A Tool for Computational Analysis of Narrative Film

Alexei (Alyosha) Efros
Frederick Alexander Hall
Maneesh Agrawala
Amy Pavel, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-102
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-102.html

August 7, 2018

Acknowledgement

Shannon Davis, Nathaniel Jeppsen, Amy Pavel, Will Crichton, Deepak Warrier, Ian Hall, & Jerica Hall

A Tool for Computational Analysis of Narrative Film

by Alex Hall

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Alyosha Efros
Research Advisor

(Date)

* * * * * * *

Professor Maneesh Agrawala
Second Reader

(Date)

A Tool for Computational Analysis of Narrative Film

Alex Hall, Amy Pavel, Alyosha Efros, Maneesh Agrawala

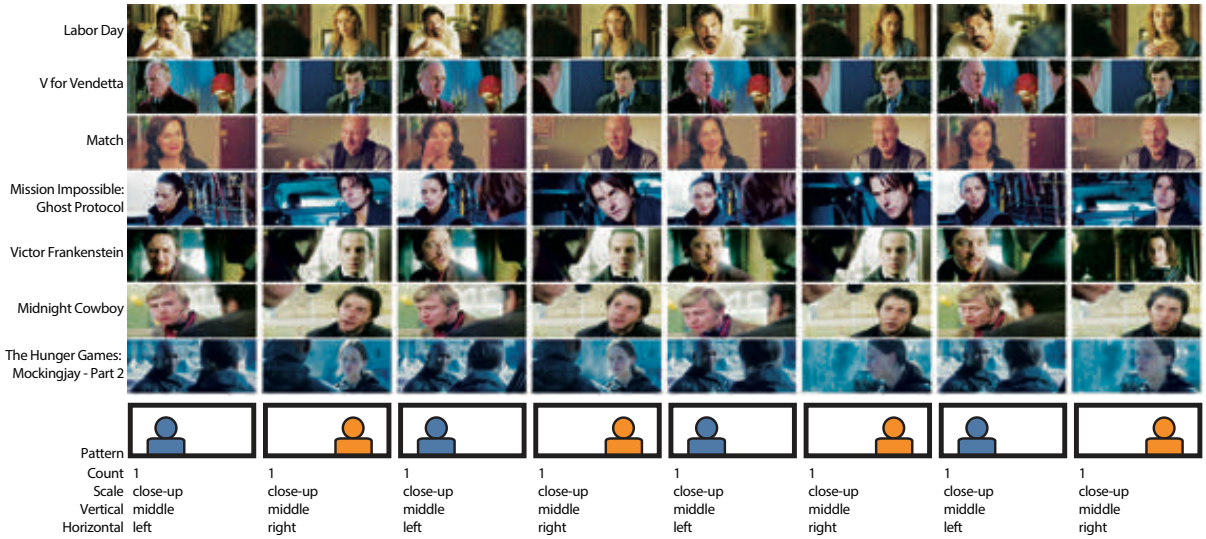

Fig. 1. Filmmakers use certain character framings (e.g. close-ups, long shots) to tell stories. We use our Film Grok pipeline, to automatically label visual and narrative features in video and then use these labels to analyze films for recurring patterns such as common sequences of character framings. Here, we show an 8-shot sequence alternating between left close up and right close up—known as the 'shot-reverse-shot' pattern. Each row is an instance of this sequence found in a different film. We find this same pattern in 75 films. Pattern discovery is only one of the types of analysis Film Grok supports.

Abstract—Film historians and filmmakers study the visual style of past films to answer research questions or gain inspiration for new projects. To help such film professionals conduct large-scale analyses of visual style in films, we present *Film Grok*, a computational tool for labeling and analyzing narrative films. We automatically label a dataset of 620 films with key features of visual style (e.g., character framing, shot sequences) derived from filmmaking texts. To study these features in the broader context of the film, we provide narrative features such as dialogue, emotional sentiment, genre, and director. For example, we use our tools to show that the rise of TV in the 1950's correlates with character framings that are on average 5% closer to the center of the screen and nearly 200% closer to the actor than they were in the 1930's and 40's. We show in another example that Westerns tend to use Extreme Long Shots at moments with 70% stronger negative sentiment than the rest of the film. Akira Kurosawa, a self-proclaimed student of American Westerns, furthers this trend, using Extreme Long Shots for moments with 400% stronger negative sentiment. We train an SVM to classify films based on genre and from this SVM extract the most discriminative shot sequences for each genre. Additionally, we use Film Grok's labels to automatically produce supercuts and *supergrids* highlighting visual features of interest based on user queries.

Index Terms—Narrative film, Digital humanities, Video, Visual analytics, Computer vision, Machine learning, Media arts.

1 Introduction

Filmmakers manipulate elements of a film's visual style, such as character framing and cut timing, to convey narrative structure, actions, and emotions in scenes. Film historians and filmmakers study the visual style of past films to answer research questions [8] or gain inspiration for new projects [2]. Some questions concern a *close-reading* [36] of visual style and therefore require only a limited amount of footage (e.g., how does Orson Welles frame close-ups of Susan Alexander in "Citizen Kane"? [23]); [1] however, many visual style questions use a *distant-reading* [36] and require analysis of a large number of films, making them difficult to answer for most researchers (e.g., 'did shots get shorter from 1915 to 2015?' [17]—they did). and, 'what aspects of visual style distinguish a director or genre's style?' [9]—shot speed and sequences are two examples.

Film researchers can answer close-reading questions about visual style by watching a few films, taking notes, and re-watching relevant scenes. But, quantitatively answering large-scale distant-reading questions remains challenging and time-consuming. In our interviews with film researchers, we found that current methods for answering large-scale questions can limit the number of films analyzed, the variety of queries posed, and the certainty of such investigations. For instance, to study a single question (how shot speed has changed since 1915), one research team labeled shot boundaries in approximately 200 films by scrubbing through the film and manually recording shot boundaries, a process that took 2-11x film runtime [15] or 400-2,200 total hours. Lengthy time requirements discourage large-scale quantitative

- *Alex Hall, Amy Pavel, and Alyosha Efros are with the University of California Berkeley. E-mail: {alexhall,amypavel,efros}@eecs.berkeley.edu.*
- *Maneesh Agrawala is with Stanford University. E-mail: maneesh@cs.stanford.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

[1] Fabe suggests that soft-focus 'shot reverse shots' with Kane and Alexander highlight their emotional isolation.

analysis, and a lack of quantitative evidence can limit the certainty of conclusions.

To help film professionals conduct large-scale analyses of visual style in films, we present *Film Grok*, a computational tool for labeling and analyzing visual and contextual features in narrative film. From filmmaking texts, we derive key elements of visual style (e.g. character framing, shot cut placement and frequency, character count) and context (e.g. scenes, sentiment, director, genre, release date) and use Film Grok to extract and analyze these features. After Film Grok labels these features, users can perform statistical analyses in external visualization tools such as Tableau. We can use Film Grok for visual inspection and search of films, classification and machine learning tasks, and video-based visualizations such as *supercuts* and our proposed *supergrids*. We evaluate Film Grok's support for realistic 'distant' visual style reading by processing 620 films, their scripts, captions, and metadata, and considering 8 open questions in film studies literature. For instance, did filmmakers alter their visual style to cater to TV audiences? We show that during the 1950's, when the percentage of U.S. households with televisions increased from 9% to 87.1% [25, 49], character framings became closer and more centralized. Filmmakers increased scale to ensure that faces had sufficient detail when shown on low-resolution televisions. They brought the characters to the frame's center to ensure that important shot contents remained in the frame when trimming from widescreen cinematic aspect ratios to TV's 4:3; wider character spacings could result in cropping part or all of an actor out of the frame. Do filmmakers get in closer to capture more emotional scenes? We show that filmmakers tend to use close-ups for negative emotions, while they use medium close-ups and medium shots for positive emotions. Do filmmakers repeat character framing patterns across different films? We show that filmmakers reuse the same types of shots (e.g. close-up, long-shot) and even the same character framing sequences (e.g., one 9 shot sequence occurs in 75 films, Figure 1).

Another open area of film studies centers on what comprises the visual style (i.e. visual fingerprint) of a genre or director. For instance, what makes an Alfred Hitchcock film look like an Alfred Hitchcock film? Or, what makes an Adventure film look like an Adventure film? We investigate this by training a one-vs-many multiclass SVM on temporal sequences of character framings—*framing n-grams*—to classify films into 12 genres as well as an additional 'directed by Hitchcock' genre. We find that Comedy films tend to have more shots with 2 people than other genres, while Adventure films make frequent use of medium long shots.

We validate Film Grok's analyses by reproducing 4 results from film studies literature, created by researchers using hand-labeled film data. Finally, we show that using visual style and context features provided by Film Grok, we can automatically produce compilations of short clips—*supercuts*—that highlight visual features of interest based on user queries. We also propose temporally aligned video grid visualizations—*supergrids*—to support direct comparison across videos.

2 Definitions

Cinematography textbooks define a set of techniques for analyzing scripted narrative on film [1, 34, 37]. To support filmmakers and film researchers, we use features derived from these texts.

Temporal units. Cinematography textbooks subdivide films into the units of the *frame* (a single static image), and the *shot* (a contiguous sequence of frames captured from a single camera). Filmmakers compose and arrange shots to create *scenes* that represent a single situation or dialogue with a fixed location and set of characters [1, 34, 37]. We similarly subdivide films into frames, shots and scenes.

Character framing. Film literature considers how filmmakers place and orient the camera to frame the characters. We consider three aspects of *character framing*: *scale* (i.e. how large the actor appears in the frame), *position* (i.e. the 2-D placement of the actors within the frame), and *count* (i.e. the number of actors in the frame).

Shot allocation. Shots are arranged temporally to compose scenes and films. We consider the duration and sequencing of shots and their character framings to compose scenes and films.

Narrative context. We seek to analyze the relationship between a film's cinematography and its narrative, including its plot, characters, and dialogue. To support this analysis, we incorporate additional non-visual signals including captions, aligned scripts, and sentiment analysis. *Captions* provide an on-screen approximation of dialogue. *Scripts* provide filmmakers with guidelines to the film's narrative contents, including details on characters, locations, dialogue, and actions. *Emotional sentiment* approximates the emotions portrayed in the dialogue.

3 Related Work

Cutting et al. [14, 16–18, 21, 51] use manually labeled shot boundaries and framings to perform a number of quantitative analyses of visual style in narrative film and its impact on how humans interpret films. We replicate several of their results on a larger dataset. Breeden et al. [10, 11] performed additional work on human perception of video by gathering eye-tracking data of subjects watching feature films to study the relationship between cinematography and viewer attention. They found that viewers commonly fixate on foreground characters even when they are out of focus. The time-intensive nature of gathering and labeling the data significantly limits the ability to process a large number of films.

Prior works address the time-intensive nature of automatically extract visual features such as shot boundaries and character framings from video. Many works consider shot boundary detection [4,13,33,47]. Avgerinos et al. [3] study the scale of faces in YouTube videos. Benini et al. [6] use 3D reconstruction to compute similar cinematographic scale classifications (e.g. close-up, long shot) and find that Michelangelo Antonioni used an unchanging shot scale distribution throughout his career. We extract similar visual features and add temporal and contextual features (e.g. scenes, dialogue, sentiment, and temporal sequence of visual features).

Some prior works do incorporate temporal and contextual features for video search (e.g. scenes and shot-type sequences) [7, 20, 26, 38, 41, 43–46, 52]. These tools support searching video for specific repeated objects such as clocks and signs [46], actors [7], emotion [52], and dialogue [38, 43, 44]. We apply temporal features (e.g., shot timings and durations) and contextual features (e.g., aligned scripts and caption sentiments) to the task of exploring and analyzing the application of cinematography across a broad set of films.

Other systems consider temporal and visual features of video to address specific video classification and prediction tasks. Some prior works use low-level visual features [42] and Convolutional Neural Networks (CNNs) [45] to address film genre classification. These tools classify films by their genres (e.g. comedy, drama, action). Hou et al. [26] consider the cinematographic properties of shots and their temporal sequence for the specific task of predicting audience 'like/dislike' responses to film trailers. Our system supports similar classification tasks and additional modes of visual feature analysis in film.

4 Feature Extraction

To compute our features, we send our films through the Film Grok pipeline, figure 2. We extract and analyze frames from films to produce shot boundaries and face detections. We classify the face detections based on cinematographic principles and compute a frame classification. Concurrently, we extract emotional sentiment data from captions. To support analyses at the shot level, we bin the caption, sentiment, and frame classification data into shots based on our shot boundary detections. To study how these properties relate to the narrative context of film, we align scripts to captions and produce an *aligned script* containing information about scenes, dialogue, locations, and characters. We combine the aligned script with our shot boundaries to estimate scene boundaries. We store the data at each step in our pipeline, thereby supporting queries across multiple relationships and granularities.

4.1 Dataset

Our dataset includes 620 feature-length films released between 1915 and 2016 spanning 10 genres and 367 directors. With few exceptions, the films were produced in the United States by major production

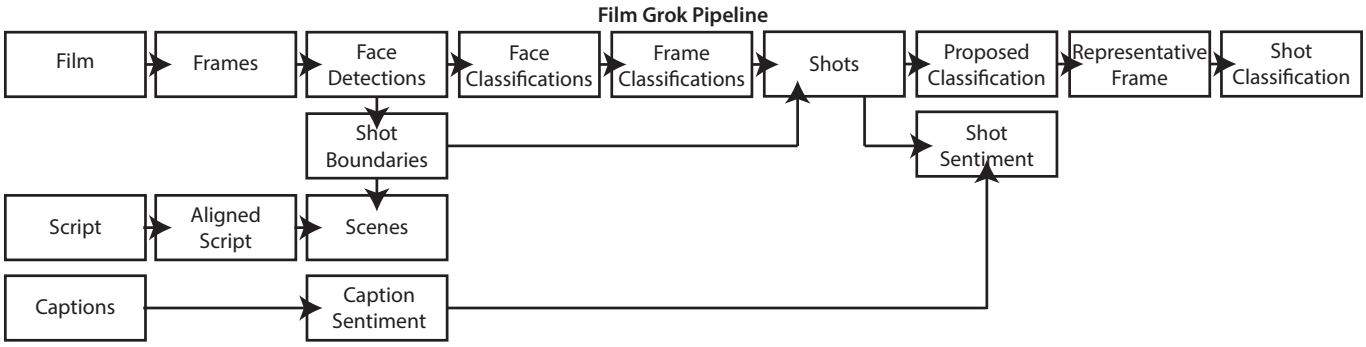

Fig. 2. The Film Grok Pipeline takes as input Films, Captions, and Scripts and visual and narrative labels for use in film research. Throughout the pipeline, we maintain all data for future analysis.

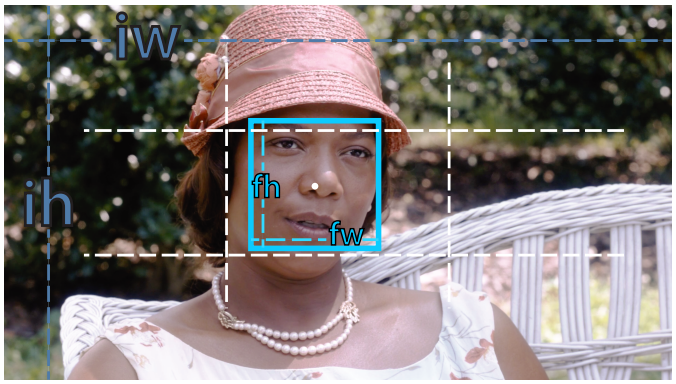

Fig. 3. We compute the height of detected face relative to the height of the frame to compute scale. We compute and store a continuous position values based on the (x,y) coordinates of the face. We also store a discretized position value based on the centroid's position on a 3×3 grid evenly dividing all possible locations in which the centroid could occur.

companies. We have aligned English language captions for 258 of the films from their original DVDs or Blu-Rays. We have 81 scripts retrieved from IMSDB [29] and DailyScript [19]. We scrape metadata for our films from the Open Movie Database (OMDB) [24]. The availability of metadata varies for each film and often includes director, release date, and genre(s). We store all available metadata for each film for later query and analysis. We provide the full list of films in the supplementary materials.

4.2 Shot Boundary Detection

We subdivide the film into shots using shot boundary detection methods from prior work [33]. The detector compares color histograms between adjacent frames and labels the peaks of these differences as shot boundaries. This shot boundary detector processes 1920x1080 progressive scan videos at more than 500 frames per second (i.e. 2 hours of film processed in under 9 minutes) with an average precision of 0.73 on our set of hand-labeled footage sampled from our dataset. By subdividing films into shots, we can perform character framing analysis on shots.

4.3 Face Detection

To capture the camera's placement relative to the characters in a frame, we first detect the faces in all sampled frames (every 24th frame) using the state of the art TinyFace detector [27]. To evaluate the face detector on our dataset, we manually label 1,000 frames with bounding boxes. The detector achieves an average precision of 0.737 on our dataset and processes 1920x1080 progressive scan video at 20fps.

4.4 Character Framing

While we maintain the continuous values from our computed bounding boxes, we also want discretized labels for use in n-gram analysis. Cinematography texts classify shots into discrete categories (e.g. close-up, long shot). Based on this, we discretize our continuous data into cinematographic classifications. Using our face detections, we compute 3 face-level labels: the *count* of people in the frame, the *position* in a 3×3 grid (e.g., Top, Middle, Bottom and Left, Center, Right),and the *scale*(i.e. how much of the character is visible). Because our face detector is limited to the bounds of the frame, bounding box centroids of large faces will occur near the center of the frame. To counteract this limitation of the face detector, we scale the 3×3 grid inversely with the height and width of the face's bounding box—Figure 3. Similar to Leake et al. [32], we compare the face height relative to frame height to assign *scale*. This *normalized face height* is invariant to aspect ratio, an important property when considering large film datasets with inconsistent aspect ratios. We discretize this face:frame ratio into 7 categories: extreme close-up (ECU), close-up (CU), medium close-up (MC), medium shot (MS), medium long shot (ML), long shot (LS), and extreme long shot (ELS).

Cinematography texts provide approximate boundaries between these character framing labels and are not always in precise agreement with one another, leaving thresholds ambiguous [1, 12, 31, 34, 35, 37]. We discretize scale (e.g., close-up, long shot) using thresholds generated by training a linear one-vs-one SVM on our set of 1,000 frames with hand-labeled face bounding boxes and discrete shot scales. The labels and their minimum thresholds are (in descending scale order): ECU 1.0,CU 0.363, MCU 0.258), MS 0.176), MLS 0.074), LS 0.058, and ELS > 0. Our thresholds achieve 82% accuracy compared to human labeled ground truth and are within 1 classification 100% of the time.

Based on all our face-level labels for each frame, we compute *frame-level* labels for all sampled frames. In cases where a frame contains multiple characters, we determine the primary character as the largest face based on Hitchcock's Rule which states that the size of the face in the "frame should equal its importance to the story at that moment" [12]. We apply the scale and position labels from the primary character to the frame. For example, in the case of a frame that contains 2 actors where one is near the camera (i.e. close-up) on the top left of the frame and the other is far from the camera at the bottom right of the frame, we label the frame as a top-left close-up of 2 characters (i.e. 2-TL-CU). Shots with faces comprise 75% of the shots found in our films, excluding credit sequences. If our detector finds no faces in a shot, we label the shot as having a count of zero and no position or scale.

Some distant readings (e.g., are close-ups or long shots used more frequently for strongly emotional scenes?—Figure 10) benefit from having a single classification for each shot. We compute a proposed shot classification as the median count, scale, and position for all frames in the shot. Using these median values directly can lead to shot

classifications that do not match any of the frames in a shot. To prevent this, we use Murch's [37] concept of the 'Representative Frame'—a frame that best summarizes the content of the entire shot. We select the frame most similar to this proposed classification as the representative frame. Finally, we assign the representative frame's labels to the shot while maintaining all frame classifications for use in finer-grained analyses (e.g., finding all face centroids—Figure 4).

4.5 Contextual Features

To supplement the context derived from film metadata (e.g. release date, director, genre), we compute script alignments, dialogue sentiment and scene boundaries. We use an edit-distance based method to align captions to scripts [22, 38]. Like Pavel et al. [38], we find scene boundaries based on caption timing before and after the script's scene headings, which name the scene as interior (INT) or exterior (EXT) and describe the scene's location (e.g., INT. RONALD'S HOUSE, EXT. OUTER SPACE). To compute the corresponding scene boundary in the film, we find the time range between end of the last caption of the previous scene and the first line of the next scene. We label the shot boundary nearest the midpoint of this range as the scene boundary. We extract dialogue sentiment from the film captions using an off-the-shelf tool [28] that assigns text measures of positive, negative, and neutral sentiment. We compute the sentiment of a shot or scene as the mean sentiment of its captions.

4.6 Shot N-Grams

To study how filmmakers select and arrange shots relative to one another and within the larger context of scenes and films, we use an n-gram representation of films. We support temporal analysis by computing the arrangement of ordered shot label sequences of varying length; we call these sequences *n-grams*. In our case, a shot is a term, a film is a document, and our dataset is the corpus. By considering films as documents, we can apply a number of traditional text analysis approaches (e.g. TF-IDF) to draw insights into how filmmakers arrange their films.

5 Results

We evaluate Film Grok's use in film research by answering open questions posed by film literature and replicating 4 existing, manually-produced quantitative results. From our dataset of 620 films, we label 10,055 scenes; 90,413 script lines; 333,347 captions; 704,683 shots; 3,109,894 faces, and 4,194,827 frames. We show the placement of all faces in our dataset in figure 4. We find that shots have gotten shorter (Figure 5) and closer (Figure 6) through the years.

5.1 Novel Results

We surveyed film research literature and found 5 predictions without prior quantitative analysis. Then, we used Film Grok to conduct the relevant query.

Shot Speed. Some film researchers suggest that the speed of shots has increased over the years [9]. We compare the mean shot duration of several genres and find Action, Adventure, and Animation films to have the fastest shots while Westerns, Romances, and Dramas tend to have the slowest shots—Figure 7. We perform the same analysis on several directors from our dataset. We find that David Ayer, Paul Greengrass, and Roland Emmerich use some of the fastest shots while John Ford, Alejandro González Iñárritu, and Steve McQueen use shots that are, on average, more than 3× as long—Figure 8.

Rise of the small screen. "The size of the television screen is small... To show things clearly, you must show them relatively large" [53]. Bordwell predicts that with the rise of TV and VHS, filmmakers started to use more close-ups [9]. The percentage of United States households owning one or more TVs increased from 8% in 1950 to 87.1% in 1960 [25, 49]. We investigate this prediction by plotting the scale of all faces in all frames in our dataset between the years 1915 and 2016—Figure 6. As expected, we see that mean scale has increased.

Bordwell also suggests that the rise of TV caused filmmakers to position characters close to the center of the frame so that all actors would remain visible when the film is cropped for TV aspect ratios.

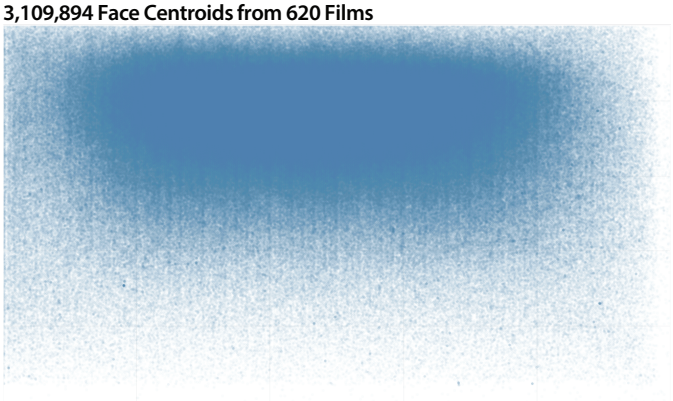

Fig. 4. The centroids of every face detected in every frame in our dataset plotted relative to their precise normalized (x,y) position in the frame. Faces in films cluster above the vertical midpoint of the frame near the horizontal center. We compute position relative to frame dimensions to normalize across different aspect ratios. Films with narrow aspect ratios (e.g. 4:3) produce vertical banding when plotted at a wider aspect ratio because we must stretch the x-axis (e.g., 4 pixels map to 16), and our face detector does not achieve subpixel precision.

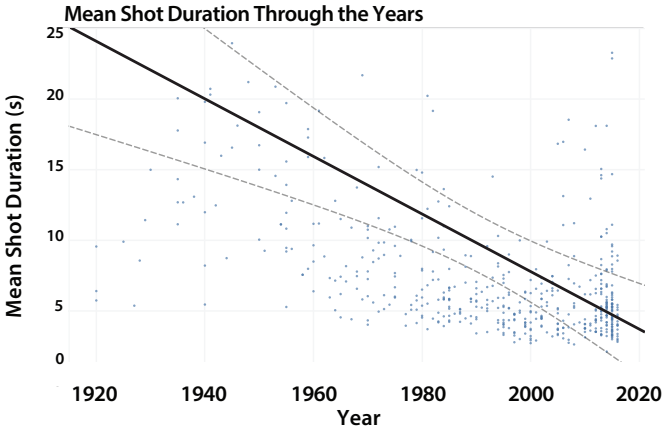

Fig. 5. Filmmakers used increasingly shorter shots over the years. Each dot represents a film, the x-axis encodes its release year, and the y-axis encodes its mean shot duration in seconds. We show a linear regression of our data with a 95% confidence interval indicated by the dashed lines. *Rope* 1948 and *Birdman* 2014 are extreme outliers, with mean shot durations on the order of hours instead of seconds and are therefore omitted from this plot.

Mean distance between the face centroids and the horizontal center of the frame decreases to its nadir in the late sixties after which time the deviation increases to a peak in 2005 (Figure 9). In 1984, RCA released 'Amarcord' and introduced 'letterboxing' in which black bars are placed above and below the film [48]. Later, 16:9 HDTV's came to market and encouraged the adoption of a widescreen format. This introduction of letterboxing gave filmmakers and broadcasters an alternative for displaying films on TV that did not cause widely spaced actors to be cropped from the image.

Sentiment and scale. Bordwell predicts that "the closest shots are reserved for the most significant facial reactions and lines of dialogue" [9]. Based on the idea that expressing emotion is "an important function of dialogue—some would say its most important" [50], we use emotional sentiment as a proxy for line importance and provide analysis of emotional sentiment and character framing (Figure 10). In our analysis, we are concerned with a line's importance relative to the other lines in the

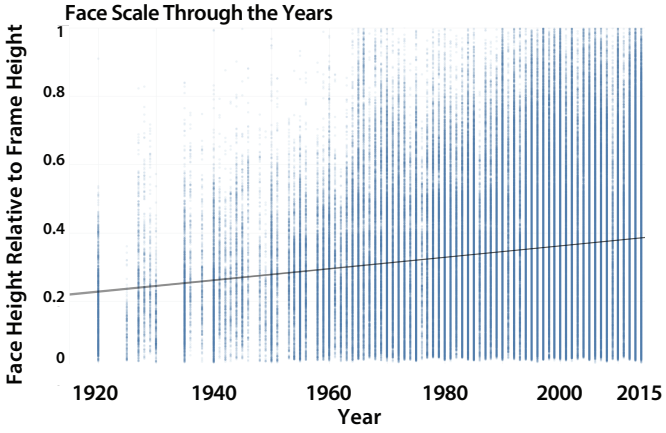

Fig. 6. Low-resolution television led cinematographers to rely on close-ups, as too much detail was lost from long shots. The x-axis encodes the year of that frame's release year, and the y-axis encodes the ratio of the detected face height to frame height. Each dot represents a primary face, and we plot a point for each frame in our dataset that contains a face. We plot points at 3% opacity; the darker the bar, the more samples at that point. Vertical bands appear because we are plotting release date by the year only.

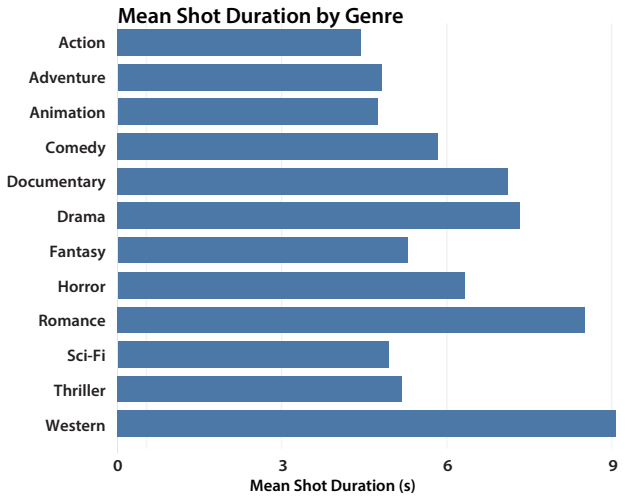

Fig. 7. The mean shot duration for each genre is shown in seconds. We take the mean over all shots in our dataset from films in the given genre. We see that Action, Animation, and Adventure films are the fastest, while Western, Romance, and Drama are the slowest.

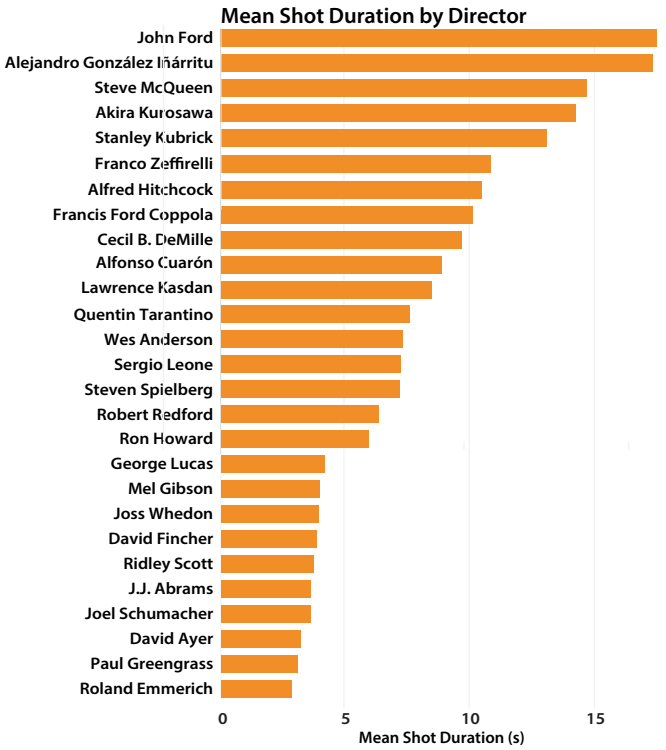

Fig. 8. The mean shot durations for each director, shown in seconds. We take the mean over all shots in our dataset from films by the given director. We see that David Ayer, Paul Greengrass, and Roland Emmerich use short, fast shots while John Ford, Alejandro González Iñárritu, and Steve McQueen use longer, slower shots.

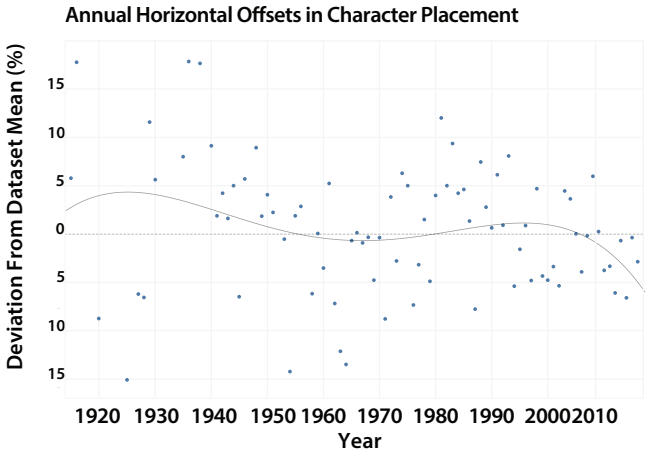

Fig. 9. We plot the mean relative horizontal distance of faces from the frame's center. During the rise of Television in the 1950's, Cinematographers started 'shooting and protecting' by placing the actors in more central locations. This ensured reasonable composition even when a film was cropped for presentation on Television's 4:3 aspect ratio.

film; the absolute sentiment of captions from a comedy may be significantly different than those found in a drama. Therefore, we compare a film's captions to its mean sentiment (i.e. the mean sentiment taken over all captions in the film). We want to compare how shots are used in relation to the mean sentiment, so we subtract the mean from our sentiment values to produce *sentiment deviation*. This deviation can tell us, for instance, that close-ups are used for shots that have $\tilde{7}\%$ more negative emotion and $\tilde{1}.5\%$ less positive emotion than average. We find filmmakers indeed choose close-ups for dialogue that expresses strongly negative sentiment. However, when filming strongly positive dialogue, filmmakers instead prefer the medium close-up. Work in neuropsychology suggests that the smile is the easiest facial expression to identify [39]. Therefore, the higher spatial resolution offered by close-ups may be more beneficial for negative emotions than for positive ones.

While these relationships between sentiment and character framing hold true for our dataset as a whole, we find that some directors and genres stray significantly from this norm (e.g., Kurosawa and Westerns). Film scholars and critics often compare the films of Akira Kurosawa to the American Western, especially the presentation of isolated characters in open spaces [30], and Kurosawa himself said "I have learned from this grammar of the Western." [40]. We use Film Grok to show

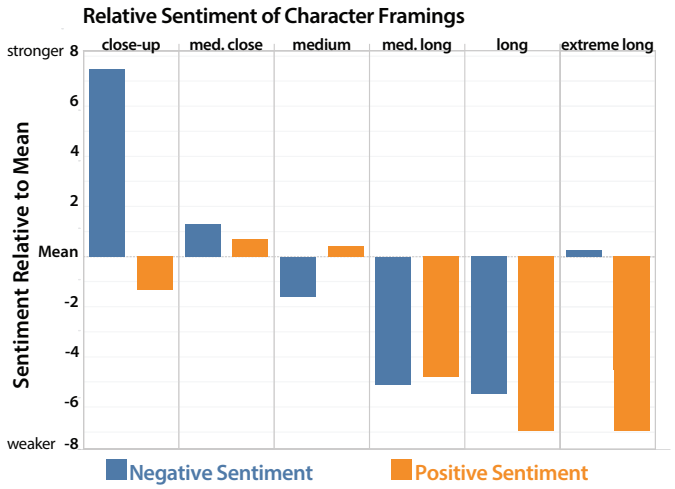

Fig. 10. Positive and Negative Emotional Sentiment relative to the film's mean sentiment (i.e. mean sentiment computed over all captions in the film). The close-up is the preferred character framing for moments of intensely negative dialogue. The medium close-up and the medium shot are preferred for moments of positive dialogue. medium long shots, long shots, are used for generally neutral dialogue (neutral sentiment is the inverse of combined positive and negative sentiment). The extreme long shot is unique in that it has low emotion over all and higher than average negativity.

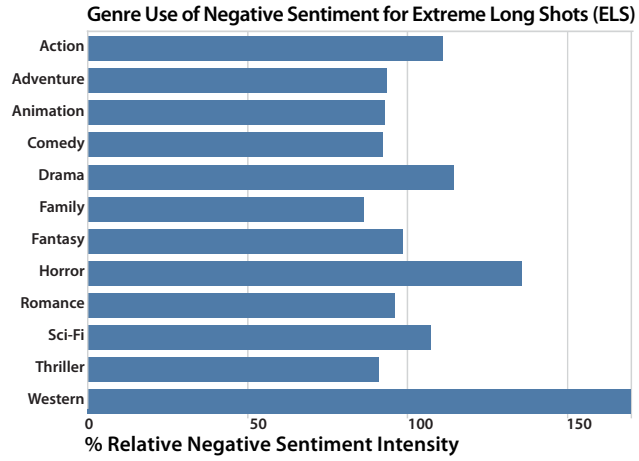

Fig. 11. Westerns are notable for their use of the extreme long shot during moments of intense negative sentiment. This technique is also employed by Akira Kurosawa.

that Westerns and Kurosawa films use extreme long shots in more emotionally negative moments than other directors and genres (Figures 11 and 12). We note that this relationship between the extreme long shot and negative sentiment is counter to the dataset mean shown in figure 10.

5.2 Film grammars

Arijon describes methods for arranging shots to effectively communicate scenes [1]. He describes both how to frame characters in a given shot, but also what sequences of framing should be used given the narrative context. For instance, during a 2-person dialogue, a filmmaker should establish the scene with a wider shot (e.g. long shot) containing both characters. After this shot, the filmmaker should alternate between close framings of one character in the right half of the frame and close framings of the other character in the left side of the frame—Figure

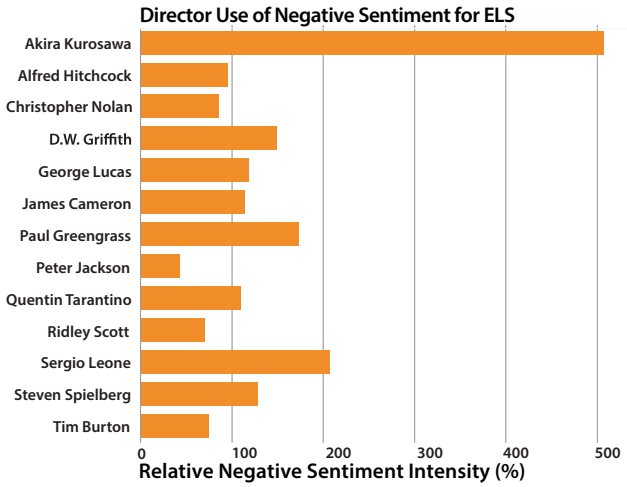

Fig. 12. Some director's make characteristic use of character framings. We see that Akira Kurosawa uses the extreme long shot for moments of intense negative sentiment. Kurosawa shares this technique with the Western genre.

Fig. 13. Examples of shot sequences using 9 centered close-ups, one after another. This pattern was found in 75 films, 26 of which are shown here.

1. We use Film Grok's features to find example shot sequences across multiple films. We do this by computing n-gram frequencies where our labels are character framings, our documents are films, and our corpus is the entire dataset. Using this n-gram analysis, we can search for specific character framing sequences. For example, we show examples of similar n-grams drawn from 75 films—Figures 1 and 13. We can also apply TF-IDF analysis to these n-grams to find sequences that are distinct to a particular group of films (e.g. genre, director).

5.3 Classification

We investigate the stylistic distinctness of film groups using shot classification n-grams as features in an SVM. Because we are interested in discriminative n-grams, we perform TF-IDF analysis where our terms are character framing n-grams, our documents are films, and our corpus is the collection of 620 films. We generate a TF-IDF vector for each film in our dataset and use these TF-IDF weighted feature vectors as inputs to the Genre classification SVM. Additionally, we have 34 of the the 55 feature length films directed by Alfred Hitchcock and include these films as an additional 'Hitchcock' genre. We train One-Vs-One SVMs for 13 genres (including Hitchcock). We compute the

Table 1. Top 3 most discriminative 4-grams per genre. Each item is a shot with the label format: count-scale-position. Count is the number of faces in the shot. Scale is the height of the face relative to the height of the frame with labels (in descending order): close-up (cu), medium close-up (mc), medium shot (ms), medium long shot (ml), and long shot (ls). Position is encoded as the vertical position—(t)op, (m)iddle, (b)ottom— and the horizontal position—(l)eft, (c)enter, (r)ight.

Action	1-cu-ml,1-cu-mr,1-cu-ml,1-cu-mr	1-cu-mc,1-cu-mc,1-cu-mr,2-cu-mr	1-ms-tc,2-ms-mc,1-ms-tc,2-ms-mc
Adventure	1-cu-tc,1-cu-mr,1-cu-mc,1-cu-mc	1-cu-mc,1-cu-mc,1-cu-mc,1-ml-tc	1-ml-tc,1-ml-tc,1-ml-tc,1-ml-tc
Animation	1-cu-tc,1-cu-tc,1-cu-mr,1-cu-mc	2-mc-tc,1-cu-mc,1-mc-mc,1-cu-mc	1-ml-tc,1-mc-tc,1-ms-tc,1-ml-tc
Comedy	2-ms-tc,1-cu-mc,1-cu-mc,1-cu-mc	2-mc-tc,1-cu-mc,1-mc-mc,1-cu-mc	2-cu-mc,2-cu-mc,2-cu-mc,2-cu-mc
Drama	2-ms-tc,1-mc-tc,1-ms-tc,1-mc-tc	1-cu-mc,1-cu-tc,1-cu-mc,1-mc-tc	1-cu-tc,1-cu-mc,1-cu-mc,2-ms-tc
Fantasy	1-ml-tc,1-mc-tc,1-ms-tc,1-ml-tc	1-mc-tc,1-cu-bc,1-cu-mc,1-cu-mc	1-ms-tc,1-cu-tc,1-cu-mc,1-cu-mc
Family	1-ms-tc,1-ml-tc,1-ms-tc,1-ml-tc	1-cu-mc,1-cu-mc,1-ml-tr,1-cu-mc	1-mc-tc,1-mc-tr,1-mc-tc,1-mc-tr
Horror	1-ml-tc,1-cu-tc,1-ml-tc,1-cu-tc	1-cu-mc,1-cu-mc,1-cu-mc,2-ml-tc	1-cu-tc,1-ml-tc,1-cu-tc,1-ml-tc
Romance	2-cu-mc,1-cu-mc,1-cu-mr,2-cu-mc	1-ms-tc,1-cu-mc,1-ms-tc,1-cu-mc	2-cu-br,1-cu-mc,2-cu-ml,1-cu-mc
Sci-Fi	1-ms-tc,1-ms-tc,1-ms-tc,1-ms-tr	2-ml-mc,1-cu-mc,1-cu-mr,1-cu-mc	1-ms-tc,1-ml-tc,1-ms-tc,1-mc-tc
Thriller	1-cu-bc,1-cu-bc,1-cu-bc,1-ml-tc	1-cu-mc,2-cu-mc,1-cu-mc,1-cu-mc	1-cu-mc,1-cu-bc,1-cu-bc,1-cu-bc
Western	1-ms-tc,1-ms-tr,1-ms-tc,1-mc-tc	1-mc-tc,2-ms-tc,1-cu-mc,1-cu-mc	2-cu-bc,1-cu-mc,1-cu-mc,1-cu-mc
Hitchcock	1-cu-mc,2-cu-mc,1-cu-mc,2-cu-ml	1-cu-mc,1-cu-mr,1-cu-tr,1-cu-mc	1-cu-tc,1-cu-mc,1-cu-mr,1-cu-mc

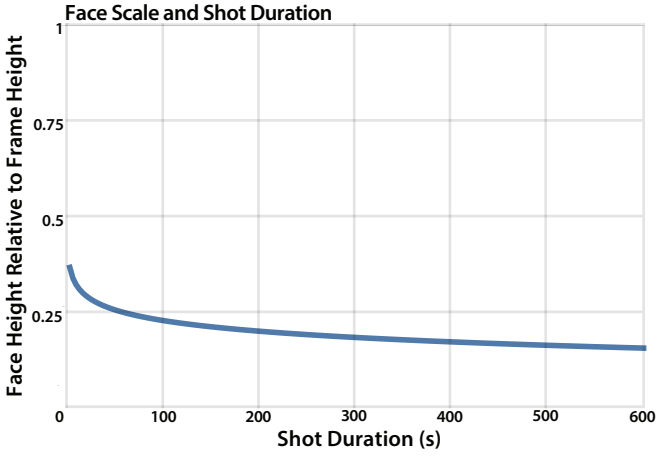

Fig. 14. We plot the correlation between the face:frame ratio and the duration of shots. Higher ratios (e.g. close-ups) tend to be shorter in duration than lower ratios (e.g. long shots).

most discriminative n-grams as the SVM's highest weighted features. Table 1 shows the top 5 most discriminating 4-grams for each genre. Our analysis suggests, for example, that Adventure films are characterized by their use of medium long shots and animation frequently uses medium close-ups instead of close-ups (perhaps due to the low detail of animated faces relative to human faces).

5.4 Replication

To validate our approach, we replicate feature analysis results produced by Cutting et al. for shot lengths, scale, and face positions. The three steps in their manual labeling—shot boundary detection (2-11x), discrete scale classification (1-2x), position labeling (0.25-0.5x)—took a total of 3.25x-13.5x film runtime. They labeled approximately 200 films, which took between 1,300 and 5,400 hours to complete. We reproduce similar quantitative results for shot length, pacing, scale, and actor placement. Our dataset includes 151 of the films from Cutting's analysis and an additional 439 not in Cutting's analysis. We find a similar decrease in shot duration over the years (Figure 5), and we show that closer shots tend to be shorter in duration than shots from farther away 14. Like Cutting, our data suggest that the pacing of shots (their duration and frequency) changes throughout the course of a film and follows a pattern similar to tension arcs commonly used in 'screenwriting'. The supplemental materials contain additional details on our replication of Cutting's findings.

6 Synthesizing Supercuts

To create a *supercut*, "some obsessive-compulsive superfan collects every phrase/action/cliche from an episode (or entire series) of their favorite show/film/game int a single massive video montage" [5]. Using Film Grok, the superfan need no longer be obsessive nor compulsive. We search and extract all clips matching a given style query (e.g. Kurosawa's characteristically emotional long shots.) from our dataset and automatically compile them into a supercut montage. We provide examples of the following supercuts in our supplementary materials: characteristic Kurosawa long shots, characteristic Kurosawa long shots—Figure 15— (highlighting the similarity between Kurosawa's style and that of the American Western), characteristic Kurosawa close-ups, emotionally negative medium shots from Horror films—Figure 16—, and emotionally positive medium close-ups from animated films.

SuperGrids. A supercut, as outlined above, is a montage or sequential arrangement of clips. For some visual style comparisons, especially those involving long sequences of clips, it can be difficult to recall the precise visual style of the previous clip, let alone a set of ten or more clips watched in sequence. To support simultaneous comparison of many sequences, we present the *SuperGrid*, a set of temporally aligned video clips arranged in a 2-dimensional matrix on the screen. We extract each shot using our shot boundaries and scale the playback speed of each shot such that it plays in a fixed duration (e.g. all shots run in 3 seconds). This way, we can compare similar sequences of differing lengths and segmentation. In our supplementary materials, we provide video of the supergrid shown in Figure 17, simultaneously visualizing an 8-gram shot-reverse-shot pattern found in 41 films.

7 Discussion

We present a feature set and system for the analysis of visual style in narrative video. We evaluate our tool by providing evidence for 4 film studies claims that lacked quantitative analysis. We automatically reproduce prior work that relied on manual labeling. We use our features set classify films based on genre and director. And, we provide sample synthesis results based on quantitative visual analysis of narrative video.

Our ongoing research adds new features (e.g., character gender/age, camera motion/angle) to investigate a broader range of social and film studies questions. Future work includes leveraging Film Grok for film synthesis tasks. Film Grok could be incorporated into a tool to recommend character framings at capture time based on an input script. Leake et al. [32] use hand-coded idioms to automatically edit films from multiple takes. The analysis from our system shows that there are distinct visual characteristics of film eras, genres, and directors. This analysis could support the creation of data-driven editing idioms for style emulation.

Fig. 15. Selected stills from a supercut of shots from Kurosawa films that exhibit his characteristic use of long shots during moments of intense negative emotion. We include this and other supercut videos in supplementary materials.

Fig. 16. Selected stills from a supercut of close-ups at moments of strong negative sentiment in horror films.We include this and other supercut videos in supplementary materials.

REFERENCES

[1] D. Arijon. *Grammar of the Film Language*. Silman-James Pr, Los Angeles : Hollywood, CA, reprint edition ed., Sept. 1991.

[2] S. Ascher and E. Pincus. *The filmmaker's handbook: A comprehensive guide for the digital age*. Penguin, 2007.

[3] C. Avgerinos, N. Nikolaidis, V. Mygdalis, and I. Pitas. Feature extraction and statistical analysis of videos for cinemetric applications. In *Digital Media Industry & Academic Forum (DMIAF)*, pp. 172–175. IEEE, 2016.

[4] J. Baber, N. Afzulpurkar, M. N. Dailey, and M. Bakhtyar. Shot boundary detection from videos using entropy and local descriptor. In *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–6. IEEE, 2011.

[5] A. Baio. Fanboy supercuts, obsessive video montages, Apr 2008.

[6] S. Benini, L. Canini, and R. Leonardi. Estimating cinematographic scene depth in movie shots. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pp. 855–860. IEEE, 2010.

[7] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2280–2287. IEEE, 2013.

[8] D. Bordwell. *On the History of Film Style*. Harvard University Press, Cambridge, Mass, Feb. 1998.

[9] D. Bordwell. *The Way Hollywood Tells It: Story and Style in Modern Movies*. University of California Press, Berkeley, Apr. 2006.

[10] K. Breeden and P. Hanrahan. Analyzing gaze synchrony in cinema: a pilot study. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 129–129. ACM, 2016.

[11] K. Breeden and P. Hanrahan. Gaze data for the analysis of attention in feature films. *ACM Transactions on Applied Perception (TAP)*, 14(4):23, 2017.

[12] B. Brown. *Cinematography: theory and practice: image making for cinematographers and directors*. Taylor & Francis, 2016.

[13] C. Cotsaces, N. Nikolaidis, and I. Pitas. Video shot detection and condensed representation. a review. *IEEE signal processing magazine*, 23(2):28–37, 2006.

[14] J. E. Cutting. The evolution of pace in popular movies. *Cognitive Research: Principles and Implications*, 1(1):30, Dec 2016. doi: 10.1186/s41235-016-0029-0

[15] J. E. Cutting. Computational film research tool. Personal Communication, Jan 2018.

[16] J. E. Cutting, K. L. Brunick, and J. E. DeLong. How act structure sculpts shot lengths and shot transitions in hollywood film. *Projections*, 5(1):1–16, 2011.

[17] J. E. Cutting, K. L. Brunick, J. E. DeLong, C. Iricinschi, and A. Candan. Quicker, faster, darker: Changes in hollywood film over 75 years. *i-Perception*, 2(6):569–576, 2011.

[18] J. E. Cutting, J. E. DeLong, and C. E. Nothelfer. Attention and the evolution of hollywood film. *Psychological Science*, 21(3):432–439, 2010.

[19] dailyscript. Daily script. `http://dailyscript.com/`. Accessed: 2018-01-22.

[20] M. Del Fabro and L. Böszörmenyi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia systems*, 19(5):427–454, 2013.

[21] J. E. DeLong, K. L. Brunick, and J. E. Cutting. finding patterns and limits. *The social science of cinema*, p. 123, 2013.

[22] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy–automatic naming of characters in tv video. 2006.

[23] M. Fabe. *Closely watched films: an introduction to the art of narrative film technique*. Univ of California Press, 2014.

[24] B. Fritz. Open movie database api. `https://omdbapi.com/`. Accessed: 2018-01-22.

[25] T. Genova. Television history. `http://www.tvhistory.tv/Annual_TV_Households_50-78.JPG`. Accessed: 2018-03-26.

[26] Y. Hou, T. Xiao, S. Zhang, X. Jiang, X. Li, X. Hu, J. Han, L. Guo, L. S. Miller, R. Neupert, et al. Predicting movie trailer viewer's like/dislike via learned shot editing patterns. *IEEE Transactions on Affective Computing*, 7(1):29–44, 2016.

[27] P. Hu and D. Ramanan. Finding tiny faces. *CoRR*, abs/1612.04402, 2016.

[28] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[29] IMSDB. Internet movie script database. `http://www.imsdb.com/`. Accessed: 2018-01-22.

[30] S. M. Kaminsky. The samurai film and the western. *Journal of Popular Film*, 1(4):312–324, 1972.

[31] S. D. Katz. *Film directing shot by shot: visualizing from concept to screen*. Gulf Professional Publishing, 1991.

[32] M. Leake, A. Davis, A. Truong, and M. Agrawala. Computational video editing for dialogue-driven scenes. 2017.

[33] J. Mas and G. Fernandez. Video shot boundary detection based on color histogram. *Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST*, 2003.

[34] J. V. Mascelli. *The Five C's of Cinematography: Motion Picture Filming Techniques*. Silman-James Pr, Los Angeles, 1st silman-james press ed edition ed., June 1998.

[35] G. Mercado. *The filmmaker's eye: Learning (and breaking) the rules of cinematic composition*. Taylor & Francis, 2011.

[36] F. Moretti. *Distant reading*. Verso Books, 2013.

[37] W. Murch. *In the blink of an eye: A perspective on film editing*. Silman-James Press, 2001.

[38] A. Pavel, D. B. Goldman, B. Hartmann, and M. Agrawala. Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 181–190. ACM, 2015.

Fig. 17. Stills of the first 4 shots taken from an 8-gram supergrid video. We create supergrids by finding similar shot sequences (n-grams) in multiple films, temporally aligning their shots, and playing them back in a grid layout. We include the video in supplementary materials.

[39] M. T. Posamentier and H. Abdi. Processing faces and facial expressions. *Neuropsychology review*, 13(3):113–143, 2003.

[40] S. Prince. *The warrior's camera: the cinema of Akira Kurosawa*. Princeton University Press, 1999.

[41] Z. Rasheed and M. Shah. Movie genre classification by exploiting audio-visual features of previews. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2, pp. 1086–1089. IEEE, 2002.

[42] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, 2005.

[43] R. Ronfard. Reading movies: an integrated dvd player for browsing movies and their scripts. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 740–741. ACM, 2004.

[44] R. Ronfard and T. T. Thuong. A framework for aligning and indexing movies with their script. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 1, pp. I–21. IEEE, 2003.

[45] K. Sivaraman and G. Somappa. Moviescope: Movie trailer classification using deep neural networks. *University of Virginia*, 2016.

[46] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, pp. 127–144. Springer, 2006.

[47] A. F. Smeaton, P. Over, and A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[48] E. Smith. The controversial history of letterboxing for movies on your tv, Apr 2016. Accessed on 03-26-2018.

[49] C. Steinberg. *TV Facts*. Facts on File, 1986.

[50] J. Thomas. *Script analysis for actors, directors, and designers*. CRC Press, 2013.

[51] H.-Y. Wu, Q. Galvane, C. Lino, and M. Christie. Analyzing elements of style in annotated film clips. In *WICED 2017-Eurographics Workshop on Intelligent Cinematography and Editing*, p. 7, 2017.

[52] M. Xu, J. S. Jin, S. Luo, and L. Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*, pp. 677–680. ACM, 2008.

[53] H. Zettl. *Television production handbook*. Wadsworth Pub. Co, Belmont, Calif, 1976.