proceedings of Machine Learning Research: Conference on Fairness, Accountability, and Transparency* 81 (2018): 1-15.

Shot classification

By using the placement of faces in a shot, we have hypothesized that we will be able to offer a classification of shot semantics. That is, we want to use the extracted data about faces and automatically determine features of a specific shot within a video file. In this final step we no longer work directly with the visual materials and build a meta-algorithm that takes detected shot breaks and faces as input and outputs categorical predictions for each shot.[34]

There are several specific types of shots commonly described within film and media studies. Typically, these depend on the relative size of characters within a frame, the number of characters, the orientation of the characters, and the angle of the camera. Many of these specify something about a particular frame, however, they do not describe in aggregate multiple elements at the shot level. For example, how should we classify a single shot that pans from a close-up of one character to an establishing shot of a building? If one character is standing over another character at a table, is this a medium long shot (most but not all of one character is in the frame) or a medium close up (as only the seated character's upper half is visible)?

In order to algorithmically classify shot types, we need precise definitions with which to build a testing dataset. After carefully parsing through many extracted scenes, we settled on the following:

- **close-up**: Only one character is present in the foreground of the entire shot and the waist of this character is never shown. Further, if the character is sitting, we can infer that the shot is framed close enough that they would leave the shot if they were to stand up.
- **close two-shot**: Exactly the same as a close-up, but there are two characters present in the shot and at some point we see each character's face.
- **group shot**: Any shot that contains at least three characters.
- **over-the-shoulder**: Two characters are shown in the shot but one of the two characters (the one closest to the camera) has their back to the camera and no face showing. This type of shot is likely to arise in the context of a shot-reverse shot editing sequence in which the camera follows a conversation between two or more characters.

[34] We are aware of only one prior attempt to offer an algorithmic taxonomy of film shots. The presented taxonomy mostly relates to camera movement in action films, and was not very relevant to the features we are interested in classifying. Hee Lin Wang and Loong-Fah Cheong, "Taxonomy of directing semantics for film shot classification," *IEEE Transactions on Circuits and Systems for Video Technology* 19, no. 10 (2009): 1529-1542.

(a) Group shot (b) Two shot

(c) Close shot (d) Over-the-shoulder shot

Figure 3: Shot type examples from the *Bewitched* episode "The Short, Happy Circuit of Aunt Clara" (Season 3, Episode 9).

Figure 3 shows examples of these four shot types from an episode of *Bewitched*. We refer to any shot that is neither a close-up nor a close two-shot as a **long shot**. This includes shots that show characters from the waist up—in other studies referred to as a "medium shot"—or the entire body.

To detect these shot types from the data, we built a hand-constructed algorithm based on the number of detected faces, the placement of the faces, and the size of the faces. Then, the shots in 8 episodes were hand labelled and compared with the results from the algorithm. Close shots were identified with a precision of 99.0% and recall of 93.5% (F1 Score: 0.962); group shots were classified with a precision of 98.25% and recall of 95.73% (F1 Score: 0.969); and over-the-shoulder shots had a precision of 88.24% and recall of 95.74% (F1 Score: 0.918). The majority of our analysis focuses on the timing and presence of close shots, for which the precision is particularly high as a result of conservatively chosen logic in our algorithm.

Analysis

Corpus

The Network Era of American television (1952-1985) was controlled by just three competing networks: ABC, CBS, and NBC. By the early 1960s, the majority of American households owned a television and had access to the affiliate station from all three major networks. With extensive market penetration and high user engagement, there was little doubt that millions of Americans watched television programming on any given evening. The goal for each of the established networks in the oligopoly was to the attract the most viewers to their programming over the competition.[35] These market forces caused networks to push out content with a wide appeal in order to attract the largest number of viewers, and in turn the largest percentage of the advertising revenue. The result was a relatively uniform stream of programming aimed primarily at white suburban middle class families.[36]

The situational comedy, or "sitcom," has been one of the dominant narrative forms of television shows from the early years of the Network Era. Sitcoms focus on a fixed set of characters, unlike sketch comedy and vaudeville, and usually

[35] Gary Edgerton, *The Columbia History of American Television* (Columbia University Press, 2007).
[36] Amanda Lotz, *The Television Will Be Revolutionized* (NYU Press, 2014).

consist of self-contained episodes. Within each episode, the cast of characters are presented with a new problem that is resolved by the episode's conclusion. While major plot events, such as extended romances, may be referenced across episodes, the structure of the sitcom makes it easy for a viewer to follow the series even if they miss a particular week's show.[37]

These features made sitcoms an attractive programming choice during the Network Era. Prior to the advent of consumer VHS recording technologies in the 1980s, Americans had to watch their favorite programs when they aired or risk missing them entirely. There was also no way to fast-forward through commercial breaks, as has become possible with VCRs, DVRs, and streaming services. These viewing patterns dictated the three-act structure and self-contained narratives that characterize network era programming.[38] The familiar recurring characters served as a draw for viewers; the fixed cast and sets kept production costs low; and, the formulaic narrative assured that consumers would not lose track of the show if they missed an episode.

We now turn to two popular American sitcoms from the 1960's. In line with the genre and time period, both feature a small, all white cast living in generic suburban neighborhoods. Each week's plot follows a formulaic model: a problem arises that threatens the status quo at home or in the workplace, the first several attempts to address it fail in some comic fashion, until ultimately a happy resolution presents itself at the episode's conclusion. Yet, we will show that the narrative and editing of these series is far from formless. There are distinct visual elements that serves to challenge and reinforce the narrative structures present in each show.

The sitcom *Bewitched* premiered on ABC in the Fall of 1964 and ran for a total of eight seasons. Its main premise focuses on the marriage of the ordinary Darrin Stevens to the supernatural witch Samantha. Samantha does her best to live a "normal" American life but difficulties from her magical world crop up to make this process difficult. The main cast is completed by Samantha's meddling mother Endora, Darrin's boss Larry Tate, and (from Season 3 onward) the Stevens' magically gifted daughter Tabitha. *Bewitched* proved remarkably popular. It was the second-highest rated show across all three networks in its first season and enjoyed wide syndication from 1972 to the early 2000s.[39]

In order to understand the typical structure of an episode of *Bewitched*, it is instructive to walk through the plot of a specific example. We can then see how this

[37] Umberto Eco concisely describes this phenomenon as "a plot which does not 'consume' itself." Umberto Eco, "The Myth of Superman," *Diacritics* 2, no. 1 (1972): 17.

[38] See Amanda Lotz, *The Television Has Been Revolutionized.*

[39] Walter Metz, *Bewitched* (Wayne State University Press, 2007), 14-16.

Figure 4: Example of the detected characters and narrative breaks from one episode of *Bewitched*.

plot is reflected through the data computationally extracted from its visual contents. Figure 4 shows all of the detected character faces in the episode "Business, Italian Style," the seventh episode in the show's fourth season, which first aired on September 21, 1967. In the opening scene Darrin and Larry meet with the assistant of a potential new client, an Italian businessman trying to expand into the U.S. market. Through a comic misunderstanding, Larry asserts that Darrin will be able to speak to the assistant's boss in his primary language. Following the credits, Darrin returns home and describes the situation to his wife Samantha, who then confides in her mother Endora. Endora casts a spell that causes Darrin to *only* speak in Italian, which the Stevens' notice the next morning over breakfast around the midway point of the episode. After the commercial break, Larry shows up at Darrin's home to find that he can no longer communicate in English. After several mishaps, Endora reverses the spell, the business is saved, and all is well with the main characters. The final scene ends with a dinner party that brings together Darrin, Samantha, Larry, and the Italian businessmen, reinforcing the successful conclusion of the episode.

In response to the success of *Bewitched*, in the Fall of 1965 NBC debuted *I Dream of Jeannie*.[40] The show features a 2,000-year-old female genie, referred to as "Jeannie," discovered by the astronaut Tony Nelson, whom she refers to as her "master." The main cast also includes Tony's friend and fellow astronaut Roger Healey

[40] Metz, *Bewitched*, 17.

and the NASA psychologist Dr. Alfred Bellows. A recurring plot device involves Tony and Roger attempting to fulfil their job responsibilities, while hiding from and dealing with Jeannie's magical antics. In later seasons, Jeannie and Tony become romantically involved and later marry. *I Dream of Jeannie* did not enjoy the same level success as *Bewitched*, running only five seasons and failing to win any major awards.

Scholarship on *Bewitched* and *I Dream of Jeannie* has in large part focused on the depiction of two prominent female characters during the early days of second-wave feminism. Often the context of these two shows are combined together, typically in a negative light, such as Stoddard's analysis on the role of magic and feminism:

> Both Samantha and Jeannie are unusual women with extraordinary powers, powers which they promise to curtail at the insistence of the men they love. Samantha's husband, a mere mortal, expects her to perform tasks in a human, rather than supernatural, manner… Similarly, Jeannie the genie is "found" by a mortal, who becomes her "master." With crossed arms and the blink of her eyes, Jeannie has the power to control everything around her. At the insistence of her master, however, she busies herself keeping his house orderly and catering to his wishes in as human a manner as possible.[41]

Others offer different readings of the show's gender and sexual politics. Lynn Spiegel noted that in *Bewitched* "the woman's alien powers serve to invert the gender relations of suburban domesticity, and with this, the consumer lifestyles that characterize the suburbs are also parodied."[42] While the male characters are attempting to subvert the powers of the female characters, at the very core of each show's premise is that women's magic or power cannot be contained and periodically "escapes" domestic or marital constraints. *Bewitched* can also be read as a queer metaphor, in which "mortal" norms takes the place of heteronormative thought.[43]

These two shows form an excellent set for our study of form in television, as Network Era sitcoms have small casts and a small repertoire of sets, making both relatively easy to analyze computationally. The prominence of both shows and their repeated presence in television studies research speaks to the cultural importance

[41] Karen M. Stoddard, "Bewitched and Bewildered," *Journal of Popular Film and Television* 8, no. 4 (1981): 50.

[42] Lynn Spigel. *Welcome to the Dreamhouse: Popular media and Postwar Suburbs* (Duke University Press, 2001), 132.

[43] Patricia Fairfield-Artman, Rodney E Lippard, and Adrienne Sansom, "The 1960s Sitcom Revisited: A Queer Read," *Taboo: The Journal of Culture & Education* 9, no. 2 (2005): 27.

of both shows and allows our work to engage with broader conversations about the medium and genre. Even better, the scholarly disagreements over the relationship between the shows and their feminist (or not) representation of women makes our analysis all the more revealing in broader conversations about television, genre, and gender.

Dominant character

An open question regarding these two series is to understand which characters are considered the "main" characters of the show. Elizabeth Montgomery (Samantha) and Barbara Eden (Jeannie) are certainly the most well-known actors in our corpus and feature prominently in the shows' marketing materials, but it is not immediately clear to what extent this visibility and name recognition translates into centrality in the narratives of their respective series. For example, a large proportion of plot lines revolve around the careers of their romantic counterparts. To what extent are these less-recognized male characters the true central points of the show? The series' titles even hint that this may be the case by establishing agency with the ordinary, career-oriented male characters: Darrin is "bewitched" by Samantha, and Tony is the one "dreaming"of Jeannie. How does visual style reflect or potentially undermine narrative centrality across these series?

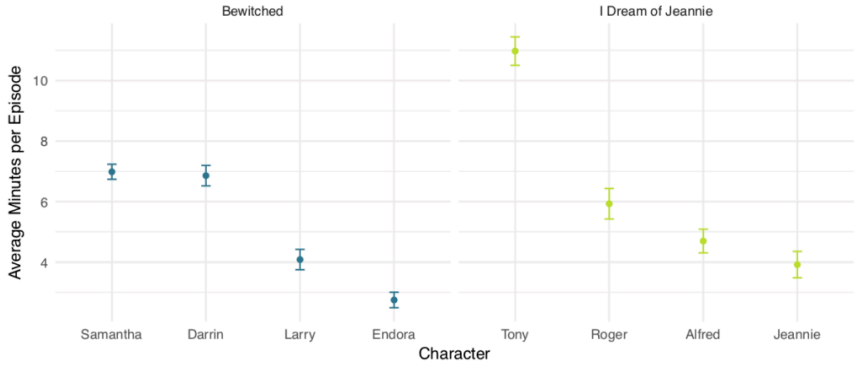

Figure 5: Average minutes per episode for which a character is visible. Error bars with 95% confidence intervals for the mean of each group.

One straightforward way to assess to what extent visual style correlates to character centrality is to quantify the frequency that each actor appears on screen,

a type of shot semantics. For every shot in our corpus, we calculated every detected character face that appeared at least once. These numbers were aggregated together with shot lengths to compute the average minutes per episode that each character was visually present.[44] Figure 5 shows the computed averages along with confidence intervals. *Bewitched*'s Samantha and Darrin share the lead with an average of 7 minutes of screen time per episode, suggesting not only that their marriage lies at the heart of the show but that theirs is an equitable and egalitarian coupling. At just under 11 minutes per episode, Tony is by far the most visually present character across both shows. On *I Dream of Jeannie*, Jeannie, seen for less than an average of 4 minutes per episode, is the least visually displayed major character. This inequality is a point in favor of *Jeannie*'s retrograde feminism: while she may act out and rebel in these episodes, the form of the show often transforms or screens her (and her controversially costume that was seen by some audiences as too sexy for TV) from the viewer. By contrast, Tony's coworker—Roger (6 minutes) and Alfred (4.5 minutes)—are seen for longer average durations than Jeannie.

These results highlight a significant difference between the two shows and raise a host of related questions. The character relationships on *Bewitched* are more balanced, indicating the show's twin focus on work and home. Larry is Darrin's boss but Endora is Samantha's mother. That said, throughout the series, Samantha is absent from the workplace and Darrin is often absent from the magical and family plot elements. Tony's persistent visual presence in *I Dream of Jeannie* underscores how all of the lead characters are connected to him, not to one another: Roger is Tony's best friend, Alfred Tony's boss and psychiatrist, and Jeannie his genie. As a result, the majority of the action, both within the work and domestic spheres, usually involves Tony in a substantial way.

Understanding when within an episode each character is most likely present indicates how the narrative structure reinforces character centrality. Each episode can be broken into four parts. The first part, a "cold open," occurs from the start of the episode until the title sequence and lasts approximately 2-3 minutes. The main actions that build and resolve the plot occur in two longer blocks of time, interrupted by a mid-episode commercial, lasting around 10 minutes each. We refer to these as Act 1 and Act 2. Following the final commercial break, a final short resolution scene reaffirms the successfully resolved plot and subsequent return to normalcy.[45] We show the proportion of episodes for which each char-

[44] The character of Darrin on *Bewitched* was played by two actors: Dick York (seasons 1-5) and Dick Sargent (season 6-8). We created separate face detection algorithms for both actors. However, in this and all other results we have combined them to their common character.

[45] This four-part structure roughly follows the chart provided by Jeremy G Butler, *Television: Visual Storytelling and Screen Culture* (Routledge, 2018); and very accurately matches the structure of

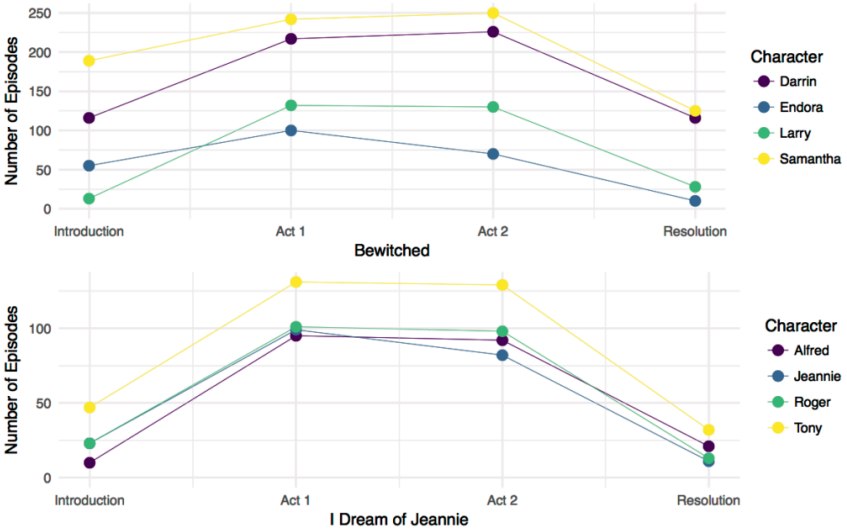

Figure 6: Number of episodes for which a character was seen.

acter was detected in each narrative part in Figure 6. By narrative part, Tony's presence dominates the other *I Dream of Jeannie* characters. Most notably, he is almost always present during Act 2. Samantha is also shown to be present in more acts that Darrin, particularly in the cold open. Though their overall screen time may be similar, we rarely go too long without seeing Samantha.[46]

One way to establish the importance of a particular character is to make a character the first person seen in the opening of an episode. From the opening shot, all of the subsequent actions and characters are linked to the perspective of this starting character. This effect is closely related to the cognitive bias of *anchoring* in which a set of potential options in a decision-making process are all judged in relationship to the first option presented.[47] The results in Figure 7 tabulate the number of times each character was the first detected face in an episode.[48] Tony

our two series. We were able to automatically extract the narrative parts through the chapter breaks encoding in our DVD materials.

[46] A common trope on the show involves fast cuts between Darrin and Samantha as they talk over the telephone between the office and home.

[47] Muzafer Sherif, Daniel Taub, and Carl I Hovland, "Assimilation and contrast effects of anchoring stimuli on judgments.," *Journal of Experimental Psychology* 55, no. 2 (1958): 150.

[48] Only 1.04% of the dataset started with a shot containing multiple faces; in these cases neither character was counted.

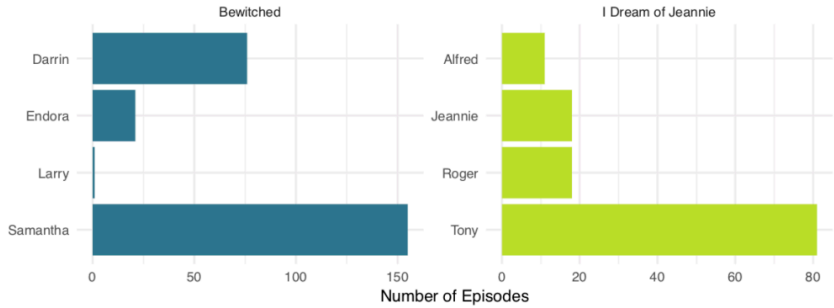

Figure 7: Number of episodes where each character is associated with the first face detected in an episode.

(81 episodes) and Samantha (153 episodes) are overwhelmingly the most likely to be first seen in an episode; Darrin is about half as likely to be featured first relative to Samantha. All of the remaining characters, including Jeannie, are featured at the start of less than 20 episodes. For *I Dream of Jeannie*, this further reinforces the centrality of Tony in the show. It also re-establishes the overall narrative: the show chronicles the life of the astronaut Tony Nelson, as his world becomes disrupted by the sudden and unwelcome intrusion of Jeannie. (This is echoed by the show's opening animated credits, in which Tony is introduced first, then Jeannie.) In *Bewitched*, while the show in many ways focuses on both characters, Samantha takes center stage more frequently than her mortal husband. To what extent that allies viewers with Samantha, inviting them to root for her subversion of her husband's will, is a question we can begin to address through the presence of close-ups.

Close-up shots are yet another visual tool for establishing characters' centrality and subjectivity. In the mid-1960's, television was distributed in standard definition, television sizes were substantially smaller, signals were sent through noisy analogue radio transmission, and many households watched television in black and white. Viewers could only get a good view of an actor only in tightly framed shots that reveal the character's sometimes private thoughts, feelings, and reactions. Figure 8 shows the average duration per episode that each character is shown in a close-up shot. Tony once again dominates the other *I Dream of Jeannie* characters, but overall, the longest duration (1.36 minutes) across both shows comes from Samantha. Along another dimension, Samantha dominates the visual space, Darrin playing a more prominent role compared to the other characters.

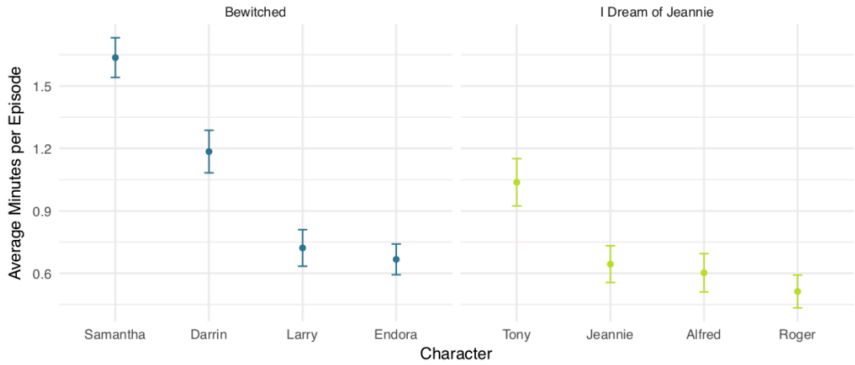

Figure 8: Average minutes per episode for which a character is visible in a close shot. Error bars with 95% confidence intervals for the mean of each group.

Taken together, these analyses of shot semantics provide insight into the prominence of the characters in our corpus. Tony appears as the center of the narrative structure, and this centrality is reinforced through the visual style— time on screen, close shots, and presence at the start of the episode—of the series. Surprisingly, the titular Jeannie is far from being the most central character, she is no more visually dominant than Roger or Alfred. On *Bewitched*, there are approximately the same number of scenes featuring either Samantha or Darrin. However, visual elements, such as screen time and the number of close-up shots, establish Samantha as the most important character. While the show focuses on the couple as a whole, Samantha is pictorially featured as the leading star character.[49]

Shot distribution and plot

The extracted locations of faces can be further used to examine how visual style communicates or complicates narrative. Tracking stylistic elements throughout each episode makes it is possible to investigate if and how producers and editors use visual elements to further the plot as it progresses from the cold open all the way through to the episode's denouement. These patterns, if present, provide further evidence of the visual complexity within the style of Network Era sitcoms.

[49] It is unclear from this analysis whether this is due to the magical abilities of Samantha Stevens or the star power of Elizabeth Montgomery. Likely both contribute in some way.

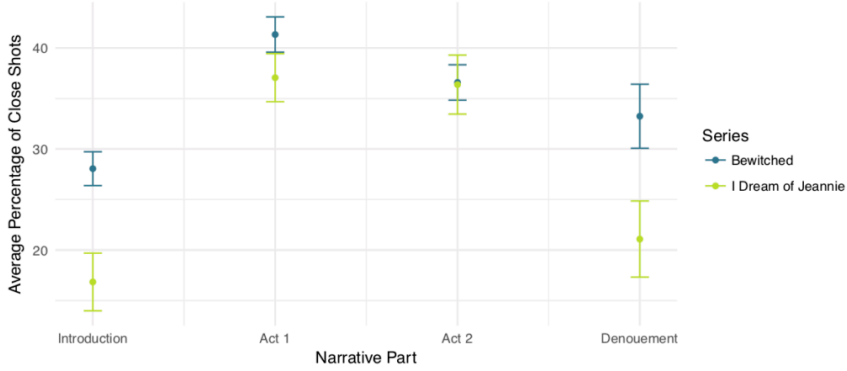

Figure 9: Percentage of all shots classified as "close" as a function of narrative act and series. Error bars with 95% confidence intervals for the mean of each group.

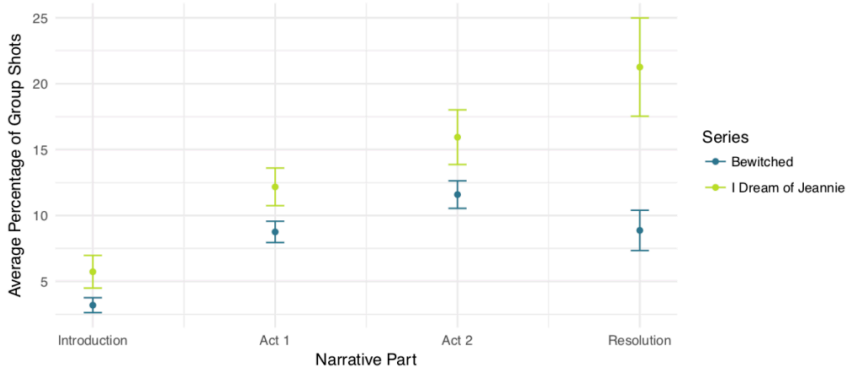

Figure 10: Percentage of all shots classified as group shots with three or more characters present, as a function of narrative act and series. Error bars with 95% confidence intervals for the mean of each group.

One way to explore shot semantics is by shot type. The distribution of shot types changes between narrative acts in both *Bewitched* and *I Dream of Jeannie*. These distributions also differ between the two series. Figure 9 gives the proportion of close shots for each narrative act and Figure 10 provides the proportion of group shots. The proportion of close shots is lowest in the Introduction and Denouement for both series, though the drop-off is noticeably larger for *I Dream of Jeannie*. This likely speaks to how these acts function as "establishing" sequences and book-ending the conflicts that characterize the drama of Acts 1 and 2. On *Bewitched*, there is also a significant drop in the proportion of close shots between Act 1 and Act 2; there is almost no difference between Act 1 and Act 2 in this metric on *I Dream of Jeannie*. One possibility for this is that Act 1 involves the women characters using their magic powers (as an instigating event) and Act 2 centers on the complications that arise from their meddling. Samantha wiggles her nose to do magic, typically revealed to the audience through a close-up on Montgomery's face; Jeannie's magic, triggered when she crosses her arms and nods her head, is better framed by a medium or long-shot, which is why we might not expect more close shots in the first act.

Across all narrative acts, the proportion of group shots on *I Dream of Jeannie* is higher compared to *Bewitched*. The rate of group shots increases for both shows from the Introduction to Act 1 and again from Act 1 to Act 2. Group shots are less common in the Denouement on *Bewitched*, but are at the highest observed level in *I Dream of Jeannie*. Overall, these results indicate that, as the plot builds, a larger proportion of wide shots are utilized. These wider shots help to convey the complexity of the conflict. For example, in the example "Business, Italian Style," wide shots are used throughout Act 2 to capture the interaction of Darrin, Larry, and their two business clients. In terms of the workplace scenes, wide shots provide the opportunity to show NASA as grand and impressive in comparison to Samantha's middle-class home or Darren's office in *Bewitched*. In particular, Act 2 in *Jeannie* has a significantly larger proportion of wide shots in which three or more characters present. In both shows, however, these shot changes serve to visually reflect the increasingly complicated relationships unfolding in the episode. The initial problematic in *Bewitched* may involve only two characters, such as Darrin and his boss Larry, but throughout the episode, additional characters are drawn in, creating additional strain and opportunity for comedy. These either extend the problem to other relationships, such as in the episode "Business, Italian Style," or signal failed initial attempts to resolve the plot. The increase in wide shots serves to visually represent and contain these increasingly involved plot lines.

There is a noticeable difference between the median shot lengths across each nar-

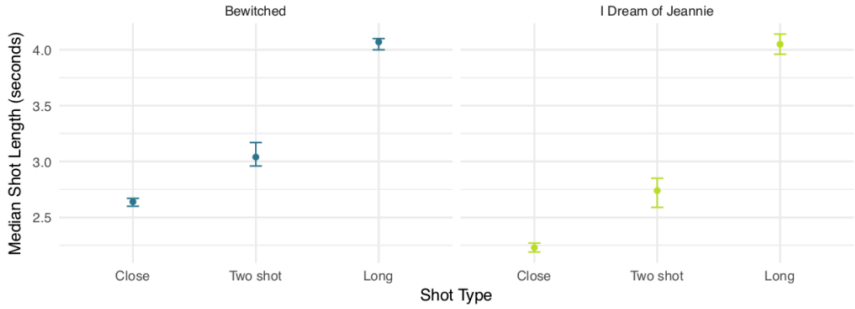

Figure 11: Median shot length for each series separated into close shots, two shots, and long shots. 95% confidence intervals for the median are given for each group.

rative act, but this relationship is deceiving; from our analysis, we can see that median shot length is closely related to shot type. There have been several studies that investigate median shot lengths in television series. It would not be surprising to find that shot lengths increase in Act 2 to capture an increasingly complex plot or, conversely, decrease in Act 2 to heighten the drama and speed of the storyline.[50] In Figure 11, the median shot lengths for each shot type are shown.[51] Shot lengths are very strongly tied to shot type in that even small differences in shot distribution can affect the median shot length. Regression analysis can be used to detect how strongly various factors influence median shot length. Predicting shot length as a function of series yielded an R2-value of only whereas a regression model using shot type provided an R2-value of .[52] The type of shot explains two orders of magnitude more variation in shot length than series. It seems, therefore, that shot lengths are not an ideal metric to use without first accounting for shot type.

The shot type chosen for the first and last shot of Act 1 and Act 2 reveal distinctive stylistic decisions. All four of these shots occur on the boundary of a commercial

[50] They actually decrease slightly, from a median of 3.3 seconds in Act 1 to a median of 3.17 in Act 2 when aggregated across both shows.

[51] Exact confidence intervals for the median are performed using a non-parametric sign test. Jean Dickinson Gibbons and Subhabrata Chakraborti, *Nonparametric Statistical Inference* (Chapman / Hall/CRC, 2010); Andri Signorell, *DescTools: Tools for Descriptive Statistics*, R package version 0.99.24 (2018).

[52] Following prior work on shot lengths, we assume that shot length follows a log-normal distribution and ran each regression of the logarithm of each shot length. Barry Salt, *The Metrics in Cinemetrics*, 2011.

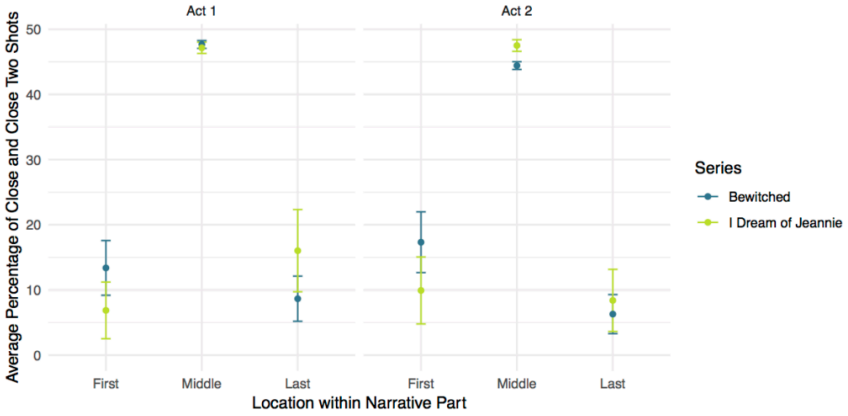

Figure 12: Percentage of shots that were classified as close or close two from both Act 1 and Act 2 as function of whether the shot was the first in the act, last in the act, or somewhere in the middle. 95% confidence intervals for the proportion are given.

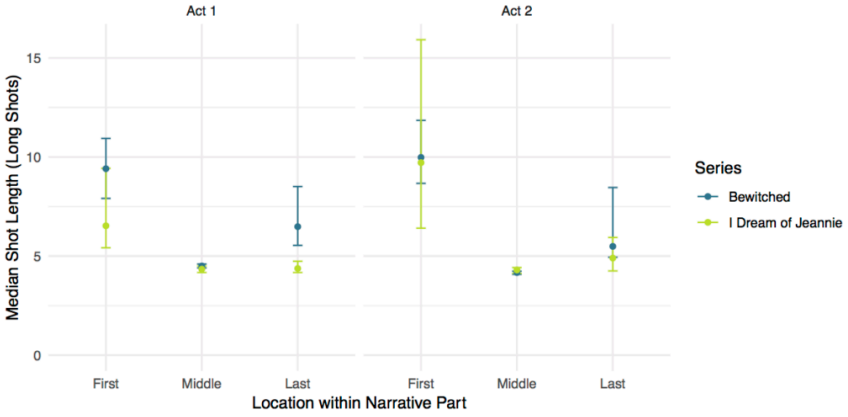

Figure 13: Median shot length of long shots for Act 1 and Act 2 as function of whether the shot was the first in the act, last in the act, or somewhere in the middle. 95% confidence intervals for the median are given.

break; the opening shots attempt to draw the viewer in and the closing shots try to keep viewers sticking around until the next act. Figure 12 gives the proportion of close and close two shots used in these opening and closing shots, as well as the proportion of other Act 1 and Act 2 cuts for comparison. While overall, across both series, over 40% of shots are framed closely, less than 20% of the first and last shots are closely framed. Figure 13 shows how these first shots also tend to have significantly longer durations. Overall, shot type and duration establish and continue the narrative by expanding the gaze of the viewer: at the start of each act, this widening establishes the mise-en-scène, in particular situating the characters in the spaces of work and home, and at the ends of each act, the wider view increases tension by gesturing to all of the characters and scenarios that remain to play out.

Magical and feminine gaze

As mentioned, the Samantha and Jeannie characters are similar or parallel characters, both magically gifted women trying to fit into stereotypical white suburban neighborhoods in the 1960's. As we have shown, a closer analysis reveals that Samantha and Jeannie actually function in dramatically different narrative roles within their respective shows. These differences, in turn, affect the way that each is visually portrayed. Despite these divergent roles, are there stylistic similarities between these two characters? To answer this question, we investigated those metrics that look at relative relationships between visual style tropes, rather than overall tabulations, as with the character centrality analysis.

The proportions of time each main character spent in close shots relative to the time seen in all shots are shown in Figure 14. By taking the relative proportion of close shots, a new pattern emerges. The three magical female characters all have a significantly larger proportion of their total screen time in close shots. The overall ratios for *Bewitched* characters are higher than those in *I Dream of Jeannie*, and Alfred is slightly more likely to be a close shot than both Tony and Roger. Given that the pattern here diverges from those seen in the character dominance metrics, it seems likely that the close shot ratios are enforcing a different visual characteristic and achieving a different narrative objective.

While there is a clear signal in the data identifying a visual aesthetic uniting Samantha, Endora, and Jeannie through the high proportion of close shots, it is not possible within this corpus to offer a confident hypothesis regarding the specific meaning of this effect. The high representation of close shots might be a common trope for the representation of feminine characters, a method for si-

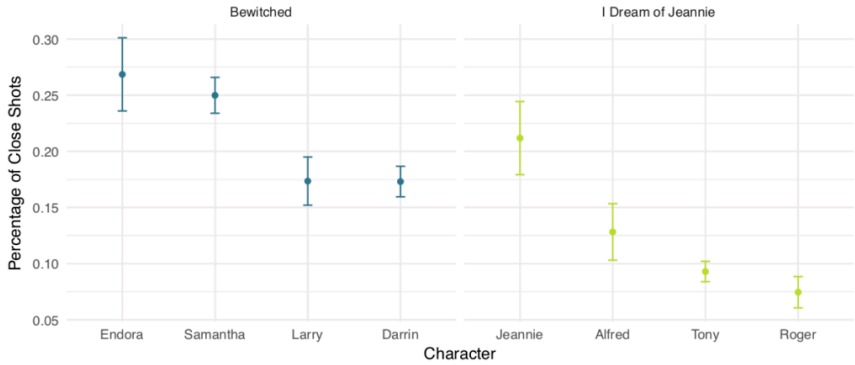

Figure 14: Proportion of time for which each character is shown in a close up shot as a ratio of the total time they are present in the show. Error bars with 95% confidence intervals for the mean of each group.

multaneously capturing feminine beauty and grace. As the three main magical characters, this style may instead be an attempt to represent their "otherness," the close shots giving the viewer a chance to gawk at the magical incantations and garish costumes while the narrative sets them apart as different, even isolated by their own powers. Another, perhaps more positively empowering reading would be that these magical women co-opt the screen, "stealing" the show from their male counterparts. Then again, the large proportion of close shots could simply be a direct function of the popularity of each actor, Montgomery and Eden being the leading ladies and Moorehead being a star in her own right. Indeed, all of these possibilities are plausible. To distinguish between them would require taking into account a larger corpus of material to isolate each hypothesis. We include this analysis because while distant viewing offers new knowledge, the insights in this section are shaped by and in conversation with the range of methodological approaches from film and media studies. Distant viewing may not always answer questions, but it can direct avenues of inquiry.

Conclusion

In many ways, the study of American television is indebted to David Bordwell, Janet Staiger, and Kristin Thompson's formulation of "classical Hollywood cinema," in which all aspects of film form are in service of conveying narrative and

"character-centered… causality is the armature of the story."[53] Put another way, all elements of film form—from editing to sound—follow character action, which in turn motivates story. The hidden complexities of Hollywood classicism have been examined in numerous cinema studies accounts: to what extent does the "invisible" style of Hollywood film allow for ideological inconsistencies, generating tensions between benign storylines and quietly subversive forms?

Through the use of facial recognition software, we bring these questions to the study of the American television sitcom, itself a character-driven dramatic form and one typically dismissed as middle-brow commercialism, lacking in formal or stylistic rigor. As Jeremy Butler writes, "all television programs employ conventions of the medium,"[54] and we ask what tensions, contradictions, and paradoxes are buried in these seemingly facile conventions. The "schema" of the situation comedy—the "bare-bones, routinized devices that solve perennial problems," to borrow Bordwell's language—can be the vehicle for messages that belie the content of the story, speaking to the fissures and frictions that characterize *Bewitched* and *Jeannie*'s fraught second-wave feminisms.[55] Given that Network Era sitcoms are often dismissed as formless middle-brow fluff, it is striking to find that the visual elements show noticeable difference between these two series and serve to reinforce character relationships and gender politics. Shot semantics offer an exciting way to analyze visual elements.

There are several avenues for extending the analysis presented here to a wider set of scholarly questions surrounding the computational analysis of moving images. The corpus, for one, could be greatly expanded in order to analyze the same features within and across time and genre. Methodologically, it is also possible to expand on the set of available features. In the visual field of the television screen, computational methods could be developed to assess aspects such as scene changes, character emotions, camera movement, camera angles, and to automatically identify minor and one-off characters. There is also the potential to incorporate audio features ranging from sound effects, music identification, and speaker resolution.

Putting all of these extensions together, the results presented here serve as an example of the possibilities for distant viewing.

[53] David Bordwell, Janet Staiger, and Kristin Thompson. *The Classical Hollywood Cinema* (Columbia University Press, 1987): 13.

[54] Jeremy Butler, *Television: Visual Storytelling and Screen Culture* (Routledge, 2018): 369.

[55] Butler, *Television*, 374.