Gaze Data for the Analysis of Attention in Feature Films

KATHERINE BREEDEN, Stanford University
PAT HANRAHAN, Stanford University

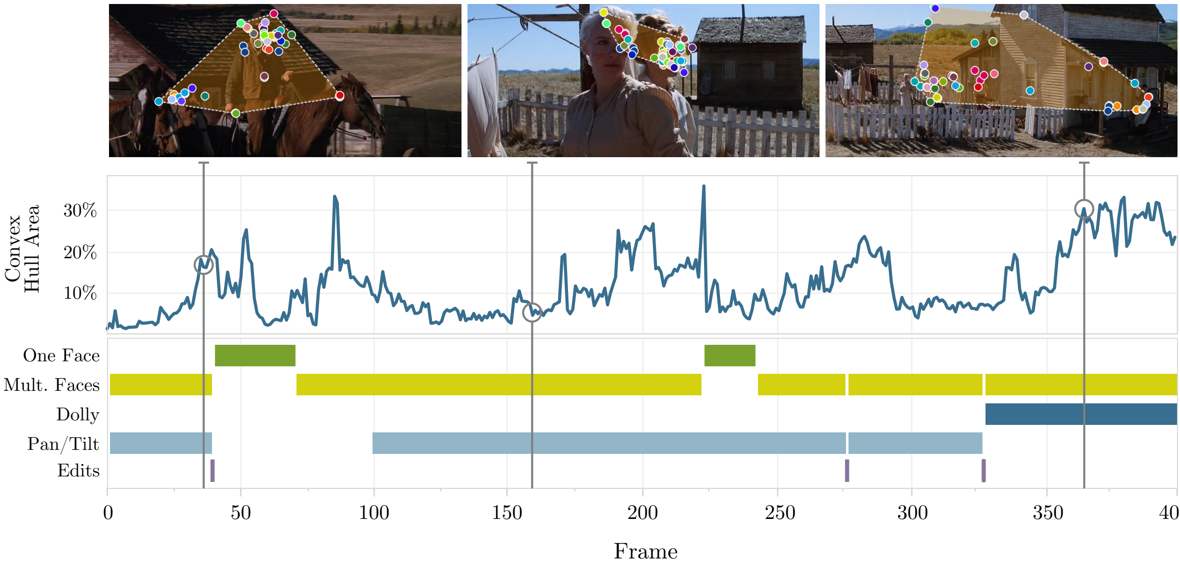


Fig. 1. Ensemble gaze behavior for one of the film clips in our data set. Top: Example frames with recorded gaze points colored by participant. Dashed boundary shows convex hull. Middle: Convex hull area varies over time. Smaller values indicate increased attentional synchrony. Bottom: Colored bars indicate the presence of hand coded features, e.g. a single face (dark green) or camera motion such as pans and tilts (light blue). Film: *Unforgiven* (Warner Bros., 1992).

Film directors are masters at controlling what we look at when we watch a film. However, there have been few quantitative studies of how gaze responds to cinematographic conventions thought to influence attention. We have collected and are releasing a data set designed to help investigate eye movements in response to higher level features such as faces, dialogue, camera movements, image composition, and edits. The data set, which will be released to the community, includes gaze information for 21 viewers watching 15 clips from live action 2D films, which have been hand annotated for high level features. This work has implications for the media studies, display technology, immersive reality, and human cognition.

CCS Concepts: • **Computing methodologies** → **Perception**; • **Applied computing** → *Media arts*; • **Human-centered computing** → Empirical studies in HCI;

Additional Key Words and Phrases: Eye tracking, gaze behavior, gaze direction, psychophysics, film studies

This work was supported by Stanford's Google Graduate Fellowship Fund and the National Science Foundation, Grant CCF-1111943.
The definitive Version of Record was published in *ACM Transactions on Applied Perception*, https://doi.org/10.1145/3127588.

ACM Reference format:
Katherine Breeden and Pat Hanrahan. 2017. Gaze Data for the Analysis of Attention in Feature Films. *ACM Transactions on Applied Perception* 1, 1, Article 1 (January 2017), 14 pages.
https://doi.org/10.1145/3127588

1 INTRODUCTION

The human eye is constantly in motion. As the eye moves, areas of visual interest are sequentially centered on the densest region of photoreceptors in the retina, called the fovea. This allows the visual system to piece together a detailed representation of the world around us. The characterization of eye movements is of great interest to a wide variety of researchers, as they provide something rare: a measurable external indication of attention. set of fields including cognitive science, psychology, and neurology. For example, involuntary eye motion and pupillometry is used to measure cognitive load [Klingner 2010; Stuyven et al. 2000] and can help characterize conditions such as schizophrenia [Levy et al. 2010] and autism spectrum disorder [McPartland et al. 2010].

A deeper understanding of the mechanisms behind various eye movements is of practical use as well. In computer graphics, systems designed for a single user can reduce computational load by employing real-time eye tracking to monitor gaze location and adaptively deliver high-resolution imagery only to the most sensitive part of the retina [Guenter et al. 2012; Patney et al. 2016]. With sufficiently high-performance eye tracking, the reduction in peripheral detail can be made imperceptible to the user. Such systems, known as foveated particular interest for mobile and augmented reality applications, as computational efficiency is critical for providing acceptable quality in a portable form factor.

Thus far, advances in foveated rendering have been limited by the unpredictable and ballistic nature of eye movements. Better knowledge of where people look in visually rich, dynamic environments would enable predictive models of gaze and attention and reduce reliance on high-performance eye tracking. Several recent efforts to improve our understanding of gaze response have focused on eye and head movements in virtual reality environments, in response to both virtual reality movies [Serrano et al. 2017] as well as omnidirectional panoramas [Sitzmann et al. 2016].

Another potential wealth of knowledge in this area comes from filmmakers, who have spent over a century exploring the communication of time and space through moving pictures. The development of modern cinematographic techniques has resulted in the emergence of a rich compendium of best practices thought to direct a viewer's attention (e.g., [Alton 2013; Block 2001; Mascelli 1998]). These conventions are known to evolve over time and as filmmaking technologies change; for example, Cutting and colleagues have identified gradual image darkening and decreased average shot length over time [Cutting 2016; Cutting et al. 2011]. Such methods are critical for producing films that are readily understood, as cinema is a complex dynamic audiovisual stimulus that contains a broad range of information for viewers to comprehend. This information includes human interactions, dialogue, narrative structure, and the spatial configurations of characters and environs.

Cinematographic conventions encompass image composition, the meaning of temporal discontinuities (edits) in different contexts, camera movements, and beyond[1]. However, there has been little quantitative analysis of how gaze responds to this type of highly structured dynamic content. In order to provide the community with a resource suitable for investigating the influence of high-level image features and cinematographic techniques on gaze position, we contribute the following data set. It contains recorded gaze information for hand curated film clips, which have been augmented via the annotation of selected high-level features. In particular, the data set contains:

- 15 film clips, of duration 1–4 minutes each, selected to represent a variety of visual and editing styles

[1] The authors recommend Bordwell and Thompson's *Film Art* [2012] for a comprehensive introduction.

- Recorded gaze behavior for 21 viewers watching those clips
- Frame by frame annotations of high-level cinematographic features

This data set is easily extendable, either with additional hand annotations, or with existing methods from machine vision. The following sections describe the collection of the data set and outline its potential uses.

2 RELATED WORK

Generally speaking, eye movements can be divided into two categories: endogenous, or internally produced, and exogenous, or externally produced [Gompel et al. 2007]. Endogenous eye movements, such as scanpaths, are shaped by both high-level cognitive processes such as search task [Yarbus 1967], as well as by neurological conditions such as injuries to the visual cortex [Jahnke et al. 1995; Pambakian et al. 2000]. Exogenous (stimulus-driven) eye movements are reflexive and affected by dynamic stimuli in a variety of complex ways. For instance, small changes to dynamic stimuli can affect saccade onset latency and saccade trajectory [Doyle and Walker 2001; Saslow 1967]. Exogenous saccades can also be induced via subtle modulations in the periphery of still images [Bailey et al. 2009; McNamara et al. 2009]. In other words, gaze location is partially directed by dynamic image content and not under the conscious control of the viewer.

Many previous studies using eye tracking to probe visual attention have focused on static images; these studies have explored the influence of both low and high-level image features. Low-level features include local image variations such as contrast, color, and edges. These types of image features have been shown to affect gaze; for example, by manipulating depth of field in order to direct fixations into the parts of an image which are in focus [Cole et al. 2006]. Low-level image features combine to form higher-level features such as people, faces, and text. Faces are especially well known to be important targets for gaze and attention (e.g., [Buchan et al. 2007; Crouzet et al. 2010; Haxby et al. 2000]). In video and film, size, duration, and number of faces have been shown to influence patterns of eye movements [Cutting and Armstrong 2016; Rahman et al. 2014]. In general, both low and high-level features affect gaze. This observation is behind the data-driven image saliency model of Judd and colleagues [2009]; it was enabled by gaze recordings for still images, and incorporates both low and high-level features.

Research focusing on gaze behavior in response to dynamic imagery has employed a variety of stimuli, such as video games, documentary footage, and narrative films [Li et al. 2010; Peters and Itti 2007]. Gaze response in these settings is challenging to study because, in most realistic scenarios, visual task is unavoidably intertwined with both low and high-level image features. This results in a mix of exogenous and endogenous factors effecting gaze.

Consequently, it is striking that examinations of gaze behavior in response to feature films have shown a great deal of similarity between viewers [Goldstein et al. 2007]. Their agreement may be due in part to a reduced diversity of endogenous influences on gaze location: when watching a film, each viewer is performing a common task, namely, following the narrative. When multiple viewers attend to the same image region simultaneously, this is known as *attentional synchrony*. Attentional synchrony is known to correlate with low-level image features such as motion, contrast, and flicker [Mital et al. 2011; Smith and Mital 2013].

Gaze response to higher-level image characteristics in moving images is less well understood. Therefore, the data set described in this paper has been designed to contribute to the understanding of higher-level features in promoting attentional synchrony during the viewing of dynamic content.

3 FILM SELECTION AND FEATURE ANNOTATION

Directors, cinematographers, and film editors exercise tight control over their content, facilitating viewer comprehension with the use of conventions in both spatial (e.g., image composition) and temporal (e.g., editing) structure. We selected films likely to furnish good examples of these techniques. From these, we selected a diverse set of representative clips which would provide a variety of visual settings, emotional tones, and temporal rhythms. High-level features that could be readily annotated were identified from previous work in both film theory and eye tracking with still imagery.

3.1 Film Selection

Because filmmaking is such a highly structured art form, even very different films can be expected to resemble one another in terms of their craft. To ensure our films exemplified high quality filmmaking, technical excellence was inferred by selecting from the nominees and winners of these film industry awards:

- American Cinema Editors Awards (The Eddies)
- The American Society of Cinematographers Awards; Category: Theatrical Release
- The British Society of Cinematographers Awards; Category: Theatrical Release
- The Hollywood Foreign Press Association Awards (The Golden Globes); Category: Best Director
- Academy of Motion Picture Arts and Sciences Awards (The Oscars);
  Categories: Film Editing, Directing, Cinematography

Films represented in these awards were then given a point score comprised of 1 point for each win and 0.5 points for each nomination. This score was normalized by the total number of awards given in each release year. Of the top 100 highest scoring films, 13 were chosen. To minimize the effect of changes in filmmaking conventions over time, these films all come from the years 1984–2014, and were approximately evenly spaced throughout that period. As we wanted low-level image features to be as similar as possible, all films are full color and live action. To minimize any bias due to individual style, we also selected for unique directors, cinematographers, and editors.

3.2 Clip Selection

Clips were chosen from candidate films with an eye towards maximizing the presence of features that distinguish narrative films from other types of dynamic imagery found in previous studies. Clips were sought which would be compelling even out of context and would not reveal key plot points. They were also selected to express the wide range of aesthetics found in modern cinema. For example, although all clips are in color, color palettes range widely from highly desaturated (*Saving Private Ryan*) to vibrant (*Slumdog Millionaire*, *The Last Emperor*). Some color palettes are naturalistic (*Amadeus*, *Unforgiven*), while others are more stylized (*Birdman*).

In nearly all modern films, dialogue is a key component in advancing the story. We sought scenes in which dialogue would be easily understood and engaging, but would not require additional knowledge of the film's plot. The data set includes clips with no dialogue (*No Country For Old Men*, Clip 2), monologue (*Amadeus*, *The King's Speech*, and *Shakespeare in Love*, Clip 2), and multi-character dialogue (*Argo*).

Of course, dialogue on screen is almost always coincident with the presence of human faces. This data set contains examples of single face close-ups (*The Departed*, *The Kings Speech*) and small groups of characters (*Argo*, *Birdman*). There are also scenes with larger numbers of faces on screen, such as ensemble dance numbers (*Chicago*), and crowd scenes (*Gladiator*, and *Shakespeare in Love*).

Clips were also selected in part to illustrate a variety of common filmmaking techniques, such as composition, editing, and camera motion. For instance, dialogue scenes are frequently accompanied by the shot/reverse shot pattern. This technique establishes an imaginary line segment connecting two characters. With the camera restricted to one side of the line, the filmmakers alternate between shots of each endpoint (character). Examples can be found in clips from *The Departed* and *No Country For Old Men* (Clip 1). We also include a range of editing styles; the effect of fast paced editing on gaze behavior can be examined in the clip from *Gladiator* (fight sequence), the opening chase scene from *Slumdog Millionaire*, and the musical number from *Chicago*. Other clips have much more modest tempos (e.g., *Shakespeare in Love*, Clip 2). Finally, both stationary (*The Departed*) and more dynamic (*Saving Private Ryan*, *Birdman*) approaches towards camerawork are represented.

Category	Coding	Description
Faces	`f frameX frameY`	A single face is visible for the specified range of frames: `[frameX, frameY]`
	`ff frameX frameY`	Multiple faces are visible
	`fa frameX frameY`	One or more non-human faces are visible
Dialogue	`don frameX frameY`	Dialogue; speaking character is on-screen
	`doff frameX frameY`	Dialogue; speaking character's face is not visible
Camera	`pt frameX frameY`	Pan, tilt, or a combination of the two
	`d frameX frameY`	Camera is moving via dolly
	`z frameX frameY`	Camera operator is zooming in or out
	`cr frameX frameY`	Crane; both camera base and mount are in motion
	`h frameX frameY`	Jerky motion of the base and/or mount, as if handheld
	`r frameX frameY`	Racking focus (i.e., focal plane in motion)
Edits	`c frameX frameY`	There is a (plain) cut between frames `frameX` and `frameY`
	`xf frameX frameY`	There is a cross-fade between frames `frameX` and `frameY`

Table 1. Binary features included in data set. For terminology, see text.

Each clip is 1–4 minutes in duration, which participants reported to be adequate time to feel immersed in the narrative. Two films, *No Country For Old Men* and *Shakespeare in Love* contributed two clips each. These provided additional instances of crowd scenes and shot/reverse shot dialogue. In total, the data set contains 15 clips from 13 films for a total of about 38 minutes of content. A table containing a brief description of each clip can be found in Appendix B[2].

3.3 Hand Annotation of Features

Typically, films focus on human stories and therefore contain many salient high-level image features, such as faces, along with other, potentially salient technical and narrative factors such as camera motion and dialogue. Unlike many low-level features such as edges, contrast, and optical flow, higher-level features like these can be difficult to accurately detect automatically. To aid in the exploration of higher-level image features and their influence on gaze, we hand coded each frame for the presence or absence of the following:

Faces: Separate codes are used to indicate zero, one, or multiple faces; we also identify frames that contain non-human faces such as animals or masks.

Dialogue: Codes specify frames in which dialogue is being delivered and whetherthe speaker's face is visible.

Camera Motion: Several types of camera motion are differentiated. These are pan/tilt (azimuthal and/or vertical rotation of a camera with fixed base), dolly (translation of the camera along a fixed path, without changes in elevation), crane (combined translation and elevation changes), zoom (change in focal length), rack focus (change in focal plane), and handheld (erratic perturbations of the camera).

Edits: We code for two common editing techniques. The first is the most common: a discontinuity in time, space, or action in which two shots are joined sequentially such that the last frame of the first shot immediately precedes

[2] Resources for data set reproduction, including DVD imprint information, clip start times, and first and last frames are available at http://graphics.stanford.edu/~kbreeden/gazedata.html.

the first frame of the second shot. The second is the cross fade; here, frames from two shots are overlaid for a short period of time, during which the preceding shot fades out and the subsequent shot fades in. Cross fades are identified by a transition window starting with the first frame of the fade out and ending with the final frame of the fade in[3].

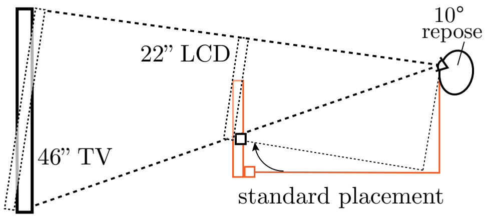
Table 1 contains a listing of these features as they are found in the data set. These are meant to provide a useful starting point for the analysis of gaze behavior in feature films. As computer vision algorithms continue to improve, it is our hope that the feature set will expand. An illustration of our hand coding is provided in Figure 1.

4 COLLECTION OF GAZE DATA

In order to facilitate eye movements that would match natural viewing conditions as closely as possible, the experimental setup was designed to approximate the typical home cinema experience. Gaze locations were recorded for 21 volunteers aged 22–73 (mean 34, 11 female). Clips were presented in a different randomized order to each viewer, and sound was delivered by external speaker or using headphones, as preferred.

4.1 Eye Tracker Configuration

Gaze location was recorded using a Gazepoint GP3 table top eye tracker[4]. This unit samples at 60 Hz and has an accuracy of between 0.5 and 1°, and is designed to be colocated with a display that is 24” or smaller. However, as noted by Troscianko and colleagues, both objective and subjective measures indicate that engagement increases with absolute screen size, even as field of view remains constant [2012]. In other words, absolute screen size must be taken into account when a laboratory environment is meant to promote an immersive viewing experience.



With this in mind, we displayed experimental stimuli on a 46” Sony Bravia wide-format television[5] at a distance of 1.8 m. The eye tracker was mounted on a portable, adjustable arm at a horizontal distance of 65 cm from the participant, with eye tracker height adjusted such that the bottom of the screen was just visible over the top of the eye tracker. This setup mimics the visual field of a 22” monitor colocated with the eye tracker. Note, it was necessary to recline the viewer 10° in order to raise the television to a workable height; this arrangement also helped reduce head motion.

4.2 Participant Instructions and Calibration

During set-up, participants were made as comfortable as possible to minimize movement. A neck pillow was used to stabilize the head, and most chose to elevate their feet. Once positioned, the eyetracker was calibrated using the 5-point least squares routine packaged with the Gazepoint software. The accompanying visualization of the calibration result was used to determine whether the setup was adequate; corrections of greater than 1° prompted the experimenters to reposition the participant and recalibrate. Calibration results were similar to those observed in the intended set-up. Participant instructions are reproduced in full in Appendix A.

Once the experiment began, advancement to the next film clip was self-paced via keyboard controls. A post-experiment questionnaire was administered to record which films had been seen previously by each viewer; see supplementary materials for details.

[3] In our data set, cross fades are not common. For the purposes of calculating shot length, the midpoint of the cross fade is used.

[4] https://www.gazept.com/product/gazepoint-gp3-eye-tracker

[5] Resolution is 1080p. Model number: KDL-46EX400.

4.3 Gaze Data Post Processing

After gaze information was recorded for all participants, the raw data was analyzed to identify several sources of error. These are classified as follows:

Eye tracker error: Due to blinks or other tracking errors, some gaze points were flagged as invalid by the eye tracker at record time (<1% of all data).

Gaze point off-screen: Gaze points recorded as falling outside of the frame — either outside the screen entirely or within the screen but inside frame letterboxing — were flagged as invalid (2.6% of all data).
Invalid subject for clip: Occasionally, the eye tracker would lose tracking for certain viewers[6]. As experimenters were able to monitor eye tracker performance in real time and adjustments were possible between trials (i.e., individual clips), it was not necessary to invalidate data collected for all clips for these participants. Instead, we invalidate all gaze points recorded for trials in which the error rate for a specific participant exceeded a given threshold. For the results shown in this paper, we use a threshold of 10%.

5 DATA SET OVERVIEW

This section provides a descriptive overview of the data set, including preliminary observations related to our hand annotated features.

5.1 Shot Length and Editing

Using our hand-coded edits, we can examine editing characteristics of our film clips. Excluding the clip from *Birdman*, which was edited to give the impression of a continuous take, the average shot duration in our data set is 3.7 seconds. Table 2 illustrates the overall distribution of shot lengths, as well as the maximum, minimum, and average shot length for each clip. Notably, of the 609 total shots in the data set, only 18 (3%) are 13 seconds

[6] This was most commonly due to the presence of eyeglass glints.

Film	Total Shots	Shortest Shot (s)	Longest Shot (s)	Average Shot Length (s)
Gladiator	105	0.3	19.8	1.8
Slumdog Millionaire	88	0.3	6.0	1.9
Chicago	65	0.3	10.6	2.3
Argo	69	1.0	10.0	3.1
Unforgiven	45	1.2	12.2	3.3
The Departed	29	0.5	8.7	3.3
No Country For Old Men (2)	14	1.6	9.1	4.0
The King's Speech	40	1.4	25.0	4.5
Amadeus	42	1.2	16.3	4.7
No Country For Old Men (1)	24	1.4	15.8	5.0
Saving Private Ryan	34	0.8	18.6	5.1
Shakespeare In Love (1)	14	1.6	15.2	5.8
The Last Emperor	21	1.3	47.3	6.4
Shakespeare In Love (2)	18	0.6	32.8	7.0
Birdman	1	246.3	246.3	246.3

Table 2. Shortest, longest, and average shot length (ASL) for each clip.

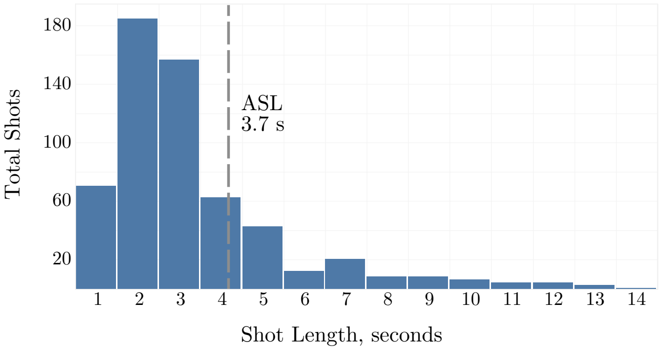


Fig. 2. Shot length distribution for all clips. Of all shots in our data set, 85% are 4 seconds or shorter. Histogram is truncated at 14 seconds, omitting a long tail which includes the clip from *Birdman*.

or longer. By contrast, 519 shots (85%) are 4 seconds or shorter[7]. Figure 2 contains a histogram illustrating the relative frequency of different shot durations for our set of clips.

5.2 Distribution of Valid Gaze Points

The process to determine gaze point reliability described in Section 4.3 voids approximately 5.7% the 2.6M recorded gaze points. This yields an average of 46 valid gaze points per frame, with between 19 and 21 unique viewers per clip.

Aggregate analysis of the spatial distribution of these gaze points reveals subtle differences when compared to eye tracking still images. Previous experiments have demonstrated that gaze point density is highest at image center for static images [Judd et al. 2009] (Figure 3, left). This pronounced bias has been attributed to compositional choices, as photographers are likely to center their images on salient features. By superimposing all recorded gaze locations for the film clips in our data set, it is apparent that the region with peak gaze point density is not the screen center, but rather just above the center (Figure 3, right). This asymmetry suggests that directors and cinematographers may be utilizing compositional conventions which differ slightly from those commonly employed by still photographers.

5.3 Attentional Synchrony

A number of ways have been proposed to measure attentional synchrony. Bivariate contour ellipse models (e.g., [Goldstein et al. 2007; Ross and Kowler 2013]) fit a single ellipsoid to the recorded gaze points, and are appropriate when there is a single area of interest and gaze points are well fit by an ellipsoid. Others have measured synchrony by comparing the scan paths taken by different viewers [Dorr et al. 2010; Meur and Baccino 2012; Taya et al. 2012]. These methods are useful when comparing the gaze behavior of multiple viewers over a period of time suffient in duration to include multiple saccades and fixations. Mital and colleagues [2011] use Gaussian mixture models to quantify attentional synchrony; Gaussians are fit to each gaze point, and high synchrony is associated with low cluster covariance. This method is well suited for situations in which robustness to noise is an issue.

Due to the aforementioned lack of temporal and spatial fidelity in our data, as well as the relatively small number of individuals recorded, we sought a measure which would be more conservative in the face of outliers. In

[7] The trend over time towards shorter average shot length described by Cutting and others (e.g. [Cutting 2016; Cutting et al. 2011]) is also visible in our data set; see our supplementary materials for a plot showing the relationship between ASL and release year.

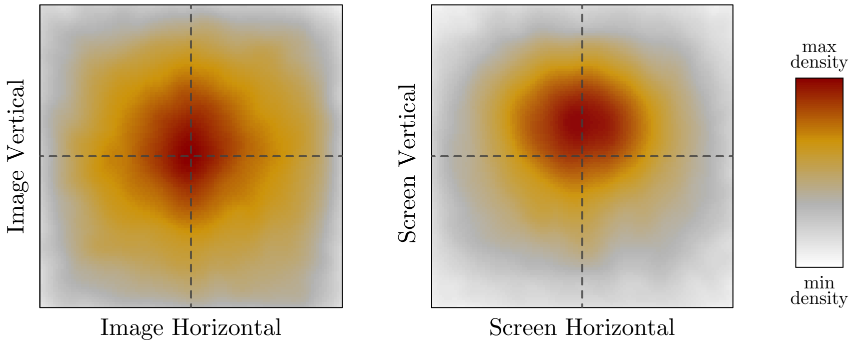


Fig. 3. Heat maps depict the distribution of all gaze points in the respective data set. (left, [2009]) compared with ours (right, scaled to square). shown in red. Left: still images used by Judd and colleagues [2009] produce gaze locations with a strong central tendency. Right: Our data (scaled to square) shows a peak gaze point density slightly above frame center, indicating that image composition in cinema can be distinguished from that in still images previously used in eye tracking experiments. As both experiments were preceded by centered fixation targets, the left plot excludes gaze points prior to the first fixation event; right plot excludes data collected during the first second of each clip.

other words, we wanted to preserve the possibility that a handful of outliers might indeed represent a meaningful region of interest. Additionally, we did not want to make assumptions about the shape of the regions of interest. Therefore we measure the size of the screen region attended to as the area of the convex hull of all valid gaze points.

Due to its sensitivity to outliers — which have the potential to drastically increase the convex hull area — this measure should be viewed as an upper bound on attentional synchrony. Even so, we observe that, on average, viewers attend to only a small portion of the total screen area. This corroborates previous work by Goldstein and colleagues [2007]. The data set average for convex hull area is 11.0%; the clip with minimum average convex hull area was *The Departed* (7.0%) and the maximum was *The Last Emperor* (14.8%).

Figure 1 illustrates an example use of this data set to examine changes in metrics associated with ensemble gaze behavior. It shows frame-wise variations in the gaze point convex hull area for a short section of the clip from *Unforgiven*. Colored bars (Figure 1, below) denote the presence of hand coded features for the indicated range of frames. While the ballistic nature of eye movements makes the measure of convex hull area inherently noisy, some observations can be made. First, we see that faces — whether human or horse — are likely gaze targets. Image composition and the number of highly attractive targets determine whether gaze points will produce a diffuse collection (Figure 1, example frame, left), or tight clusters (Figure 1, example frame, center). The impact of camera motion can be studied similarly. For instance, the dolly shot in Figure 1 (dark blue bar, example frame at right) appears to be associated with reduced attentional synchrony.

5.4 Gaze Response to Faces

Of the approximately 54,000 frames in our data set, hand annotations indicate that 41% contain a single face, 35% contain multiple faces, and 23% of frames contain no faces. As previously observed in still images, viewers' gazes are strongly attracted to human faces. In our experiments, high gaze synchrony frequently correlates with the presence of one or more faces; for example, the clip with lowest average gaze point convex hull area, *The Departed*, also had one or more faces present in every frame. Attraction to faces is especially noticeable across cut boundaries; when present, a face will almost always be the first target following a cut.

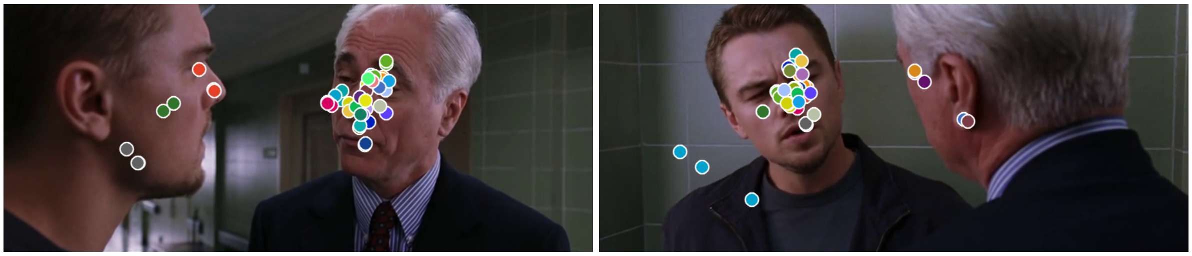


Fig. 4. During shot/reverse shot dialogue, viewers may fixate on the foreground character even when the visible character is speaking. Gaze points colored by participant. Film: *The Departed* (Warner Bros., 2006)

5.5 Dialogue

Dialogue was coincident with 37% of data set frames; in 27% of frames the speaker's face was visible (coded as "Dialogue On Screen"), and in 10% of the frames it was not (coded as "Dialogue Off Screen"), typically due to voiceover or because the speaker was facing away from the camera. Interestingly, during shot/reverse shot dialogue, it was not uncommon for viewers to fixate on the foreground character's ear or head, even when they are out of focus and/or typically salient regions of the face are obscured (Figure 4).

5.6 Camera Motion

Hand coding can be used to categorize frames according to any camera motion present, in order to examine its effect on gaze behavior in isolation. As an illustration of this, Figure 5 shows overlays of all recorded gaze points for frames in the data set containing pans and tilts, categorized by the direction of motion, either leftward or rightward. A comparison of the density of gaze points reveals what appears to be anticipatory clustering in the direction of motion; for example, pans to the left lead to higher gaze point density on the left hand side of the image. Lateral bias during pans and tilts can be contrasted with the overall upper-center bias shown in Figure 3. We hypothesize that filmmakers may be successfully using camera motion as a subtle cue to guide attention.

6 DISCUSSION AND FUTURE WORK

This paper introduces a data set containing human gaze data in response to short film clips, along with frame-by-frame annotations of camera motion, shot boundaries, faces, and dialogue. The clips were curated to provide strong examples of filmmaking techniques, and the gaze data has been post-processed to flag erroneous or suspect gaze points. It is our hope that this data will enable a variety of experiments to examine the relationship between gaze position and dynamic image content, the cognitive processes of following a narrative, and the impact of various filmmaking techniques on gaze response.

However, a number of important limitations remain, particularly concerning the Gazepoint GP3 eye tracker. This unit, while affordable and easy to use, does not have the fine spatial accuracy needed to investigate the influence of low-level image features. A more accurate eye tracker, paired with film clips transcoded from higher resolution sources, might allow for more detailed analyses in future experiments. Similarly, the relatively low sampling rate (60Hz) prohibits the classification of microsaccades; better temporal fidelity would also be helpful in understanding fine-scale motion of the eye.

Looking forward, we can investigate whether combinations of features reinforce one another or provide conflicting cues about which region of the screen is most engaging at a given instant in time. For instance, we might ask whether dialogue provides an additional signal drawing the gaze towards a speaking character's face, or if it promotes additional saccades towards other characters on screen. Additionally, the cognitive process of

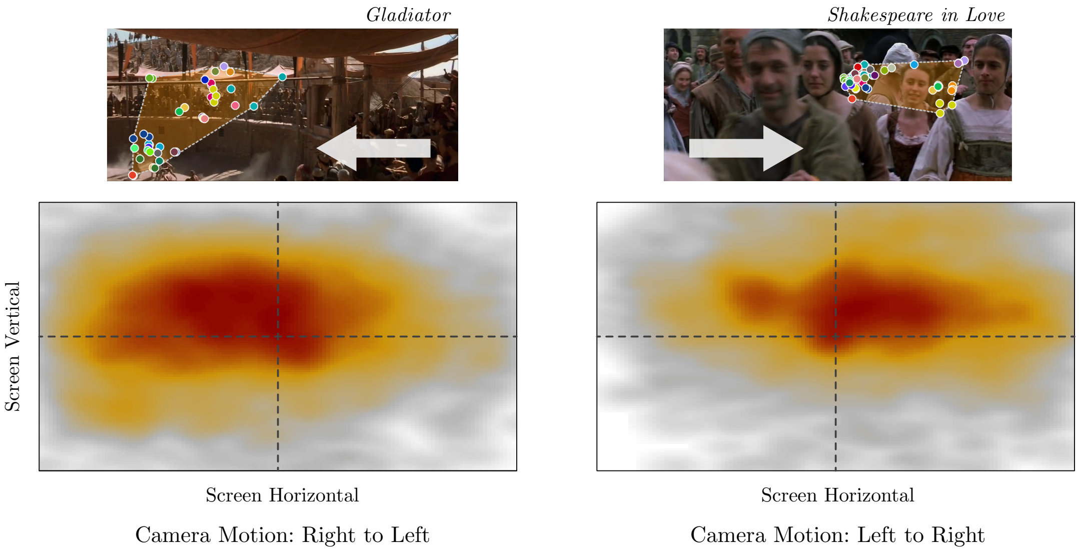


Fig. 5. Lateral camera motion is associated with anticipatory clustering in the direction of the pan; heat maps show gaze point density as in Figure 3. Below: pan and dolly shots moving leftward (left) and rightward (right). Above: example frames demonstrate the effect. Films: *Gladiator* (Dreamworks, 1999) and *Shakespeare in Love* (Miramax, 1998).

following the narrative should be examined. How narrative influences gaze patterns, for example, in situations of repeat viewing, is an opportunity for further study[8].

For cinema scholars, this data could represent an additional source of information in understanding the evolution of filmmaking conventions. One such progression that has been well established is the reduction of average shot length. The effect that this trend has on gaze location and target acquisition time is something that could be measured directly from data sets like this one. This kind of evidence has implications for establishing data-driven rules of thumb for filmmakers and film editors.

Finally, we encourage others to augment the feature set presented here; these could be hand annotated or even generated using automatic techniques. Aligning clips with their conformed screenplays, which has previously been shown effective in the automatic recognition of human actions from video [Marszalek et al. 2009], could enable the use of methods from Natural Language Processing to automatically discern mood and tone and correlate these with image features and gaze behavior. Similarly, pixel-level local motion (optical flow) could be measured within these clips and be used to correlate image dynamism with gaze fixations and smooth pursuit.

Looking ahead, an improved understanding of where people look in information rich, dynamic environments would impact multiple areas of computer science. Predictive models of gaze behavior could drive adaptive compression methods for streaming videos in order to preserve detail within regions where gaze fixations are likely. Predictive models might also support more efficient foveated displays, and dynamic user interfaces could be made easier to parse, with improved usability and accessibility. Clearly, there is much to be done towards understanding the complex ways in which visual content, gaze direction, and attention are interconnected.

[8] For researchers interested in these effects, information on which viewers had previously viewed the films used is provided in the supplementary materials.

ACKNOWLEDGMENTS

The authors would like to thank our experimental participants as well as the team at Gazepoint for their assistance with our unusual set up. We also thank Tilke Judd and colleagues for making their data available. Video encoding and frame extraction were performed using FFmpeg (http://ffmpeg.org/) and HandBrake (https://handbrake.fr). This work was supported by Stanford's Google Graduate Fellowship Fund and NSF Grant CCF-1111943.

REFERENCES

John Alton. 2013. *Painting with Light.* University of California Press.

Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. 2009. Subtle Gaze Direction. *ACM Trans. Graph.* 28, 4, Article 100 (Sept. 2009), 14 pages.

Bruce Block. 2001. *The Visual Story: Seeing the Structure of Film, TV, and New Media.* Focal Press.

David Bordwell and Kristin Thompson. 2012. *Film Art: An Introduction.* McGraw-Hill Education.

Julie N Buchan, Martin Pare, and Kevin G Munhall. 2007. Spatial Statistics of Gaze Fixations During Dynamic Face Processing. *Soc Neurosci* 2, 1 (2007), 1–13.

Forrester Cole, Doug DeCarlo, Adam Finkelstein, Kenrick Kin, Keith Morley, and Anthony Santella. 2006. Directing Gaze in 3D Models with Stylized Focus. *Eurographics Symposium on Rendering* (June 2006), 377–387.

Sébastien M. Crouzet, Holle Kirchner, and Simon J. Thorpe. 2010. Fast Saccades Toward Faces: Face Detection in Just 100 ms. *Journal of Vision* 10, 4 (2010), 16.

James E Cutting. 2016. The Evolution of Pace in Popular Movies. *Cogn Res Princ Implic* 1, 1 (2016), 30.

James E Cutting and Kacie L Armstrong. 2016. Facial Expression, Size, and Clutter: Inferences from Movie Structure to Emotion Judgments and Back. *Atten Percept Psychophys* 78, 3 (Apr 2016), 891–901.

James E Cutting, Kaitlin L Brunick, Jordan E Delong, Catalina Iricinschi, and Ayse Candan. 2011. Quicker, Faster, Darker: Changes in Hollywood Film Over 75 Years. *Iperception* 2, 6 (2011), 569–576.

M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth. 2010. Variability of Eye Movements when Viewing Dynamic Natural Scenes. *Journal of Vision* 10, 10 (aug 2010), 28–28. https://doi.org/10.1167/10.10.28

M Doyle and R Walker. 2001. Curved Saccade Trajectories: Voluntary and Reflexive Saccades Curve Away from Irrelevant Distractors. *Exp Brain Res* 139, 3 (Aug 2001), 333–344.

Robert B Goldstein, Russell L Woods, and Eli Peli. 2007. Where People Look When Watching Movies: Do All Viewers Look at the Same Place? *Computers in Biology and Medicine* 37, 7 (07 2007), 957–964.

Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill (Eds.). 2007. *Eye Movements: A Window on Mind and Brain* (1st edition ed.). Elsevier.

Brian Guenter, Mark Finch, Steven Drucker, Desney Tan, and John Snyder. 2012. Foveated 3D Graphics. *ACM Trans. Graph.* 31, 6, Article 164 (Nov. 2012), 10 pages.

James V. Haxby, Elizabeth A. Hoffman, and M. Ida Gobbini. 2000. The Distributed Human Neural System for Face Perception. *Trends Cogn Sci* 4, 6 (Jun 2000), 223–233.

M T Jahnke, P Denzler, B Liebelt, H Reichert, and K H Mauritz. 1995. Eye Movements and Fixation Characteristics in Perception of Stationary Scenes: Normal Subjects as Compared to Patients with Visual Neglect or Hemianopia. *Eur J Neurol* 2, 4 (Sep 1995), 275–295.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to Predict Where Humans Look. In *IEEE International Conference on Computer Vision (ICCV).*

Jeff Klingner. 2010. *Measuring Cognitive Load During Visual Tasks by Combining Pupillometry and Eye Tracking.* Ph.D. Dissertation. Stanford University.

Deborah L. Levy, Anne B. Sereno, Diane C. Gooding, and Gilllian A. O'Driscoll. 2010. Eye Tracking Dysfunction in Schizophrenia: Characterization and Pathophysiology. *Curr Top Behav Neurosci* 4 (2010), 311–347.

Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. 2010. Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video. *International Journal of Computer Vision* 90, 2 (2010), 150–165.

M. Marszalek, I. Laptev, and C. Schmid. 2009. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.* 2929–2936. https://doi.org/10.1109/CVPR.2009.5206557

Joseph V. Mascelli. 1998. *The Five C's of Cinematography: Motion Picture Filming Techniques.* Silman-James Pr.

Ann McNamara, Reynold Bailey, and Cindy Grimm. 2009. Search Task Performance Using Subtle Gaze Direction with the Presence of Distractions. *ACM Trans. Appl. Percept.* 6, 3, Article 17 (Sept. 2009), 19 pages.

James C. McPartland, Sara Jane Webb, Brandon Keehn, and Geraldine Dawson. 2010. Patterns of Visual Attention to Faces and Objects in Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 41, 2 (may 2010), 148–157.

Olivier Le Meur and Thierry Baccino. 2012. Methods for Comparing Scanpaths and Saliency Maps: Strengths and Weaknesses. *Behavior Research Methods* 45, 1 (jul 2012), 251–266. https://doi.org/10.3758/s13428-012-0226-9

Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson. 2011. Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. *Cognitive Computation* 3, 1 (2011), 5–24.

A L M Pambakian, D S Wooding, N Patel, A B Morland, C Kennard, and S K Mannan. 2000. Scanning the Visual World: A Study of Patients with Homonymous Hemianopia. *Journal of Neurology, Neurosurgery & Psychiatry* 69, 6 (2000), 751–759.

Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. 2016. Towards Foveated Rendering for Gaze-tracked Virtual Reality. *ACM Trans. Graph.* 35, 6, Article 179 (Nov. 2016), 12 pages.

Robert J. Peters and Laurent Itti. 2007. Congruence Between Model and Human Attention Reveals Unique Signatures of Critical Visual Events. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Curran Associates Inc., USA, 1145–1152.

Anis Rahman, Denis Pellerin, and Dominique Houzet. 2014. Influence of Number, Location and Size of Faces on Gaze in Video. *Journal of Eye Movement Research* 7, 2 (2014).

N. M. Ross and E. Kowler. 2013. Eye Movements While Viewing Narrated, Captioned, and Silent Videos. *Journal of Vision* 13, 4 (mar 2013), 1–1. https://doi.org/10.1167/13.4.1

M. G. Saslow. 1967. Effects of Components of Displacement-Step Stimuli Upon Latency for Saccadic Eye Movement. *J. Opt. Soc. Am.* 57, 8 (Aug 1967), 1024–1029.

Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. 2017. Movie Editing and Cognitive Event Segmentation in Virtual Reality Video. *ACM Transactions on Graphics (SIGGRAPH 2017)* 36, 4 (2017).

Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, and Gordon Wetzstein. 2016. Saliency in VR: How do people explore virtual environments? *CoRR* abs/1612.04335 (2016). http://arxiv.org/abs/1612.04335

T. J. Smith and P. K. Mital. 2013. Attentional Synchrony and the Influence of Viewing Task on Gaze Behavior in Static and Dynamic Scenes. *Journal of Vision* 13, 8 (jul 2013), 16–16. https://doi.org/10.1167/13.8.16

E Stuyven, K Van der Goten, A Vandierendonck, K Claeys, and L Crevits. 2000. The Effect of Cognitive Load on Saccadic Eye Movements. *Acta Psychol (Amst)* 104, 1 (Mar 2000), 69–85.

Shuichiro Taya, David Windridge, and Magda Osman. 2012. Looking to score: the dissociation of goal influence on eye movement and meta-attentional allocation in a complex dynamic natural scene. *PLoS One* 7, 6 (2012), e39060. https://doi.org/10.1371/journal.pone.0039060

Tom Troscianko, Timothy S. Meese, and Stephen Hinde. 2012. Perception While Watching Movies: Effects of Physical Screen Size and Scene Type. *i-Perception* 3, 7 (2012), 414–425.

Alfred L. Yarbus. 1967. *Eye Movements During Perception of Complex Objects*. Springer US, Boston, MA, 171–211.

A PARTICIPANT INSTRUCTIONS

The following starting instructions were given to each participant:

In this experiment, you will be asked to watch clips from a variety of feature films. Your eye movements will be recorded and these measurements will be analyzed in aggregate with that of other participants. Your name and other identifying information will not be associated with this data.

Please be advised that the types of eye movements we are interested in are generally not under conscious control. Therefore, you should view each clip as naturally as possible. Do your best to follow the plot and dialogue, as you would do in a movie theater or in your own home.

The clips you are about to see were selected in part because they did not require extensive knowledge of the surrounding plot. Many are taken from near the beginning of the film, when the intended audience would not have any additional insight than you do today. None of the clips reveal critical plot information. A list of the films used will be available for you to take home after the experiment.

You will see 15 clips from 13 films, and the total duration is approximately 30 minutes. In between each clip, you will have the opportunity to rest and readjust yourself. To begin, you will see a screen containing the name of the film, followed automatically by a small crosshairs in the center of the screen. Please focus your attention at the X until the video begins, at which point you should watch the clip normally. The next segment of the experiment will only begin after you press the Spacebar. If you need to adjust the volume, you can use the keyboard controls. At any time, you may press the Spacebar to exit the clip.

B CLIP EXTRACTION AND DESCRIPTIONS

Film clips were transcoded from DVDs using HandBrake (https://handbrake.fr, H264 codec at 23.976 fps). Individual frames were extracted using FFmpeg (http://ffmpeg.org). Table 3 contains a description of the cinematographic features present in each clip. Information necessary for the generation of our clips can be found in the supplementary materials and at http://graphics.stanford.edu/~kbreeden/gazedata.html.

Film	**Clip Details**		**Description**
Amadeus (1984)	*Director:* *Cinematographer:* *Duration:*	Miloš Forman Miroslav Ondříček 3:16 (4699 frames)	Salieri learns that Mozart has been involved with the woman he loves. Contains: close-up, monologue, dialogue.
Argo (2012)	*Director:* *Cinematographer:* *Duration:*	Ben Affleck Robert Prieto 3:35 (5157 frames)	Mendez arrives and lays out his plan to get the Americans home. Contains: shots with 3+ characters, multi-character dialogue.
Birdman (2014)	*Director:* *Cinematographer:* *Duration:*	Gonzalez Iñaritu Emmanuel Lubezki 3:29 (5864 frames)	Shortly after Riggan arrives at rehearsal, an accident occurs on stage. He makes a hasty exit. Contains: multi-character dialogue, complex camera movements.
Chicago (2002)	*Director:* *Cinematographer:* *Duration:*	Rob Marshall Dion Beebe 2:30 (3590 frames)	Velma performs “All That Jazz!” as Roxie watches, captivated. Contains: close-ups, shots with 3+ characters, rapid editing.
The Departed (2006)	*Director:* *Cinematographer:* *Duration:*	Martin Scorsese Michael Ballhaus 1:36 (2304 frames)	Billy learns he won’t be allowed to become a police officer and reflects on his family and their history of crime. Contains: close-ups, shot/reverse shot dialogue.
Gladiator (1999)	*Director:* *Cinematographer:* *Duration:*	Ridley Scott John Mathieson 3:03 (4386 frames)	An eager crowd cheers as gladiators battle in an arena. Contains: crowd scenes, rapid editing, complex camera movements.
The King’s Speech (2010)	*Director:* *Cinematographer:* *Duration:*	Tom Hooper Danny Cohen 3:01 (4293 frames)	The Duke of York struggles to deliver a message from the King. Contains: close-ups, monologue, dialogue.
The Last Emperor (1987)	*Director:* *Cinematographer:* *Duration:*	Bernardo Bertolucci Vittorio Storaro 2:12 (3189 frames)	3-year-old Puyi is introduced in the Forbidden City. Contains: crowds, complex camera movements.
No Country For Old Men (2007)	*Director:* *Cinematographer:* *Duration 1:* *Duration 2:*	Joel and Ethan Coen Roger Deakins 0:56 (1345 frames) 1:59 (2855 frames)	Clip 1: A sheriff and deputy discuss a mysterious murder. Contains: shot/reverse shot dialogue. Clip 2: Anton Chigurh creates a diversion at the pharmacy. Contains: close-up.
Saving Private Ryan (1998)	*Director:* *Cinematographer:* *Duration:*	Steven Spielberg Janusz Kamiński 2:53 (4141 frames)	American soldiers land on Omaha beach. Contains: shots with 3+ characters, complex and handheld camera movements.
Shakespeare In Love (1998)	*Director:* *Cinematographer:* *Duration 1:* *Duration 2:*	John Madden Richard Greatrex 1:20 (1920 frames) 2:05 (2987 frames)	Clip 1: Theatergoers of all stripes stream into the Globe to see “Romeo and Juliet”. Contains: crowds. Clip 2: Viola steps in for the final scenes of the play. Contains: crowds, close-up, monologue.
Slumdog Millionaire (2008)	*Director:* *Cinematographer:* *Duration:*	Danny Boyle Anthony Dod Mantle 2:43 (3910 frames)	Young children playing cricket attract the ire of local police; a chase ensues. Contains: shots with 3+ people, rapid editing, complex camera movements.
Unforgiven (1992)	*Director:* *Cinematographer:* *Duration:*	Clint Eastwood Jack N. Green 2:26 (3486 frames)	Mike and Davey lead a string of ponies into Big Whiskey. Davey offers one to Delilah as a gesture of goodwill. Contains: shots with 3+ people, dialogue.

Table 3. Descriptions of the clips used in our data set.