Behavior Research Methods (2023) 55:2940–2959
https://doi.org/10.3758/s13428-022-01949-7

Where to look at the movies: Analyzing visual attention to understand movie editing

Alexandre Bruckert[1] · **Marc Christie**[1] · **Olivier Le Meur**[1]

Accepted: 29 July 2022 / Published online: 24 August 2022

Abstract
In the process of making a movie, directors constantly care about where the spectator will look on the screen. Shot composition, framing, camera movements, or editing are tools commonly used to direct attention. In order to provide a quantitative analysis of the relationship between those tools and gaze patterns, we propose a new eye-tracking database, containing gaze-pattern information on movie sequences, as well as editing annotations, and we show how state-of-the-art computational saliency techniques behave on this dataset. In this work, we expose strong links between movie editing and spectators gaze distributions, and open several leads on how the knowledge of editing information could improve human visual attention modeling for cinematic content. The dataset generated and analyzed for this study is available at https://github.com/abruckert/eye_tracking_filmmaking

Keywords Eye-tracking · Film editing · Visual saliency

Introduction

In order to deal with the incredibly large amount of data coming from our visual environment, human beings, as most animals, have developed a biological mechanism called overt visual attention. While watching a scene, the eye makes sudden noncontinuous movements called saccades, only stopping during events called fixations. Eye fixations occur so that regions of visual interest are centered on the densest zone in photoreceptors of the retina, called the fovea. This area is heavily packed with cone cells, which allows maximum visual acuity, even if it only represents approximately one degree of the visual field. Several studies have shown that eye fixations and visual attention are closely associated (Findlay, 1997). Therefore, studying gaze pattern is of great interest in a wide range of fields (Duchowski, 2002; Zhang et al., 2020). In image and video processing, visual attention and saliency have been widely used in compression algorithms (Yu & Lisin, 2009; Zünd, Pritch, Sorkine-Hornung, Mangold, & Gross, 2013; Hadizadeh & Bajic, 2014). In the medical field, eye-tracking devices are used to help the communication in cases of locked-in syndrome (Majaranta & Räihä, 2002) or for diagnosis purposes (Harezlak, Kasprowski, Dzierzega, & Kruk, 2016); for more applications in medicine, see for instance Harezlak and Kasprowski (2018).

The factors explaining where people look in a video are usually divided into two categories: *bottom-up* and *top-down* factors. Top-down characteristics refer to observer-dependent properties, such as the age of the observers, their cultural background, or the task at hand. These factors have been shown to be the cause of sometimes-extreme discrepancies in gaze patterns; see for instance Le Meur et al., (2017) for an exploration of the age factor, or Chua, Boland, and Nisbett (2005) and Rayner, Castelhano, and Yang (2009) for the cultural parameter. Bottom-up factors refer to stimuli characteristics, such as the spatial properties of the visual scene or the temporal characteristics of a video. It also includes the implicit properties of the stimuli, such as the presence of faces (Cerf, Harel, Einhaeuser, & Koch, 2008) or text in the scene. Most of the visual attention models are purely bottom-up models, meaning that they only extract information from the stimulus. Indeed, bottom-up visual saliency has proven to be a reliable predictor of fixations location in images (Borji, Sihite, & Itti, 2013).

Over the last century, filmmakers have developed an instinctive knowledge of how to guide the gaze of the

✉ Alexandre Bruckert
alexandre.bruckert@irisa.fr

[1] University of Rennes 1, IRISA, CNRS, Rennes, France

audience, manipulating bottom-up characteristics, such as visual cuts, camera movements, shot composition and sizing, and so on Smith, Levin, and Cutting (2012), and Bordwell, Staiger, and Thompson (1985). This empirical knowledge contributes to building a set of cinematographic rules and conventions designed to accommodate the artistic intention of the director with the perception of the audience. However, formalizing these visual tools is not an easy task, and several frameworks and languages have been proposed (Ronfard, Gandhi, & Boiron, 2013; Wu, Galvane, Lino, & Christie, 2017; Wu, Palù, Ranon, & Christie, 2018). Such languages help quantifying common cinematographic rules, and allow automated models to be used in the movie production process.

As a consequence, studying the quantitative perceptual effects of the visual tools available to filmmakers is of great interest, both for understanding the way humans perceive movies and also for the filmmakers themselves, who could get quantitative feedback on the effects of their work and techniques. Understanding the mechanisms underlying the visual attention on movies can also be of help for computational models related to movie production, such as automated camera placement, automated editing, or 3D animated scenes design.

In this paper, we extend the work of Breeden and Hanrahan (2017) by proposing a new eye-tracking database on 20 different movie clips, of duration 2–7 min each. For each clip, we provide cinematographic feature annotations drawn from Wu, Galvane, Lino, and Christie (2017), such as the camera movement and angle, the framing size, and the temporal location of cuts and edits. Alongside a comprehensive analysis of the collected data, we expose several strong correlations between high-level cinematographic features and gaze patterns, which can be easily used to improve visual attention modeling. We also perform a benchmark of visual attention models on this database, and we show that state-of-the-art models often struggle to grasp and use these high-level cinematographic characteristics. Finally, we discuss several leads on how that information could be included in human visual attention models in order to improve their performances on cinematic content.

Related work

In this section, we provide a quick overview of the recent works in visual attention modeling, and especially bottom-up approaches, such as visual saliency modeling. We then give a very brief review of the field of visual attention in the context of cinematography, and of the databases available to conduct such studies.

Modeling visual attention

As mentioned earlier, eye movements rely on two kinds of attention mechanisms: top down (or endogenous) influences, which are shaped by high-level cognitive processes, such as the task at hand, the cultural background of the observer, or its medical condition, and bottom-up (or exogenous) movements, which are driven by the features of the stimulus itself. The most common way of representing attention, whether it is endogenous or exogenous, is through a representation called *saliency map*, which is a distribution predicting the likelihood of an eye fixation to occur at a given location. In this work, we will mostly focus on this representation, even if it is not the only one, nor does it capture the full range of human visual attention mechanisms (Foulsham & Underwood, 2008; Koehler, Guo, Zhang, & Eckstein, 2014).

There have been only a few studies dedicated to model top-down visual attention in scenes. For instance, Kanan, Tong, Zhang, and Cottrell (2009) proposed a top-down saliency detector based on object appearance in a Bayesian framework. Other attempts of such models, by Jodogne and Piater (2007) or Borji, Ahmadabadi, and Araabi (2011) for example, yield decent predictive results, considering that the internal cognitive state of an observer is extremely hard to predict, and can lead to less coordination and congruency among gaze patterns of observers (Mital, Smith, Hill, & Henderson, 2011; Smith & Mital, 2013; Bruckert, Lam, Christie, & Le Meur, 2019).

On the other hand, many attention models deal with bottom-up features (see for instance Borji, Sihite, and Itti (2013) Review, Borji (2019), and Wang, Shen, Guo, Cheng, and Borji (2018)for extensive reviews). Early models focused on static images, using linear filtering to extract meaningful feature vectors, which are then used to predict a saliency map (Itti, Koch, & Niebur, 1998; Bruce & Tsotsos, 2005; Le Meur, Le Callet, Barba, & Thoreau, 2006; Harel, Koch, & Perona, 2006; Gao, Han, & Vasconcelos, 2009). Those meaningful visual features include contrast, orientation, edges, or colors, for instance. In the case of dynamic scene viewing, the early investigations underlined the importance of temporal features, such as optical flow or flicker (Guo & Zhang, 2010; Mahadevan & Vasconcelos, 2010; Mital et al., 2011; Rudoy, Goldman, Shechtman, & Zelnik-Manor, 2013). Most of the early dynamic saliency models are however extensions of existing static models, and are limited by the representation power of the chosen hand-crafted features, therefore not grasping the full amount of information delivered by ground-truth saliency.

Recently, deep learning approaches managed to significantly improve performances of attention models.

The first attempt of using automatically extracted features was conducted by Vig, Dorr, and Cox (2014), and managed to outperform most models of the state of the art at the time. Later on, several deep learning models using transfer-learning were proposed, where features learned on large-scale classification datasets were used, like DeepFix (Kruthiventi, Ayush, & Babu, 2017), SALICON (Huang, Shen, Boix, & Zhao, 2015), DeepNet (Pan, Sayrol, Giro-I-Nieto, McGuinness, & O'Connor, 2016), or Deep Gaze II (Kümmerer, Wallis, Gatys, & Bethge, 2017). More recently, the emergence of large-scale fixation datasets allowed for end-to-end approaches, in order to learn features more specific to visual saliency. These new models, like SalGan (Pan et al., 2017), SAM-VGG and SAM-Resnet (Cornia, Baraldi, Serra, & Cucchiara, 2018), or MSI-Net (Kroner, Senden, Driessens, & Goebel, 2020), exhibit great predictive behaviors, and constitute a very strong baseline for modeling human visual attention. Dynamic models followed the same path towards deep learning, with models such as DeepVS (Jiang, Xu, Liu, Qiao, & Wang, 2018), ACLNet (Wang, Shen, Guo, Cheng, & Borji, 2018), Bak, Kocak, Erdem, and Erdem (2018), or Gorji and Clark (2018). Similarly to the static case, they exhibit significantly better predictive performances than earlier approaches (see for instance Wang et al., (2019) for a more detailed review).

Visual attention and movies

Studying film perception and comprehension is still an emerging field, relying on broader studies on scene perception (Smith, Levin, & Cutting, 2012; Smith, 2013). While the effects of low-level features have been studied in great detail, in part thanks to the progress of saliency models, the effects of higher-level film characteristics are far less well understood. Loschky, Larson, Magliano, and Smith (2014) showed that the context of a sequence is particularly relevant to understand the way humans are viewing a particular shot, thus underlying the need for a better comprehension of the high-level features. Valuch and Ansorge (2015) studied the influence of colors during editorial cuts, showing that continuity editing techniques result in faster re-orientations of gaze after a cut, and that color contributes to directing attention during edits. Other studies showed strong relationships between eye-movement patterns and the number and the size of faces in a scene (Rahman, Pellerin, & Houzet, 2014; Cutting & Armstrong, 2016).

A few studies focused on gaze congruency, or attentional synchrony. Goldstein, Woods, and Peli (2007) showed that observers tend to exhibit very similar gaze patterns while watching films, and that the inter-observer agreement would be sufficient for effective attention-based applications, like magnification around the most important points of the scene. Mital, Smith, Hill, and Henderson (2011) and Smith and Mital (2013) later showed that attentional synchrony was positively correlated with low-level features, like contrast, motion, and flicker. Breathnach (2016) also studied the effect of repetitive viewing on gaze agreement, showing a diminution of the inter-observer congruency when movie clips were watched several times.

More generally, it appears that understanding human visual attention while watching movies ultimately requires a framework combining both low- and high-level features. From a cognitive point of view, Loschky, Larson, Smith, and Magliano (2020) recently proposed a perception and comprehension theory, distinguishing between the front-end processes, occurring during a single fixation, and back-end processes, occurring across multiple fixations and allowing a global understanding of the scene. From a computational and modeling point of view, no model combining low- and high-level film characteristics has yet been proposed. Alongside with Breeden and Hanrahan (2017), this paper aims to facilitate the development of such a model.

Movie eye-tracking datasets

In the field of visual attention modeling for videos, the majority of the large-scale databases used to train various models contain mostly non-cinematographic stimuli. As we show in "Visual attention modeling" section, this leads to consistent errors when saliency models are used on film sequences. Moreover, most studies involving visual attention and movies use their own collected eye-tracking data, as the experimental setups are often very specific to the characteristics studied. However, there exists a few available eye-tracking databases on movie scenes, which can be general enough for modeling purposes.

Hollywood-2 (Mathe & Sminchisescu, 2015) includes 1707 movie clips, from 69 Hollywood movies, as well as fixation data on those clips from 19 observers. Observers were split into three groups, each with a different task (three observers free-viewing, 12 observers with an action recognition task, and four observers with a context recognition task). Each group being relatively small, the common way to use this data for visual attention modeling is by merging those groups, thus introducing potential biases. The large scale of this dataset (around 20 h of video) is well fit for training deep saliency models, however few conclusions regarding gaze patterns on movies can be drawn from the data itself, since it mainly focuses on task-driven viewing mode, and that each clip is only around 15 s long.

SAVAM (Gitman, Erofeev, Vatolin, Bolshakov, & Fedorov, 2014) includes 41 high-definition videos, 28 of which are movie sequences (or use movie-like realization, like commercials for instance). Eye fixations are recorded from 50 observers, in a free viewing situation. As for Hollywood-2, each clip is quite short, only 20 s on average.

The **DIEM** project (Mital et al., 2011) is an investigation of gaze patterns on videos. The authors first released a dataset composed of eye-tracking records of 42 observers, on 26 movie sequences, for a total of 2605 s of content. In their study, the authors showed that temporal features were the most predictive of eye fixations, compared to spatial and static features. Since then, the dataset has grown, and now includes data from over 250 observers on 85 videos. These videos cover a large range of genres, including advertisements, movie trailers, music clips, or sports videos.

Study Forrest (Project, 2014) is a large-scale project centered on the movie *Forrest Gump*, and dedicated to understanding a large spectrum of the sensory impact of the movie. It includes a huge amount of data, including extensive neurological imagery, movie-related annotations, and gaze-tracking data (Hanke et al., 2016). The gaze pattern dataset includes eye-tracking data of 30 observers watching the movie, 15 of the participants being in a fMRI scanner, and the other 15 in a lab setting.

Breeden and Hanrahan (2017) proposed eye-tracking data from 21 observers, on 15 clips from 13 films, for a total of 38 min of content. Each clip is between 1 and 4 min. Alongside this data, they also provide high-level feature annotations, such as the camera movements in shots, the temporal location and types of edits, the presence or absence of faces on screen, and whether or not the characters are speaking. However, the main limitations of this dataset are the relatively low precision of the eye-tracking device used, and the duration of the total content of the base itself.

It follows that the saliency modeling community, as well as cinematographic studies, would greatly benefit from an extension of Breeden and Hanrahan's work, i.e., a relatively large-scale eye-tracking database on movies sequences, including a large diversity of editing styles, genres and epochs, alongside with high-level features annotations, related to different film-making parameters. In this work, we propose such a database, and the conclusions that we can draw from it.

Dataset overview

Films and clips selection

In Wu et al., (2017), the authors proposed a language called *Film Editing Patterns* (FEP) to annotate the production and edition style of a film sequence. Alongside this formalization of cinematographic rules, they present an open database of annotations on several film sequences, for pattern analysis purposes. In order to simplify the annotation process of our dataset, we decided to use the same clips.

We selected 20 clips, extracted from 17 different movies. The movies span different times (from 1966 to 2012) and genres, and are from different directors and editors, in order to eliminate bias coming from individual style. Table 1 gives an overview of the selected clips. The sequences were selected as they were the most memorable or famous sequences from each movie, based on scenes that users uploaded to YouTube, indicating popularity and interest to the general public.

Here we give a small description of each scene, and its most remarkable characteristics:

- **American History X**: Flashback scene, dialogue between characters seated at a table. Mostly static shots on the faces of the characters. This scene is in black and white.
- **Armageddon**: Action scene, high frequency of edits. The shot size varies a lot, from extreme closeups to large establishing shots. A lot of camera movements.
- **Benjamin Button**: Flashback scene. A lot of camera movements tracking the characters. A narrator comments the whole sequence. Some of the shots are replicated, with variations, in order to indicate alternative possibilities in the unfolding of the narrated story.
- **Big Fish**: Crowd scene, with two main characters walking through the crowd. A few shots take place in a whole different location, with only the two characters conversing.
- **The Constant Gardener**: Dramatic scene, the camera is handheld, and follows a single character throughout the sequence.
- **Departures** : Closing scene, alternation of static camera shots. Three characters are present, but no dialogue.
- **Forrest Gump**: Flashback scene, narrated by a character. Camera movements are used to reveal actors in the scene.
- **Gattaca (1)**: Dialogue scene between two characters. A lot of play on camera angles, since one of the characters is in a wheelchair, and the other one is standing.
- **Gattaca (2)**: Dialogue scene between three characters.
- **The Godfather** : Dramatic sequence, where the edits alternate back and forth from one central quiet scene to several simultaneous dramatic situations.
- **The Good, The Bad and The Ugly**: Mexican standoff scene, with three characters, where the frequency of the edits accelerate and the shot sizes go from larger to closer as the tension builds up.
- **The Hunger Games**: Dramatic scene, alternating a lot of different camera movements, angles and shot sizes.

Table 1 Overview of the selected clips. The framerate is set at 24 fps

Title	Director	Genre (IMDb)	Nb. Frames	Aspect ratio	Year
American History X	Tony Kaye	Drama	5702	1.85	1998
Armageddon	Michael Bay	Action, Adventure, Sci-Fi	4598	2.39	1998
The Curious Case of Benjamin Button	David Fincher	Drama, Fantasy, Romance	4666	2.40	2008
Big Fish	Tim Burton	Adventure, Drama, Fantasy	3166	1.37	2003
The Constant Gardener	Fernando Meirelles	Drama, Mystery, Romance	5417	1.85	2005
Departures	Yôjirô Takita	Drama, Music	10117	1.85	2008
Forrest Gump	Robert Zemekis	Drama, Romance	2689	2.39	1994
Gattaca (1)	Andrew Niccol	Drama, Sci-Fi, Thriller	3086	2.39	1997
Gattaca (2)	Andrew Niccol	Drama, Sci-Fi, Thriller	3068	2.39	1997
The Godfather	Francis Ford Coppola	Crime, Drama	1918	1.37	1972
The Good, The Bad & The Ugly	Sergio Leone	Western	9101	2.35	1966
The Hunger Games	Gary Ross	Action, Adventure, Sci-Fi	5771	2.35	2012
Invictus	Clint Eastwood	Biography, Drama, History	2203	2.39	2009
LOTR : The Fellowship of the Ring	Peter Jackson	Action, Adventure, Drama	5109	2.40	2001
Pulp Fiction	Quentin Tarantino	Crime, Drama	3211	2.39	1994
The Shawshank Redemption (1)	Frank Darabont	Drama	5374	1.85	1994
The Shawshank Redemption (2)	Frank Darabont	Drama	4821	1.85	1994
The Shining	Stanley Kubrick	Drama, Horror	4781	1.33	1980
The Help (1)	Tate Taylor	Drama	4151	1.85	2011
The Help (2)	Tate Taylor	Drama	5244	1.85	2011

A crowd is present, but several tricks (colored clothing, focus) are used to distinguish the main characters.

- **Invictus**: Contemplative scene, starting in a cell and ending in outdoors. Characters appear and disappear as ghosts. A narrator reads a poem.
- **Lord of The Rings**: Dialogue scene between two characters, alternating with flashbacks, mostly of action scenes. Different camera movements, angles and shot sizes.
- **Pulp Fiction**: Dialogue scene between two characters seated face to face. The exact same camera angle is used throughout the scene.
- **Shawshank Redemption (1)**: Dialogue between several characters, various camera movements, angles and shot sizes.
- **Shawshank Redemption (2)**: Flashback scene, following a single character, explaining a prison escape. A narrator comments a part of the sequence. Various camera movements, angles and shot sizes.
- **The Shining**: Dialogue scene between two characters. Very low frequency of edits, and abundant presence of the color red in the scene.
- **The Help (1)**: Flashback scene, dialogue between two characters.
- **The Help (2)**: Flashback scene, in between a dialogue scene between two characters. A lot of faces and colored clothing.

The length of the clips varies from 1 min 30 s to 7 min. This length is intentionally higher than in the other datasets presented in "Movie eye-tracking datasets" section, in order to allow the observer to feel immersed in the sequence, and thus exhibiting more natural gaze patterns. In total, the dataset contains roughly 1 h of content. Table 2 show the lengths of the average shots for each sequence. The high diversity in terms of shot lengths underlines the diversity in terms of editing styles.

High-level features annotations

Films typically contain many high-level features aiming to attract or to divert the observers' visual attention (Smith et al., 2012). These features can be of different sorts: the presence of faces or text, the framing properties, the scene composition, or the camera motion and angle, for instance. The timing of the shots, the selection of the shots from rushes by the editor, and the narrative it creates are also high-level features specific to films. Audio cues, like the presence of music or dialogue can also be considered as a form of high-level movie features, and have been increasingly studied as a way to improve visual attention models (Tavakoli, Borji, Rahtu, & Kannala, 2019). However, all of those features can prove very challenging to extract automatically, which can explain why saliency models seem to only learn non-temporal image characteristics, at the scale of

Table 2 Lengths of the sequences, and of the longest, shortest, and average shots of each sequence

Sequence	Sequence length (s)	Longest shot (s)	Shortest shot (s)	Average shot (s)
Armageddon	191.8	12.1	0.0	1.6
The Hunger Games	240.8	16.7	0.6	2.4
The Curious Case of Benjamin Button	194.7	11.8	0.3	2.5
The Godfather	80.0	6.8	0.5	2.7
Big Fish	132.1	7.6	0.7	2.8
The Constant Gardener	226.0	13.8	0.4	3.5
LOTR : The Fellowship of the Ring	213.1	8.4	0.5	3.6
The Good, The Bad & The Ugly	379,7	36.5	0.2	3.8
The Help (2)	218.8	14.0	1.0	4.0
Invictus	91.9	8.6	1.8	4.2
American History X	237.9	14.7	1.0	4.2
Pulp Fiction	134.0	12.2	1.4	4.6
The Shawshank Redemption (1)	224.2	19.2	0.8	4.7
The Help (1)	173.2	17.7	1.8	6.0
Gattaca (1)	128.7	23.7	0.2	6.1
Departures	422.0	21.8	1.8	6.6
Forrest Gump	112.2	16.6	1.8	6.7
Gattaca (2)	128.0	17.1	1.8	6.7
The Shawshank Redemption (2)	201.1	18.0	1.8	7.7
The Shining	199.5	107.1	8.8	39.9

the frame, like contrast- or texture-like information. We then used the database of Film Editing Patterns described in Wu, Palù, Ranon, and Christie (2018) to select a hand-crafted set of high-level annotations that can help in the study of visual attention and gaze patterns on films. More particularly, such annotations enable us to conduct quantitative analysis on the influence of these cinematographic features over visual attention.

Camera motion

Camera motion is an efficient tool used on set by the filmmaker to direct attention. For each shot of the database, we differentiate several possible camera motions:

- *Static*: The camera is mounted on a stand and does not move.
- *Track*: The camera moves in order to keep an object or a character in the same region of the image
- *Zoom*: The camera operator is zooming in or out
- *Pan*: The camera rotates on the horizontal plan
- *Tilt*: The camera rotates on the vertical plan
- *Dolly*: The camera is being moved using a dolly
- *Crane*: Complex camera motion, where both the camera base and the mount are in motion
- *Handheld*: The camera operator holds the camera by hand, creating a jerky motion
- *Rack focus*: The focus of the lens shifts from one point of the scene to another

Those features are binary for each shot, and a single shot can include different camera motions.

Camera angle

In order to convey the emotional states of the characters, or power relationships, filmmakers often use camera angles (Thompson & Bowen, 2009). For instance, a rolled plan will often indicate that the characters are lost, or in an unstable state of mind, while filming actors with a low angle will give

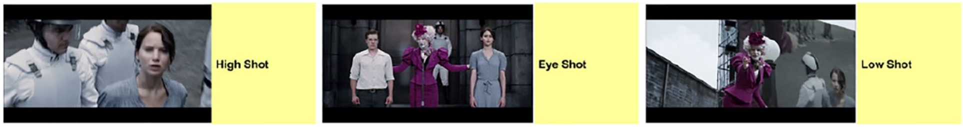


Fig. 1 Examples of different camera angles, extracted from *The Hunger Games* (from Wu et al., (2017))

them an impression of power over the other characters, as they tower over the scene. We relied on six different degrees of camera angles (Wu et al., 2017) (Fig. 1):

- *Eye*: The camera is at the same level as the eyes of the actors
- *Low*: The camera is lower than the eyes of the actors, pointing up
- *High*: The camera is higher than the eyes of the actors, pointing down
- *Worm*: The camera is on the ground, or very low, pointing up with a sharp angle
- *Bird*: The camera is very high, pointing down with a sharp angle
- *Top*: The camera is at the vertical of the actors, pointing straight down

Shot size

The size of a shot represents how close to the camera, for a given lens, the main characters or objects are, and thus how much of their body area is displayed on the screen. Shot size is a way for filmmakers to convey meaning about the importance of a character, for instance, or the tension in a scene. Very large shots can also be used to establish the environment in which the characters will progress. To annotate the shot sizes, we use the 9-size scale defined by Thompson and Bowen (2009). Figure 2 shows the differences between those shot sizes.

Faces

As explained by Cerf, Harel, Einhaeuser, and Koch (2008), the presence of faces in images is very important high-level information to take into account when studying visual attention. We then provide bounding boxes delimiting each face on each frame. Recent state-of-the-art face-detection models show that deep learning models extract this information very well. It is then probable that deep visual attention models are also great at extracting faces features, making this hand-crafted feature redundant. However, we include it, as it permits an easier analysis of the editing style: for instance, continuity edits will often display faces on the same area of the image, while shot/reverse shots often display faces on opposite sides of the image.

Eye-tracking data collection

Participants and experimental conduct

We have collected eye-tracking data from 24 volunteers (11 female and 13 male), aged 19 to 56 (average 28.8). Participants were split into two groups, each group watching half of the videos. A few observers were part of both groups, and viewed the whole dataset. In total, we acquired exploitable eye fixation data for 14 participants for each video. Details of the sequences viewed by each group can be found in https://github.com/abruckert/eye_tracking_filmmaking.

When recruiting volunteers, we specifically required that they could understand non-subtitled English movie clips, as sequences were extracted from the English version of the movies. We also welcomed them and described the experiment in English, ensuring that their comprehension would be sufficient. Viewers were asked to fill out an explicit consent form, and to perform a pre-test form. The objective of the pre-test form was to detect any kind of visual impairment that could interfere with the conduct of the experiment (color blindness or strabismus, for instance). Participants were informed that they could end the experiment at any moment.

During a session, subjects viewed the ten movie sequences assigned to their group in random order. Sound was delivered by a headset, and the volume was set before

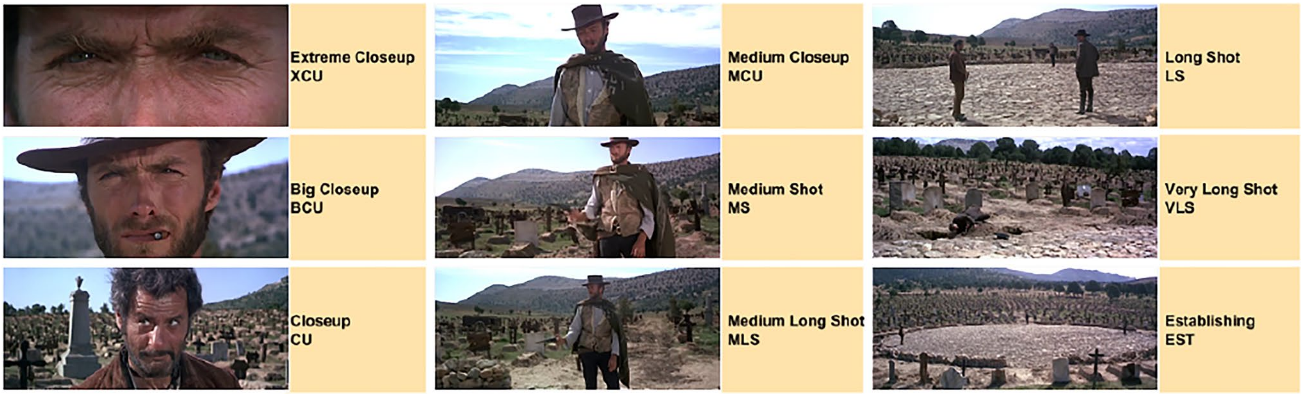


Fig. 2 The nine framing sizes, all appearing in *The Good, The Bad & The Ugly* (from Wu et al., (2017))

the first sequence. They could also adjust the volume at will during the experiment. After each sequence, a 15-s dark-gray screen was displayed. After a series of five clips (around 15 to 20 min of video), participants were asked to make a break, as long as they needed, and then fill out a form, recording whether or not they could recall the scenes they saw, whether or not they had seen the movies previously, or if they recognized any actors in the scenes. After the second series of five clips, at the end of the experiment they were asked to fill out the same form. The total duration of the experiment for a participant was between 50 min and 1 h.

Recording environment and calibration

Eye movements were recorded using a Tobii X3-120 eye tracker, sampling at 120 Hz. The device was placed at the bottom of a 24,1” screen with a display resolution of 1920 × 1200 pixels. All stimuli had the same resolution (96 dpi), and were displayed respecting the original aspect ratio, using letterboxing. The participants were asked to sit at a distance of 65 cm from the screen. They were asked to sit as comfortably as possible in order to minimize head movements. In order to replicate natural viewing conditions, we did not use chin rests. Stereo sound, with a sampling frequency of 44100 Hz, was delivered to the participant, using a headset. Calibration was performed using the 9-points Tobii calibration process. In the case of errors of more than one degree, the participant was asked to reposition and recalibrate. After the break, before viewing the five last clips, participants were asked to validate the previous calibration, and to recalibrate if necessary.

After recording the data for all participants, we used the following cleaning procedure. First, we ensured that every participant had a gaze sampling rate of more than 90% (i.e., more than 90% of the sampled points were considered as valid). We then kept only points that were flagged as fixations, eliminating tracking errors due to blinks or other factors, as well as points recorded during saccades. This choice was motivated by the relatively low frequency rate of the eye-tracker, making the analysis of saccadic data impossible. Then, we discarded all points that fell in the letterboxing or outside the screen. Finally, we used the position of the remaining raw points to construct binary fixation maps: for each frame, we create an image the same size of the frame where we give the value 1 to each pixel where a fixation point was flagged during the time the frame was on screen (i.e., 1/24th of a second), and 0 to each pixel where no fixation occurred (Fig. 3).

Open-source data package

Overall, here is a summary of what can be found in our dataset, available at https://github.com/abruckert/eye_tracking_filmmaking.

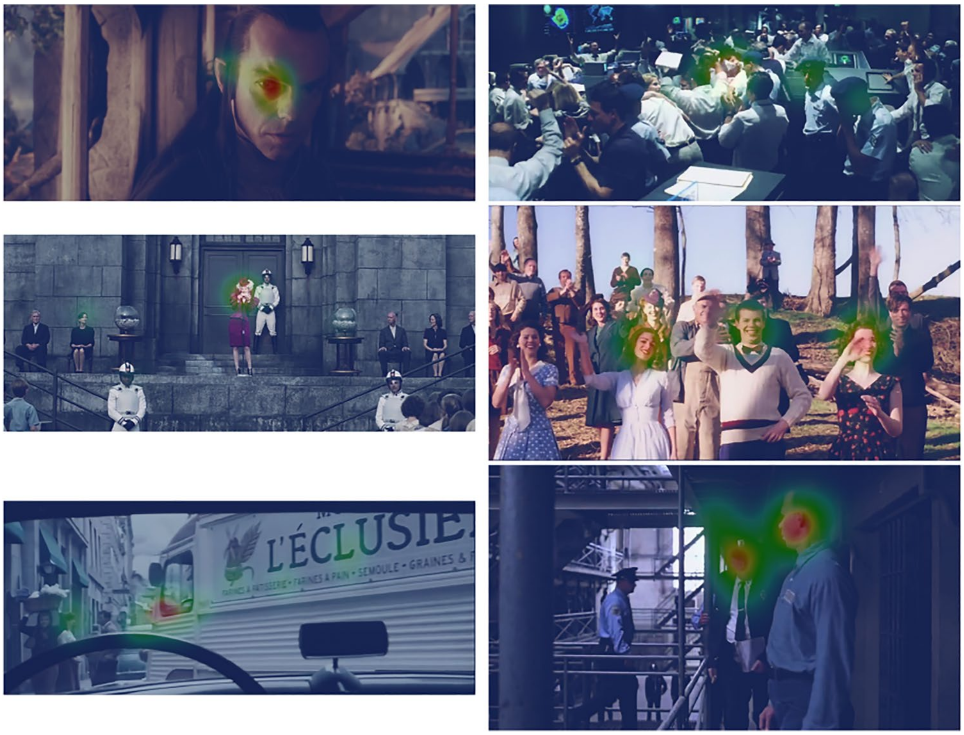


Fig. 3 Examples of saliency heatmaps created from the collected fixation points

- Firsts and lasts frames of each stimulus used, alongside with timestamps and release details from where the clips can be sourced.
- Eye-fixation maps for each frame of each stimulus.
- The derived saliency maps for each frame of each stimulus.
- The extracted fixation sequences of each individual observer on each clip.
- Hand-crafted feature annotations on the editing characteristics of each clip, described in “High-level features annotations” section
- Various details about the movie from which the clip is extracted (director, genres, aspect ratio, date)

Exploring the effects of film making patterns on gaze

In this section, we explore several characteristics throughout our database, and analyze underlying relationships between editing patterns and eye-fixation patterns. In the following, we will often refer to *fixation maps* and *saliency maps*. For each frame, the fixation map is the binary matrix where each pixel value is 1 if a fixation occurred at the pixel location during the frame, and 0 if not, as described previously. Saliency maps are obtained by convolving the fixation maps with a 2-D Gaussian kernel, where the variance is set to 1 degree of visual angle (in our case, 1 degree of visual angle equals roughly 45 pixels), in order to approximate the size of the fovea.

Editing-induced visual biases

Studying the average of the saliency maps usually reveals strong attentional biases. For instance, on static images, Tatler (2007) showed that humans tend to look at the center of the frame. That center bias is also commonly used as a lower baseline for saliency models. In order to avoid recording this bias too much, we did not take into account for our analysis the first ten frames of each clip, as people tend to look in the middle of the screen before each stimulus. This center bias is also strong on video stimuli: for instance, Fig. 4a, b shows the average saliency map on our dataset and on the DHF1K dataset (Wang et al., 2019), respectively. However, the latter is composed of YouTube videos, with a great diversity in the content, and no cinematographic scenes, which might cause a different viewing bias. Figure 4a shows a peak density slightly above the center of the frame, which would indicate that filmmakers use a different composition rule. Figure 4c shows a centered Gaussian map, often used as a baseline for centered bias. Correlation between the average saliency map on our dataset and this centered Gaussian is 0.81, whereas the correlation between the average map on DHF1K and the centered Gaussian is 0.84, which highlights this position discrepancy between the two average saliency maps. This is consistent with the findings of Breeden and Hanrahan (2017), and is most likely due to the rule of thirds (Brown, 2016) stating that in cinematography, important elements of the scene should be placed on thirds lines, i.e., lines dividing the frame in thirds horizontally and vertically.

We also observe disparities in this bias depending on the size of the shot: the wider the shot, the more diffuse that bias is, indicating that directors tend to use a bigger part of the screen area when shooting long shots, while using mostly the center of the frames for important elements during closeups and medium shots (Fig. 5a, b, c). We also observe a leftward (resp. rightward) bias during pans and dolly shots, where the camera moves towards the left (resp. right), as exposed in Fig. 5d, e. This confirms that camera movements are an important tool for filmmakers to guide the attention of the spectators.

Inter-observer visual congruency

Inter-observer congruency (IOC) is a measure of the dispersion of gaze patterns between several observers watching the same stimulus. In other words, it measures how well gaze patterns from a subset of the observers is predictive of the whole set of observers. Thus, it has been used in saliency modeling as an upper baseline. This characteristic is very similar to attentional synchrony (Smith & Henderson,

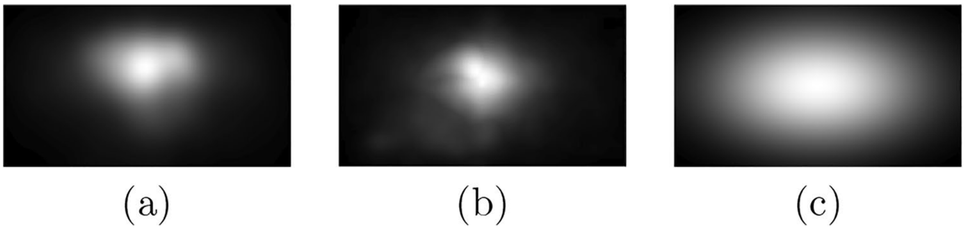


Fig. 4 Average saliency map of our dataset (**a**) compared to DHF1K (Wang et al., 2019) dataset (**b**) and to a centered gaussian map (**c**). Both average maps exclude the first ten frames of each clip

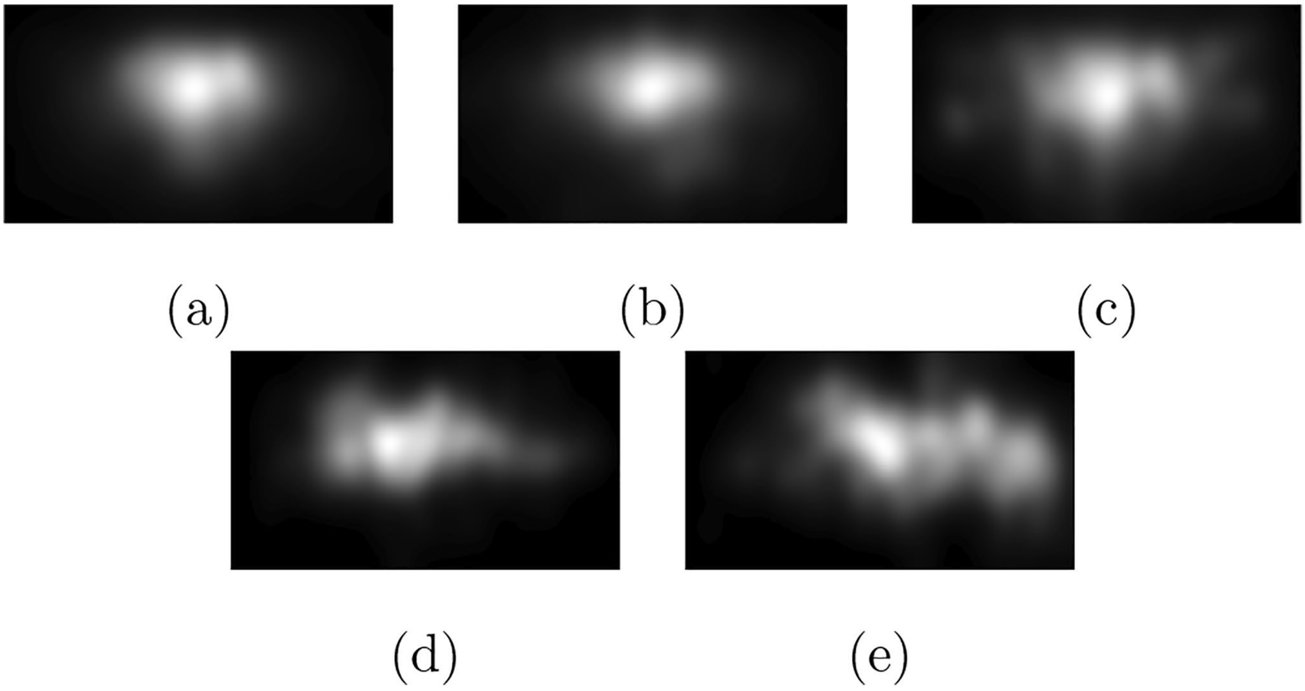


Fig. 5 Average saliency maps for closeup shots (XCU-BCU-CU) (**a**), medium shots (MCU-MS-MLS) (**b**) and long shots (LS-VLS-EST) (**b**). Subfigure (**d**) is the average saliency map during pans and dolly shots moving to the left, and (**e**) is the average saliency map during pans and dolly shots moving to the right

2008), and many methods have been proposed to measure it. For instance, Goldstein et al., (2007) measure how well gaze points are fitted by a single ellipsoid, while Mital et al., (2011) use Gaussian mixture models, associating low cluster covariance to high attentional synchrony.

In their work, Breeden and Hanrahan (2017) use the area of the convex hull of the fixation points, for each frame of their dataset. This allows to take into account all the fixation points, and requires no prior hypothesis about the shape of the regions of interest. However, as they mention, this approach can only be viewed as an upper bound on IOC, as it is very sensitive to outliers. Using it on each frame also does not take into account the temporal aspect of movie viewing: if several observers watch the same two or three points of interest, chances are, if the points are spatially distant from one another, that the convex hull area will be high, even though all the observers exhibited similar gaze patterns in a different order, in terms of fixation locations.

In order to remedy this issue, we used a leave-one-out approach, over a temporal sliding window. Assuming that there is N observers, we gather the locations of all the fixation points of $(N-1)$ observers during a window of n frames, as well as the locations of the fixation points of the left out observer. We then can build a saliency map using the fixation points of the $N-1$ observers, by convolving the fixation map with a 2-D Gaussian kernel, and use any saliency metric to compare it to the fixation map (or saliency map) of the left out observer. The process is then iterated and averaged over all observers. To compare the saliency map of the $N-1$ observers to the left-out one, we used the Normalized Scanpath Saliency metric (NSS); more details about can be found in Le Meur and Baccino (2013). The choice of the NSS metric makes the IOC scores independent to the number of gaze clusters: indeed, the NSS metrics is used after normalizing the whole saliency map such that it has a zero mean and unit standard deviation.

A high value of this score will mean that people tend to look in the same region, and a low value will indicate a higher dispersion in the fixations. The main drawback of this way of computing IOC, especially for large-scale datasets, is its computational cost, as the process is iterated over every observers, and every n-frame window of the dataset.

The size n of the sliding window can be adjusted, depending on the number of observers, and the studied characteristics. In this work, we chose two window sizes: five frames and 20 frames (roughly 200 and 800 ms). Five frames correspond roughly to the average fixation time, and 20 frames allows for a wider point of view, with less noise. While a shorter window allows for more noise, especially with a relatively short number of observers, it can also underline short-timed patterns. For instance, using the shorter window size, we noticed on every stimulus a significant peak of inter-observer congruency during the five frames consecutive to a cut (see Fig. 7).

Longer window sizes would be better suited for analyzing more general trends, at the scale of a shot, for instance. In the following, we will use the values computed using a 20-frame window, since our editing annotations are at the level of the shot.

As we suspected, the IOC values are relatively high (average of 4.1 over the whole database), especially compared to IOC values in the static case; see for instance Bruckert, Lam, Christie, and Le Meur (2019) for IOC distributions on static datasets. This corroborates the findings of Goldstein et al., (2007) and Breeden and Hanrahan (2017), that viewers only attend to a small portion of the screen area. We also observe a disparity in IOC scores between the movies: the scene with the highest score on average is the clip from *The Shining* (5.76), and the lowest score on average is the clip from *Armageddon* (3.41). This would tend to indicate that inter-observer congruency reflects certain features in terms of editing style; for instance, Fig. 6 shows a correlation (0.35) between the average IOC score of a sequence and the average length of the shots of this sequence. However, due to the low number of samples, this correlation is not significant ($p = 0.19$).

Inter-observer congruency and cinematographic features

We then considered the effects of directors' choices and editing characteristics on inter-observer congruency.

Camera movements and IOC

In order to evaluate the impact of camera motion on inter-observer congruency, we compared the IOC values on shots that contain at least one camera movement with fully static shots, using one-way ANOVA. On average, static shots show slightly higher IOC values ($M = 4.331$, $SD = 1.60$) than shots exhibiting camera motion ($M = 4.025$, $SD = 2.310$) ($p << 10^{-5}$), but it is worth noting that standard deviation is significantly higher in the shots where the camera moves, indicating that camera motion plays an important role in increasing or decreasing IOC. This is consistent with the findings of Mital et al., (2011), showing that motion-related features (not specifically camera motion) are a good predictor of eye fixations clustering. However, this high standard deviation comes from the fact that some strong camera motions, where the focus is transferred from an object to another for instance (like in Fig. 9), can create very high IOC scores. Indeed, the camera motion is leading enough

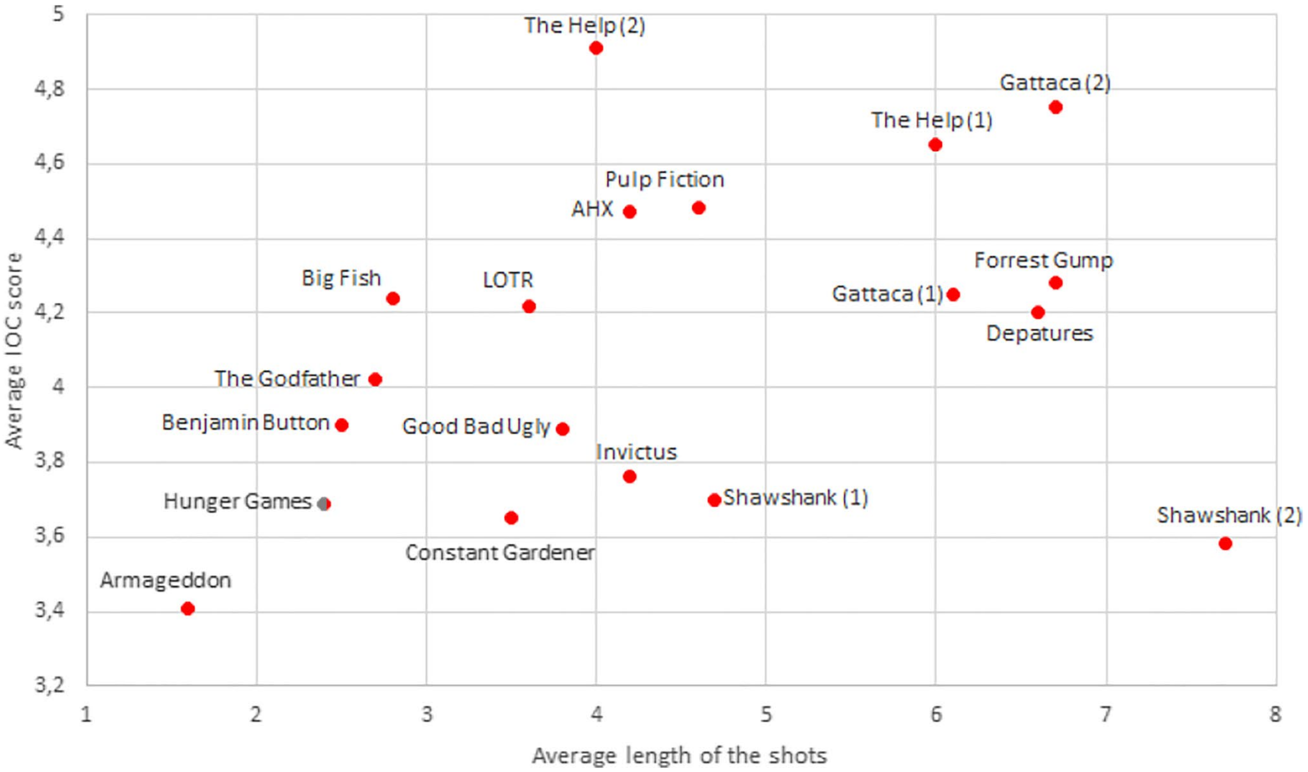


Fig. 6 IOC scores depending on the average length of the shots for each sequence (excluding *The Shining*, as it is an outlier in terms of length of the shots)

so that viewers anticipate the appearance of a very salient area, and thus cluster their gazes in the border of the frame in the direction of the camera motion.

Considering this, we performed post hoc pairwise *t* tests (using Bonferroni correction for multiple tests, i.e., multiplying the *p* values by the number of comparisons) between the annotation groups for camera movement, showing significant differences ($p << 10^{-5}$) between most shot characteristics, except between static and dolly shots ($p = 0.026$). As expected, the highest average IOC values are in zoom shots and rack focuses, which are camera features specifically designed to direct visual attention; these values are shown in Fig. 8a.

Camera angles (Fig. 8b) show no significant differences between the choice of camera angle and inter-observer congruency. At first glance, it may seem that extreme camera angles (bird shots, worm shots, and top shots) are associated with higher IOC values, but this might just be an artifact due to the relatively low number of such shots in the dataset.

Shot size and IOC

Similarly, we looked at the average scores depending on the size of the shots (Fig. 8c). Extreme closeup shots are associated with the highest IOC scores ($M = 4.863$, $SD = 1.758$), while interestingly, big closeups have the lowest IOC averages ($M = 3.967$, $SD = 1.840$). This difference is confirmed by a *t* test ($p << 10^{-5}$), and might be explained by the way these shots are used in the overall scene: on average, extreme closeup shots are very short ($M = 34.111$, $SD = 26.434$ frames) compared to big closeups ($M = 78.683$, $SD = 67.476$ frames), thus leaving more time for exploration. A perfect example of this is theMexican standoff scene from *The Good, the Bad and the Ugly* (Sergio Leone, 1966): the shots come closer and closer to the characters as the tension builds up; when it reaches the big closeup shot size, a little bit of time is given to the spectator to read the characters faces. After that, a series of very short shots show extreme closeups on the eyes and guns of the characters, forcing eye fixations on the salient elements (i.e., the eyes and the guns).

Medium shot categories (MCU, MS, MLS, and LS) show little to no significant differences of IOC. This might be due to categories sometimes not very well defined, as it can be hard distinguishing between a medium shot and a medium-long shot, for instance.

Smith (2013) showed however that IOC have a U-shaped relationship with the size of the shots, with the extreme closeups and very long shots exhibiting a low IOC, and medium shots a higher IOC. We think the discrepancies in results that we obtained are explained by several effects, like the relatively low number of extreme closeup or establishing shots in our dataset, or the difficulty of defining shot sizes categories. The explanation given for this U-shaped relationship is the fact that on extreme closeups, focal objects are so close to the camera that it fragments into several focal objects, like the eyes, nose, or mouth of an actor. In this context, IOC scores could be useful to objectively define the difference between close shot size categories.

Cuts and edits

As mentioned just before, the rhythm of the cuts and edits play an important role in directing attentional synchrony. A well-known effect (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Mital et al., 2011; Smith & Mital, 2013)is the sudden augmentation in inter-observer agreement immediately following a cut. This tendency is observable in our dataset, when taking into account the binning effect linked to the size of the temporal window used to compute IOC scores. Figure 7 shows this effect on a clip from *Armageddon* (Michael Bay, 1998). We observe a significant difference ($p << 10^{-5}$) between the IOC scores of the frames within the first 500 ms immediately following a cut ($M = 5.712$, $SD = 1.882$) and the rest of the frames ($M = 3.510$, $SD = 1.974$).

Visual attention modeling

In this section, we evaluate several visual saliency models on our database, and highlight certain limitations of current dynamic saliency models. We also discuss how editing patterns can explain some of the failure cases of the models.

Saliency models

In order to evaluate the predictive power of visual saliency on movie clips, we selected nine models from the literature, which are representative of the field. We include five dynamic saliency models, and four static models.

Static models

- We first evaluate the performances of the seminal saliency model of Itti, Koch, and Niebur (1998). In this approach, each individual frame is passed through a linear filter, and low-level features are extracted, namely color features, intensity features and orientation features. The resulting feature maps are then combined into three conspicuity maps, which are ultimately merged together into the final saliency representation.
- Salgan (Pan et al., 2017) is a deep learning model relying on generative adversarial networks. The frames are passed through a deep convolutional feature extractor (VGG16), and the resulting features are passed into a succession of convolutions and upsamplings, resulting in a saliency prediction. Then, a discriminative convolution

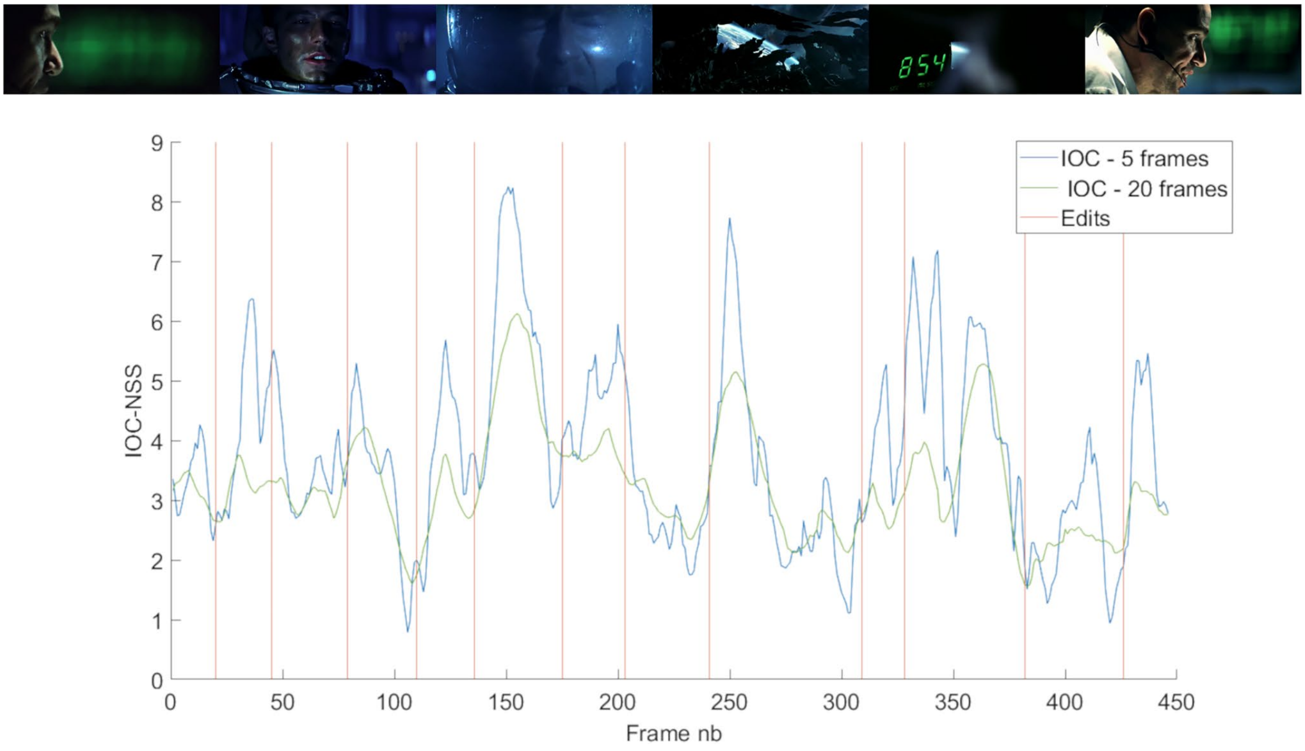
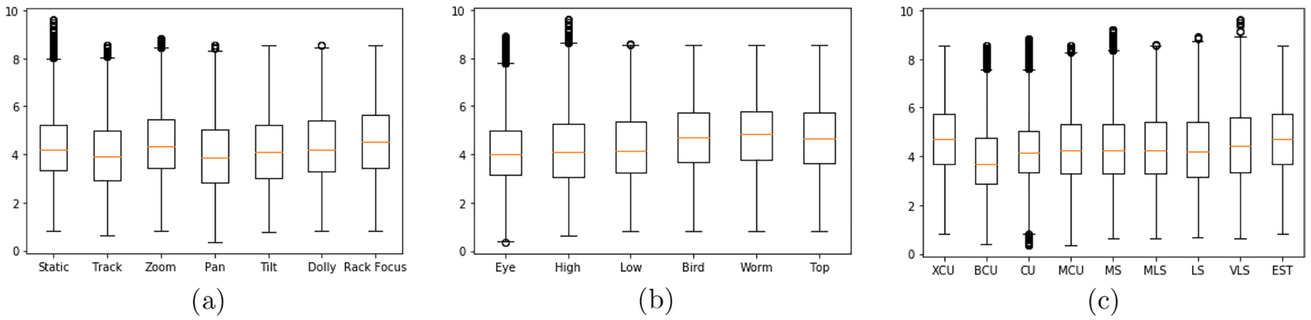


Fig. 7 Example of the evolution of inter-observer congruency over a sequence of *Armageddon*. *Blue* is the IOC value computed with a five-frame time window ($n = 5$), *green* is using a 20-frame time window ($n = 20$). *Red lines* show the locations of the edits; notice the characteristic peak in IOC values after each cut (for the five-frames window)



(a) (b) (c)

Fig. 8 IOC scores depending on camera movement features (**a**), camera angles (**b**), and shot size (**c**)

network is trained to differentiate between the saliency predictions and the ground-truth maps. Both networks are then trained alternatively, the generative network improving the prediction, and the discriminator refining the cost function.

- DeepGaze II (Kümmerer et al., 2017) is one of the leading models in terms of performances on static stimuli. This model is formulated as a probabilistic model, meaning that instead of predicting the saliency topographic map, it outputs a fixation probability density, which can then be converted into a saliency map. It consists of a deep convolutional feature extractor (VGG16) where the parameters are frozen during the training, and a trained readout network, with four 1 × 1 convolutions. The resulting map is then blurred and combined with a center bias.
- MSINet (Kroner et al., 2020) once again consists in an encoder–decoder architecture, with a deep VGG encoder. However, the feature maps are extracted at different depths of the encoder, and passed through Atrous Spatial Pyramid Pooling. The resulting features are concatenated and passed through a convolutional decoder with upsampling at the end of each convolution block. This model has proven to be a very efficient tradeoff between

the number of parameters to optimize and the overall performances.

Dynamic models

- PQFT (Guo & Zhang, 2010) is a non-deep dynamic saliency model relying on phase spectrums of the quaternion Fourier transform. Each frame is transformed into a quaternion representation, with two color dimensions, an intensity dimension, and a motion dimension. Then, by applying a quaternion Fourier transform to this representation, the phase spectrum in the frequency domain is extracted, and used to reconstruct the quaternion. The saliency map is finally created by convolving this reconstruction with a Gaussian filter.
- The two-stream approach (Bak, Kocak, Erdem, & Erdem, 2018) is a deep learning model based on the separation of image and motion. The frames are first passed individually through a deep encoder–decoder model, similarly to the static encoder–decoder models, outputting a spatial saliency representation. In parallel, the optical flow is also passed through a similar encoder–decoder, thus exploiting only temporal information, and outputting a temporal saliency representation. The two encoders are trained separately, and then fused altogether by a convolution fusion model.
- DeepVS (Jiang et al., 2018) is also composed of two subnets, one treating motion and the other objectness. The objectness network is based on the YOLO model (Redmon & Farhadi, 2018), designed for object detection, while the motion networks extract temporal features at different depths from pairs of neighboring frames, using the FlowNet architecture (Dosovitskiy et al., 2015), a deep convolution network designed to infer optical flow. All of those features are later combined together using a LSTM network, predicting a sequence of saliency maps.
- ACLNet (Wang et al., 2019) also relies on the succession of a convolution encoder with a LSTM unit. A static convolution module is trained on images to extract intra-frame static features. These features are then passed into a LSTM, learning the sequential saliency representations.
- Finally, Zhang and Chen (2019) also proposed a two-stream approach, similar to the DeepVS architecture, but this time using 3D convolution networks as feature encoder. The use of 3D convolution networks, while computationally expensive, allows for a better representation of temporal characteristics.

Performance results

In Table 3, we show the performances of state-of-the-art static and dynamic saliency models. In order to evaluate the models, we used the following six classic saliency metrics, described in Le Meur and Baccino (2013):

- Pearson's correlation coefficient ($CC \in [-1,1]$) evaluates the degree of linear correlation between the predicted saliency map and the ground truth map.
- SIM ($SIM \in [0,1]$) evaluates the similarity between two saliency maps through the intersection between their histograms.
- AUC (AUC-J, AUC-B $\in [0,1]$) is the area under the receiver operator curve (ROC). Differences between AUC-J and AUC-B rely on the way true and false positive are computed (see Le Meur and Baccino (2013) for more details).

Table 3 Scores of several saliency models on the database

	Model	CC ↑	SIM ↑	AUC-J ↑	AUC-B ↑	NSS ↑	KLD ↓
Baselines	Leave-n-out standard*	0.974	0.932	0.946	0.929	4.523	0.033
	Center prior*	0.398	0.302	0.859	0.771	1.762	2.490
Dynamic models	PQFT* (Guo and Zhang, 2010)	0.146	0.189	0.702	0.621	0.783	2.948
	Two-stream (Bak et al., 2018)	0.404	0.329	0.873	0.830	1.738	1.410
	DeepVS (Jiang et al., 2018)	0.457	0.361	0.880	0.829	2.270	1.245
	ACLNet (Wang et al., 2019)	0.544	0.429	0.892	0.858	2.54	1.387
	ACLNet (retrained)‡	0.550	0.423	0.890	0.858	2.592	1.408
	(Zhang & Chen, 2019)	**0.608**	**0.454**	**0.903**	0.881	2.847	**1.154**
Static models	Itti* (Itti et al., 1998)	0.208	0.195	0.756	0.640	1.005	2.573
	SalGAN (Pan et al., 2017)	0.533	0.390	0.897	0.781	2.622	1.372
	DeepGaze II (Kümmerer et al., 2017)	0.584	0.362	0.846	0.774	**3.188**	2.307
	MSINet (Kroner et al., 2020)	0.597	0.417	0.901	**0.893**	2.893	1.226

Non-deep models are marked with *. Best performances are in bold. ‡Note that the testing dataset for the retrained ACLNet model is not exactly the same as the other models, as it is a subset of half of our dataset. CC: Person's correlation coefficient; NSS: normalized scanpath saliency; SIM: similarity; AUC-J: area under curve (Judd); AUC-B: area under curve (Borji); KLD: Kullback–Leibler divergence

- Normalized scanpath saliency ($NSS \in [0, +\infty[$) is computed between the predicted saliency map and the ground truth fixation map by measuring the saliency values at the locations of the fixations.
- Kullback–Lieber divergence ($KLD \in [0, +\infty[$) between the two probability distributions represented by the saliency maps.

In general, those results are quite low, compared to performances on non-cinematic video datasets (see for instance Wang et al., (2019)).

It is also worth noting the important gap between the performances of the model and the leave-one-out standard results. The leave-n-out standard is obtained by comparing the fixations of n individuals to the aggregated fixations of all other observers, using various metrics, and cross-validated over all the groups of n individuals. The IOC score can be seen as this leave-n-out standard, with $n = 10$. However, on a single frame, which is necessary to compare it to the performances of saliency models, we cannot rely on a single observer, as it is likely that there will be too few fixation points during the time where the frame is displayed, thus creating a lot of noise. We then use $n = 10$ to compute the leave-n-out standard. Overall, this creates a strong upper bound, as the gaze distribution of another group of humans is ultimately more predictive than any model as of today.

This would indicate, in the case of deep-learning models, that either the training sets do not contain enough of videos with features specific to cinematic stimuli or the deep neural networks cannot grasp the information from some of those features. Even though the best performing model is a dynamic one (Zhang & Chen, 2019), we observe that static models (DeepGaze II and MSINet) performances are quite close to those of dynamic models. This might support the latter hypothesis that dynamic models fail to extract important temporal features.

Recent work from Tangemann, Kümmerer, Wallis, and Bethge (2020) on the failure cases of saliency models in the context of dynamic stimuli also highlight this point, listing cases like appearing objects, movements, or interactions between objects as some of the temporal causes of failure. Figure 9 shows an example from our database of such a failure case. It should be noted that all the deep learning models are trained on non-cinematic databases, with the exception of ACLNet, which include the Hollywood 2 dataset in its training base. However, this base is not well-fit to learn meaningful cinematographic features, as explained in the "Movie eye-tracking datasets" section.

In order to confirm this hypothesis, we retrained the ACLNet model using the same training procedure described in Wang et al., (2018). For the static batches, we used the same dataset (SALICON Huang, Shen, Boix, and Zhao (2015)), and for the dynamic batches we created a set composed of half of our videos, randomly selected, leaving the other videos out for testing (roughly 490,000 frames for training, and 450,000 frames for test). We only obtained marginally better results on some of the metrics (0.550 instead of 0.544 on the correlation coefficient metric, 2.592 instead of 2.54 on the NSS metric), and did not outperform the original model settings on the other. All of this would tend to indicate that some features, specific to cinematographic images, could not be extracted by the model.

Edition annotation and model performances

We also studied how the two best dynamic models, Zhang et al., (2020) and ACLNet (Wang et al., 2019), performed on our database, depending on shot, camera motion, and camera angle characteristics. Table 4 shows the average results of the models depending on the annotation characteristics. Similarly to "Inter-observer visual congruency" section, we performed one-way ANOVAs to ensure that results within each table would yield significant differences. In all cases, p values were under 10^{-5}.

As shown in Table 4(a), it appears that saliency models perform relatively well on static scenes, or when the camera movement tracks an actor, or an object on screen. Performances are also quite good on shots including rack focuses, which was expected, as this is a very strong tool for the filmmaker to use to direct attention, and deep feature extractors distinguish very well blurry background from clear objects. However, when a more complex camera motion appears, like pans or tilts, models seem to fail more often; this might indicate that saliency models are unable to anticipate that an object is likely to appear in the direction of the motion, which humans usually do.

With Table 4(b), we observe that camera angles show little variations in the performances of the models. However, it seems that scenes with high amplitude angles (Bird or Worm) are easier for a model to predict. This is probably due to the fact that those camera angles are often used when filming characters and faces, in order to convey a dominant or a submissive feeling from the characters (Thompson and Bowen, 2009); since deep learning models are very efficient at recognizing faces, and faces tend to attract gaze, saliency models naturally perform better on those shots.

Finally, looking at Table 4(c), saliency models seem to exhibit great performances on closeup scenes. It is surprising to see that closeup scenes exhibit the best performances saliency-wise, while their IOC scores are amongst the lowest. This could be explained by this profusion of faces in closeup shots, which makes them particularly good candidates for deep saliency models, compared to longer shots, which allows the director to add more objects or actors on screen. Indeed, as shown by Tangemann et al., (2020), interactions between objects are often a failure case for deep

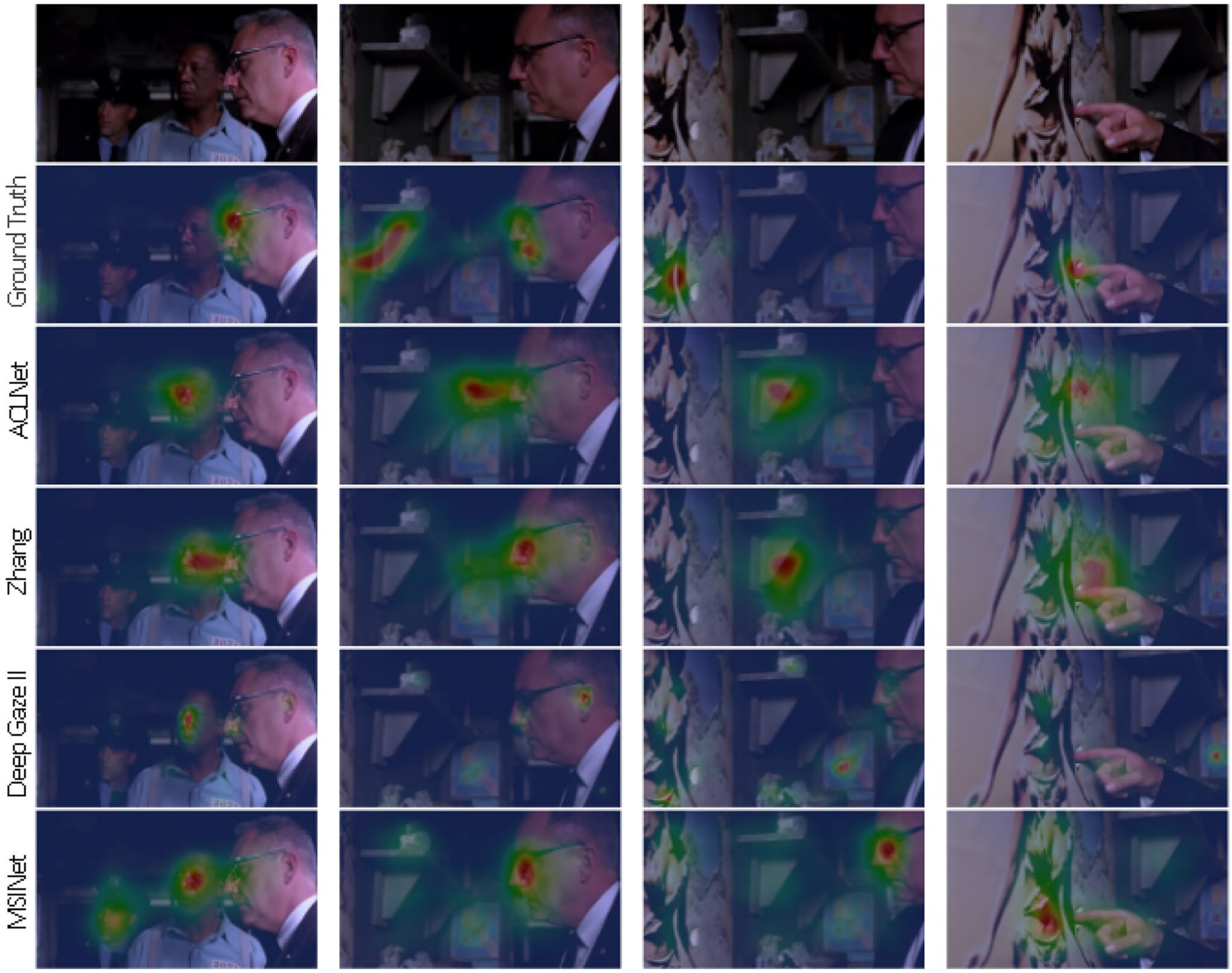


Fig. 9 An example of failure case in *Shawshank Redemption*. Here, the camera pans from the face of the prison director to the poster on the wall. While observers quickly shift their attention towards the poster, as suggested by the camera movement, even though it is not yet on screen, models tend to predict areas of interest on the faces

saliency models. Such situations would then generate high IOC scores, while being particularly difficult for visual saliency models to perform well on.

Conclusions and future work

In this work, we introduced a new eye-tracking dataset dedicated to studying visual attention deployment and eye fixation patterns on cinematographic stimuli. Alongside with the gaze points and saliency data, we provide annotations on several film-specific characteristics, such as camera motion, camera angles, or shot size. These annotations allow us to explain a part of the causes of discrepancies between shots in terms of inter-observer visual congruency, and in terms of performances of saliency models.

In particular, we highlight the conclusions of Tangemann et al., (2020) regarding failure cases of state-of-the-art visual attention models. Video stimuli sometimes contain a lot of non-static information, that, in some cases, is more important for directing attention than image-related spatial cues. As directors and editors consciously include a lot of meaning with their choices of cinematographic parameters (camera motion, choice of the shots within a sequence, shot sizes, etc.), we would advocate researchers in the field of dynamic saliency to take a closer look at movie sequences in order to develop different sets of features to explain visual attention.

Looking forward, we can investigate whether or not the high-level cinematic features that we provided would be of help to predict visual deployment by building a model that includes this kind of metadata at the shot level. Another crucial point that we did not pursue is the context of the shot : the order of the shots within the sequence has been proven

Table 4 Scores of two saliency models on the database, depending on hand-crafted editing features

(a) Scores depending on camera motion										
Model	Metric	Static	Track	Zoom	Pan	Tilt	Dolly	Rack Focus		
ACLNet	CC	**0.561**	0.545	0.538	*0.466*	0.488	0.517	0.545		
	NSS	**2.631**	2.610	2.523	*2.138*	2.269	2.481	2.610		
Zhang et al.	CC	0.637	0.608	0.643	*0.556*	0.584	0.615	**0.675**		
	NSS	3.014	2.908	3.118	*2.615*	2.797	3.022	**3.338**		
(b) Scores depending on camera angles										
Model	Metric	Eye	High	Low	Bird	Worm	Top			
ACLNet	CC	0.552	*0.500*	0.525	**0.544**	0.532	0.540			
	NSS	2.602	*2.343*	2.465	**2.699**	2.679	2.628			
Zhang et al.	CC	0.621	*0.582*	0.605	0.648	**0.679**	0.672			
	NSS	2.932	*2.777*	2.918	3.286	**3.513**	3.375			
(c) Scores depending on shot size										
Model	Metric	XCU	BCU	CU	MCU	MS	MLS	LS	VLS	EST
ACLNet	CC	0.526	0.532	**0.586**	0.549	0.497	0.510	*0.473*	0.520	0.512
	NSS	2.596	2.271	**2.689**	2.677	2.497	2.481	*2.255*	2.478	2.543
Zhang et al.	CC	0.656	0.607	**0.663**	0.645	0.580	0.615	*0.567*	0.628	0.636
	NSS	**3.320**	*2.679*	3.099	3.186	2.889	3.027	2.733	3.089	3.221

Highest score for each metric and each model is in bold, lowest score is italicized. CC: Person's correlation coefficient; NSS: normalized scanpath saliency

to influence gaze patterns (Loschky, Larson, Magliano, & Smith, 2014; Loschky, Larson, Smith, & Magliano, 2020). As these questions have been tackled from a psychological or cognitive point of view, they remain to be studied by the computer science part of the field, and to be included in visual attention models. This would greatly benefit multiple areas in the image processing field, like video compression for streaming, or automated video description.

Furthermore, we hope that these data would help cinema scholars to quantify potential perceptual reasons to filmmaking conventions, assess continuity editing on sequences, and hopefully improve models of automated edition (Galvane, Christie, Lino, & Ronfard, 2015).

While we proposed an exploration through the lens of deep saliency models, it should be noted that this dataset can be of use for researchers in other fields. The annotated shots and edits can be used by film researchers to explore and propose new definitions for cinematic conventions (shot size or camera motion categories, for instance). From the perspective of filmmakers, this dataset gives interesting feedback of the audience's point of view, validating or not their artistic intentions. For instance, one could compare the obtained gaze-tracks to the expectations of professional filmmakers.

Finally, developing automated tools to extract similar high-level cinematic information could be particularly of interest, both for the design of such tools, as it would give cues on the way to design better visual attention models on cinematographic data, but also with its outcome, as it would allow the provision of large-scale annotated cinematic databases, which would give a new—quantitative—dimension to research on movie contents by cinema scholars.

While some annotations can already be extracted automatically today, such as bounding boxes for characters, shot boundaries, and to a lesser extent camera motion, many cinematographic properties still require the human eye to be recognized, like the size of the shot for instance. The recent work of Courant, Lino, Christie, and Kalogeiton (2021) show promising results, being able to accurately detect camera motion, but also frame layering, which could be used to infer a shot size.

We identify two different ways to incorporate information extracted from these cinematic features in visual attention models.

The first one would be to train deep learning models on cinematic style-related problems, such as director recognition, or style classification, using large-scale datasets such as MovieNet (Huang, Xiong, Rao, Wang, & Lin, 2020). The resulting deep representations of the movie clips could then be concatenated with more traditional image processing deep features to train a visual saliency model. However, the main drawback of this approach is its lack of explainability: establishing the relationship between a deep representation and understandable features is indeed a very difficult task.

The second way would be to identify important cinematic features beforehand, which is what we have initiated in this work. Once an interesting feature has been identified, one can design a way of automatically extracting it. For instance, image segmentation models can be used to infer the visible area of the characters on screen. Combining this information with the frame layering (i.e., which characters are in the foreground or in the background) could reliably give the shot size. Once those features are extracted, they could easily be included in visual attention models as biases.

Declarations

Open practices statement The study was not preregisered; the dataset generated and analyzed during the current study is available on the following page: https://github.com/abruckert/eye_tracking_filmmaking .

Conflict of Interests The authors declare that they have no conflicts of interest.

References

Bak, C., Kocak, A., Erdem, E., & Erdem, A. (2018). Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, *20*(7), 1688–1698. https://doi.org/10.1109/TMM.2017.2777665.

Bordwell, D., Staiger, J., & Thompson, K. (1985) *The classical Hollywood cinema: Film style & mode of production to 1960*. New York: Columbia University Press. https://books.google.fr/books?id=OtDJktQYxo0C.

Borji, A. (2019). Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2019.2935715.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89.

Borji, A., Ahmadabadi, M.N., & Araabi, B.N. (2011). Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications*, *22*, 61–76. https://doi.org/10.1007/s00138-009-0192-0.

Borji, A., Sihite, D.N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69. https://doi.org/10.1109/TIP.2012.2210727.

Breathnach, D. (2016). Attentional synchrony and the effects of repetitive movie viewing. In *AICS*. http://ceur-ws.org/Vol-1751/AICS_2016_paper_57.pdf.

Breeden, K., & Hanrahan, P. (2017). Gaze data for the analysis of attention in feature films. *ACM Transactions on Applied Perception*, *14*(4). https://doi.org/10.1145/3127588.

Brown, B. (2016). Cinematography: theory and practice: image making for cinematographers and directors, Taylor & Francis, chap 2, p. 26.

Bruce, N.D.B., & Tsotsos, J.K. (2005). Saliency based on information maximization. In *Proceedings of the 18th international conference on neural information processing systems, NIPS'05*. https://doi.org/10.5555/2976248.2976268(pp. 155–162).

Bruckert, A., Lam, Y.H., Christie, M., & Le Meur, O. (2019). Deep learning for inter-observer congruency prediction. In *2019 IEEE international conference on image processing (ICIP)*. https://doi.org/10.1109/ICIP.2019.8803596 (pp. 3766–3770).

Cerf, M., Harel, J., Einhaeuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems*, (Vol. 20 pp. 241–248).

Chua, H.F., Boland, J.E., & Nisbett, R.E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences*, *102* (35), 12629–12633. https://doi.org/10.1073/pnas.0506162102.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Predicting human eye fixations via an LSTM-based saliency attentive model. *IEEE Transactions on Image Processing*, *27*(10), 5142–5154. https://doi.org/10.1109/TIP.2018.2851672.

Courant, R., Lino, C., Christie, M., & Kalogeiton, V. (2021). High-level features for movie style understanding. In *ICCV 2021 Workshop on AI for creative video editing and understanding*.

Cutting, J.E., & Armstrong, K.L. (2016). Facial expression, size, and clutter: Inferences from movie structure to emotion judgments and back. *Attention, Perception, & Psychophysics*, *78*, 891–901. https://doi.org/10.3758/s13414-015-1003-5.

Dorr, M., Martinetz, T., Gegenfurtner, K.R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 28–28. https://doi.org/10.1167/10.10.28.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., ..., Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2758–2766).

Duchowski, A.T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods Instruments, & Computers*, *34*, 455–470. https://doi.org/10.3758/BF03195475.

Findlay, J.M. (1997). Saccade target selection during visual search. *Vision Research*, *37*(5), 617–631. https://doi.org/10.1016/S0042-6989(96)00218-0.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2). https://doi.org/10.1167/8.2.6.

Galvane, Q., Christie, M., Lino, C., & Ronfard, R. (2015). Camera-on-rails: Automated computation of constrained camera paths. In *Proceedings of the 8th ACM SIGGRAPH conference on motion in games*. https://doi.org/10.1145/2822013.2822025 (pp. 151–157).

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(6), 989–1005. https://doi.org/10.1109/TPAMI.2009.27.

Gitman, Y., Erofeev, M., Vatolin, D., Bolshakov, A., & Fedorov, A. (2014). Semiautomatic visual-attention modeling and its application to video compression. In *2014 IEEE international conference on image processing (ICIP)*. https://doi.org/10.1109/ICIP.2014.7025220 (pp. 1105–1109).

Goldstein, R.B., Woods, R.L., & Peli, E. (2007). Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine*, *37*(7), 957–964. https://doi.org/10.1016/j.compbiomed.2006.08.018.

Gorji, S., & Clark, J.J. (2018). Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *2018 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00783 (pp. 7501–7511).

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing*, *19*(1), 185–198. https://doi.org/10.1109/TIP.2009.2030969.

Hadizadeh, H., & Bajic, I.V. (2014). Saliency-aware video compression. *IEEE Transactions on Image Processing*, *23*, 19–33. https://doi.org/10.1109/TIP.2013.2282897.

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F., ..., Stadler, J. (2016). A StudyForrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data, 3*.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Proceedings of the 19th international conference on neural information processing systems, NIPS'06*. https://doi.org/10.5555/2976456.2976525(pp. 545–552).

Harezlak, K., & Kasprowski, P. (2018). Application of eye tracking in medicine: A survey, research issues and challenges. In *Computerized medical imaging and graphics*. https://doi.org/10.1016/j.compmedimag.2017.04.006, (Vol. 65 pp. 176–190).

Harezlak, K., Kasprowski, P., Dzierzega, M., & Kruk, K. (2016). Application of eye tracking for diagnosis and therapy of children with brain disabilities. In *Intelligent decision technologies*. https://doi.org/10.1007/978-3-319-39627-9_28 (pp. 323–333).

Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D. (2020). Movienet: A holistic dataset for movie understanding. pp 709–727. https://doi.org/10.1007/978-3-030-58548-8_41.

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE international conference on computer vision (ICCV)*. https://doi.org/10.1109/ICCV.2015.38 (pp. 262–270).

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558.

Jiang, L., Xu, M., Liu, T., Qiao, M., & Wang, Z. (2018). DeepVS: A deep learning-based video saliency prediction approach. In *2018 European conference on computer vision (ECCV)*. https://doi.org/10.1007/978-3-030-01264-9_37 (pp. 625–642).

Jodogne, S., & Piater, J.H. (2007). Closed-loop learning of visual control policies. *Journal of Artificial Intelligence Research*, *28*(1), 349–391. https://doi.org/10.5555/1622591.1622601.

Kanan, C., Tong, M.H., Zhang, L., & Cottrell, G.W. (2009). Sun: Top-down saliency using natural statistics. *Visual Cognition*, *17*(6-7), 979–1003. https://doi.org/10.1080/13506280902771138.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M.P. (2014). What do saliency models predict? *Journal of Vision, 14*(3). https://doi.org/10.1167/14.3.14.

Kroner, A., Senden, M., Driessens, K., & Goebel, R. (2020). Contextual encoder–decoder network for visual saliency prediction. *Neural Networks*, *129*, 261–270. https://doi.org/10.1016/j.neunet.2020.05.004.

Kruthiventi, S.S.S., Ayush, K., & Babu, R.V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, *26*(9), 4446–4456. https://doi.org/10.1109/TIP.2017.2710620.

Kümmerer, M., Wallis, T.S.A., Gatys, L.A., & Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE international conference on computer vision (ICCV)*. https://doi.org/10.1109/ICCV.2017.513 (pp. 4789–4798).

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, *45*(1), 251–266. https://doi.org/10.3758/s13428-012-0226-9.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 802–817. https://doi.org/10.1109/TPAMI.2006.86.

Le Meur, O., Coutrot, A., Liu, Z., Rämä, P., Le Roch, A., & Helo, A. (2017). Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing*, *26*(10), 4777–4789. https://doi.org/10.1109/TIP.2017.2722238.

Loschky, L., Larson, A., Magliano, J., & Smith, T.J. (2014). What would jaws do? the tyranny of film and the relationship between gaze and higher-level comprehension processes for narrative film. *Journal of Vision, 14*(10). https://doi.org/10.1167/14.10.761.

Loschky, L.C., Larson, A.M., Smith, T.J., & Magliano, J.P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, *12*(1), 311–351. https://doi.org/10.1111/tops.12455.

Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 171–177. https://doi.org/10.1109/TPAMI.2009.112.

Majaranta, P., & Räihä, K.J. (2002). Twenty years of eye typing: Systems and design issues. In *Proceedings of the 2002 symposium on eye tracking research & applications, ETRA '02*. https://doi.org/10.1145/507072.507076 (pp. 15–22).

Mathe, S., & Sminchisescu, C. (2015). Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(7), 1408–1424. https://doi.org/10.1109/TPAMI.2014.2366154.

Mital, P.K., Smith, T.J., Hill, R.L., & Henderson, J.M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, *3*(1), 5–24. https://doi.org/10.1007/s12559-010-9074-z.

Pan, J., Sayrol, E., Giro-I-Nieto, X., McGuinness, K., & O'Connor, N.E. (2016). Shallow and deep convolutional networks for saliency prediction. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2016.71 (pp. 598–606).

Pan, J., Canton, C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., & i Nieto, X.G. (2017). Salgan: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081.

Project, S.F. (2014). StudyForrest. https://www.studyforrest.org/.

Rahman, A., Pellerin, D., & Houzet, D. (2014). Influence of number, location and size of faces on gaze in video. *Journal of Eye Movement Research*, *7*(2), 891–901. https://doi.org/10.16910/jemr.7.2.5.

Rayner, K., Castelhano, M.S., & Yang, J. (2009). Eye movements when looking at unusual/weird scenes: Are there cultural differences? *Journal of Experimental Psychology: Learning Memory, and Cognition*, *35* (1), 254–259. https://doi.org/10.1037/a0013508.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv.

Ronfard, R., Gandhi, V., & Boiron, L. (2013). The prose storyboard language: A tool for annotating and directing movies. *AAAI Workshop on Intelligent Cinematography and Editing*.

Rudoy, D., Goldman, D.B., Shechtman, E., & Zelnik-Manor, L. (2013). Learning video saliency from human gaze using candidate selection. In *2013 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2013.152 (pp. 1147–1154).

Smith, T.J. (2013). Watching you watch movies : Using eye tracking to inform cognitive film theory, pp. 165–192. https://doi.org/10.1093/acprof:oso/9780199862139.003.0009.

Smith, T.J., & Henderson, J. (2008). Attentional synchrony in static and dynamic scenes. *Journal of Vision*, *8*(6), 774. https://doi.org/10.1167/8.6.773.

Smith, T.J., & Mital, P.K. (2013). Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of Vision*, *13*(8), 16–16. https://doi.org/10.1167/13.8.16.

Smith, T.J., Levin, D., & Cutting, J.E. (2012). A window on reality: Perceiving edited moving images. *Current Directions in*

Psychological Science, *21*(2), 107–113. https://doi.org/10.1177/0963721412437407.

Tangemann, M., Kümmerer, M., Wallis, T.S.A., & Bethge, M. (2020). Measuring the importance of temporal features in video saliency. In *2020 European conference on computer vision (ECCV)*. https://doi.org/10.1007/978-3-030-58604-1_40 (pp. 667–684).

Tatler, B.W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4–4. https://doi.org/10.1167/7.14.4.

Tavakoli, H.R., Borji, A., Rahtu, E., & Kannala, J. (2019). Dave: A deep audio-visual embedding for dynamic saliency prediction. arXiv:1905.10693.

Thompson, R., & Bowen, C. (2009). Grammar of the Shot.

Valuch, C., & Ansorge, U. (2015). The influence of color during continuity cuts in edited movies: an eye-tracking study. *Multimedia Tools and Applications*, *74*, 10161–10176. https://doi.org/10.1007/s11042-015-2806-z.

Vig, E., Dorr, M., & Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *2014 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2014.358 (pp. 2798–2805).

Wang, W., Shen, J., Guo, F., Cheng, M.M., & Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00514 (pp. 4894–4903).

Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., & Borji, A. (2019). Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2019.2924417.

Wu, H.Y., Galvane, Q., Lino, C., & Christie, M. (2017). Analyzing elements of style in annotated film clips. In *Eurographics workshop on intelligent cinematography and editing*. https://doi.org/10.2312/wiced.20171068.

Wu, H.Y., Palù, F., Ranon, R., & Christie, M. (2018). Thinking like a director: Film editing patterns for virtual cinematographic storytelling. *ACM Transactions on Multimedia Computing, Communications, and Applications, 14*(4). https://doi.org/10.1145/3241057.

Yu, S.X., & Lisin, D.A. (2009). Image compression based on visual saliency at individual scales. In *Advances in visual computing (ISVC)*. https://doi.org/10.1007/978-3-642-10331-5_15 (pp. 157–166).

Zhang, K., & Chen, Z. (2019). Video saliency prediction based on spatial-temporal two-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(12), 3544–3557. https://doi.org/10.1109/TCSVT.2018.2883305.

Zhang, R., Saran, A., Liu, B., Zhu, Y., Guo, S., Niekum, S., ..., Hayhoe, M. (2020). Human gaze assisted artificial intelligence: A review. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20*. https://doi.org/10.24963/ijcai.2020/689 (pp. 4951–4958).

Zünd, F., Pritch, Y., Sorkine-Hornung, A., Mangold, S., & Gross, T. (2013). Content-aware compression using saliency-driven image retargeting. *2013 IEEE International Conference on Image Processing (ICIP)*, pp. 1845–1849. https://doi.org/10.1109/ICIP.2013.6738380.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.