Automated Saliency Prediction in Cinema Studies

Using AI to Map Early Cinema's Use of Visual Saliency

Lein de Leon Yong and Suren Jayasuriya

Abstract: In visual cognition research, saliency refers to the prominence of specific elements in a scene. Moreover, saliency guidance is part of a filmmaker's toolset to build narratives and guide the audience into emotive responses. This article compares two Convolutional Neural Network (CNN) saliency mapping models with viewers' eye-position mapping to investigate the potentiality of automated saliency mapping in moving image studies by analyzing saliency's role during cinema's transition from one-shot to multiple-shot. Although the exact moment when montage and editing methods appeared cannot be identified with precision, findings suggest one of the reasons for this transition was saliency guidance, hence its preponderance.

Keywords: cinema, editing, AI, montage, saliency, silent

Cinema has become a profitable worldwide industry because its comprehension is based on innate mechanisms of human cognition (Carroll 1996, 80–81). Saliency detection, our innate capacity of categorizing certain areas of a scene, is part of this process of object identification and environmental cognition. Furthermore, film editing is one of the most powerful tools a filmmaker has to direct and redirect our attention by nudging the viewer toward affective responses. This article aims to study the use of attention guidance to build narratives by comparing saliency maps produced by three sources: two convolutional neural network (CNN) saliency mapping models: the Contextual Encoder-Decoder Network for Visual Saliency Prediction (CENDVS) (Kroner et al. 2020, 261–270) and the Temporally-Aggregating Spatial Encoder-Decoder network for video saliency detection (TASED) (Min and Corso 2019). On the other hand, viewer-based eye-position mapping will be employed to test the correlation between gaze behavior and CNN models' automated predictions. The present anal-

doi: 10.3167/proj.2023.170303
Volume 17, Issue 3, Winter 2023: 43–63
ISSN 1934-9688 (Print), ISSN 1934-9696 (Online)

ysis will track the role saliency played in the historical transformation of early cinematic narratives between 1895 and the mid-1920s, when films went from one-shot attractions to increasingly sophisticated stories composed of multiple shots (Bottomore 1990, 107). Looking at the selected films, representative of distinct stages of this evolution, the limits and possibilities of automated saliency mapping tools to study the moving image will be tested.[1]

Saliency

The information perceived through the eyes is much larger than what the brain can process and understand. To address this issue, our brains have evolved to prioritize certain parts of the visual landscape and simplify cognition. The fovea, the eye region with the highest density of photosensitive cells, has the function of accomplishing the initial selection from the visual field through eye movement toward the most prominent or salient elements in a scene, fixating on the area with repeated saccadic eye movements (American Academy of Ophthalmology 2017; Itti et al. 1998, 1254–1259). Our eyes move continuously for eighty percent of our awake time, even when our gaze appears as fixed our eyes continue to move in unnoticeable micro-saccades crucial for image perception (Martinez-Conde and Macknik 2013, 95–114).

Gaze allocation has three stages. The first, known as saccadic, is when the eye quickly moves to an initial specific salient area. The second phase is when the eye gradually starts to explore the scene by looking at more areas around the initial point. The third phase is when the eye reaches a steady state and mostly stays within a specific area (Schütt et al. 2019).

This process happens quickly and automatically when something catches our eye, and more slowly and deliberately when searching for something specific (Itti et al. 1998, 1254–1259). However, our brains do not process most of the information that our eyes behold. Instead, the brain has to filter the visual information and select only a small portion of information to be processed in more detail to understand it. Attention can be either overt, that is, effected by moving our eyes, or covert, when we detect something peripherally while our eyes are fixated on another object (Itti and Koch 2000, 1489–1506). Additionally, researchers acknowledge that saliency detection may not be uniform across viewers and that individuals' subjective interests, preferences, aspirations, and cultural context may impact what they perceive as salient (American Academy of Ophthalmology 2017). As a result, this process of saliency detection involves both the fovea's quick bottom-up mechanism and the brain's slower psychological, top-bottom activity that includes cultural memory, goal pursuit, and other cognitive processes (Neisser 1964; Tatler et al. 2011; Veale et al. 2017).

Typically, formal features of the image thar will immediately trigger the bottom-up focus include high contrast objects, highly lighted areas, movement, human figures, faces and letters. However, an essential factor to consider when analyzing saliency is center bias: a significant percentage of eye movements will be directed toward the center of the screen. In a similar way, algorithmic saliency models also have this centered tendency, implying a potential risk of center biased gaze predictions (Tseng et al. 2009).

Automated Saliency Detection Mapping

A saliency detection model is an automated computer vision algorithm that identifies and predicts which regions within an image a viewer most likely will look at. These saliency mapping models require training through an extensive collection of images previously analyzed by showing them to viewers while tracking their gaze to generate eye-position maps; this set of images and saliency maps is known as the ground truth data set. The automated model reads in the pixel values of the input images, and then performs a series of filtering and processing to eventually produce a saliency map that is identical to the ground truth. For CNN-based saliency detection models, the different network layers perform low-level feature extraction and high-level context mapping to identify patterns in the pixel values that correlate to saliency. The learned weights of the neural network are then subsequently used by the network to produce a saliency map for a given image (e.g., an image it has never seen before in training). The resulting output is a potentially reliable representation of the most prominent regions of the input image. Saliency detection models are used for various applications such as image compression, object detection, and image segmentation.

A saliency detection model is an automated computer vision algorithm that identifies and predicts which regions within an image a viewer most likely will look at.

TASED, one of two CNN-based saliency mapping models used in this article, is a network architecture created for video saliency detection and is comprised of two main components: an encoder network that extracts low-resolution spatial (compositional) features from a sequence of frames and a prediction network that creates saliency maps. The authors of TASED assert that their model can predict the salient areas of any given frame considering a limited number of past frames (Min and Corso 2019), estimating visual saliency with consideration of change over time or movement.

The second saliency model used in this article is the Contextual Encoder-Decoder for Visual Saliency (CEDNVS). This neural network model emulates human responses to visual stimuli and predicts which areas of an image a human viewer will perceive as salient. The TASED and CEDNVS models predict saliency on semantic information in addition to low-level

feature contrasts in the pixel values (Kroner et al. 2020). Although training both models involved datasets that incorporate attention (a top-bottom process), it is noteworthy that the TASED model utilized an eye tracking methodology while the CENDVS model employed an attention mapping scheme.

Methods

The research corpus, composed of eight silent films, was analyzed frame by frame at a normalized speed of twenty-four frames per second using saliency mapping methods: CENDVS, and TASED; and with the viewer's eye-position online mapping tool, realeye.io. This web sourced eye-position mapping data was gathered by presenting the corpus to a sample of ten participants. The testing group age ranged between twenty-six and seventy-three years old, including three males and seven females, living in four different countries. The real eye online eye tracking apparatus uses the participants' personal computers built-in webcams through a web browser to record gaze movement. In our case, the viewers were provided a participation link, and simply instructed to watch the videos, performing a free viewing task. In order to calibrate the system, each participant was asked to grant access to their webcam and then to follow with their eyes a red dot that appeared on their screen on three different backgrounds: white, black, and gray, in order to limit the potential influence of the monitor's light intensity on the test results. Subsequently, the calibration was checked using a set of nine points, where the viewer had to fixate their gaze until the gaze was recognized by the webcam. The gaze position data's sampling rate range was between thirty and sixty hertz, capturing the eye-position of the viewer between thirty to sixty times per second, depending on the specifications of the webcam and the quality of the internet connection. This data included the coordinates of the eyes and cursor, with the top left corner of the screen serving as the reference point (0,0) (Wisiecka et al. 2022). The videos were constrained to a sixty-second time frame due to the limitations of the eye tracking platform. Finally, video files with gaze-based saliency maps were exported from the website and evaluated for accuracy through two methods: a quantitative correlation coefficient comparison with data derived from saliency mapping models and qualitative visual estimation of the assembled videos on a non-linear editing software (NLE). A visual representation of this workflow is shown in Figure 1.

Furthermore, to perform the quantitative comparison, three versions of each clip output—CENDVS, TASED, and eye-position mappings—were exported from the NLE with their respective saliency map masks, frame by frame, in .png format, a total of 24,296 image files. The data set was

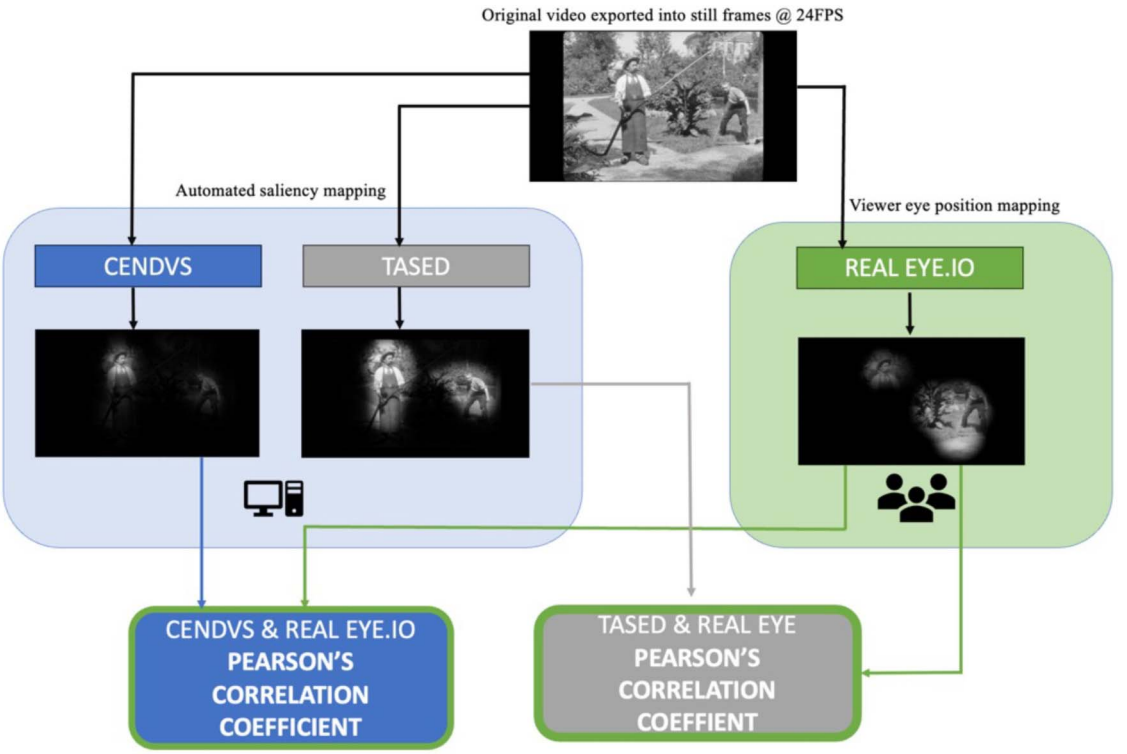


Figure 1. To perform the tests, first the original videos were exported frame by frame at 24 FPS. Afterwards, they were analyzed by two CNN automated saliency and one viewer position mapping methods. Finally, the Pearson's correlation coefficient between the automated saliency mapping and the eye-position mapping was calculated.

then analyzed with a Python code for calculating correlative statistics. Using an open-source Python code repository for computing saliency metrics (Sharma 2017), the statistical Pearson's correlation coefficient of the similarity or dissimilarity between the saliency maps and the eye-position maps was computed. The resulting values of the Pearson's correlation coefficient range from –1 to 1. A score of –1 indicates a perfect opposite correlation; a negative and its positive print would be an example of this; 0 indicates random or no correlation at all, and 1 indicates perfect positive correlation, or that the examples are identical. Moreover, a correlation within the range of 0.1 to 0.3 indicates a weak correlation, 0.3 to 0.5 indicates a moderate correlation, and 0.5 to 1 indicates a strong correlation (Kent State University 2017; Kirch 2008, 1090–1091). Finally, standard deviation was measured to reflect the variability of the correlation coefficient values. A low standard deviation suggests that the values in a dataset exhibit proximity to the mean, also known as the anticipated value, and correspondingly indicate a greater probability. In contrast, a large standard deviation signifies that the

values are dispersed across a broader range, implying a lower probability (Bland and Altman 1996).

Of the eight films analyzed, all scored above the random correlation range. Besides the CENDVS mapping of *Arroseur et Arrosé* (Lumière Brothers, 1895), which produced a Pearson's correlation coefficient lower than 0.3, a weak correlation score, all the results were in the moderate-to-strong correlation ranges. Moreover, the later version of *Arroseur et Arrosé* (Lumière Brothers, 1897) displayed the most significant correlation, 0.6, followed by *Taking President McKinley's Body* (Edison, 1901), with 0.6, which are considered strong correlation scores. A complete visual survey of these results can be seen in Figure 2.

Thus, automated saliency detection has a moderate-to-strong accuracy as an analysis tool in moving image studies. This technique, an automated form of predicting how viewers would attend to specific features

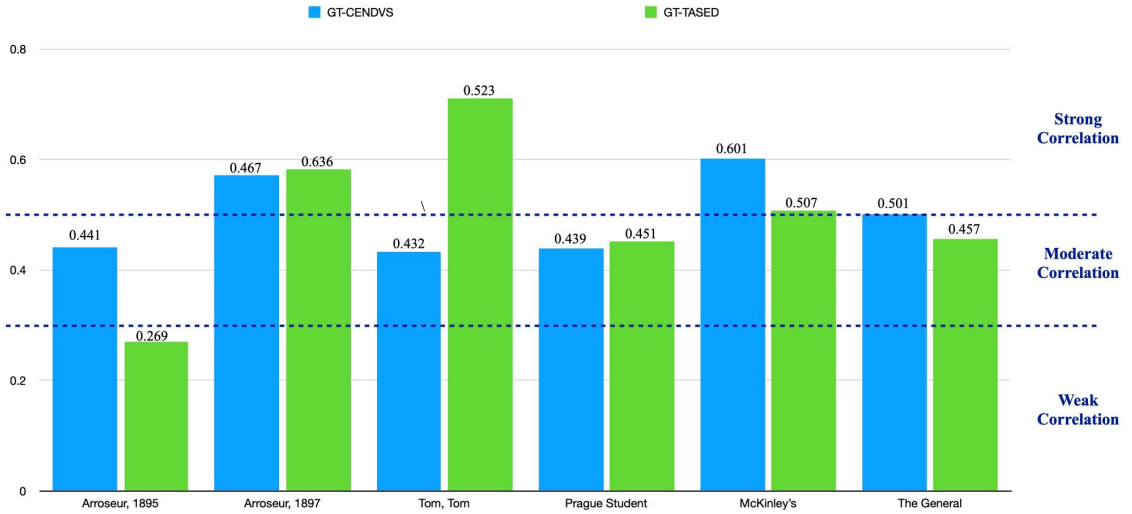


Figure 2. The accuracy with which saliency maps (CENDVS/TASED) approximate human eye-position map is measured by their Pearson's correlation coefficients, reported by GT-CENDVS and GT-TASED respectively for the two neural network models. As evidenced by the data, a majority of the saliency maps exhibited moderate correlation.

Pearson's CC Mean and Standard Deviation

	MEAN		STANDARD DEVIATON	
FILM	**GT-CENDVS**	**GT-TASED**	**GT-CENDVS**	**GT-TASED**
Arroseur, 1895	0.441	0.269	0.211	0.247
Arroseur, 1897	0.572	0.582	0.180	0.379
Tom, Tom	0.432	0.710	0.125	0.187
Prague Student	0.439	0.451	0.208	0.274
McKinley's	0.601	0.507	0.191	0.233
The General	0.501	0.457	0.189	0.212
MEAN	0.498	0.496	0.184	0.255

on a shot or series of shots, reduces cost and time that eye tracking-based research would consume. Furthermore, automated mapping methods allow the analysis of vast collections of audiovisual pieces, facilitating the possibility of different approaches such as a statistical analysis of a large corpus.

Saliency and Film

Film style is partially constructed from evolutionary adaptation responses. Rapid movements, facial expressions, and other people's gazes are universal attention-getters. A recorded image can produce eye behavior similar to the one induced by reality with its dynamic features. However, culture teaches people to pay attention to certain things, but such learned skills involve fine-tuning of preexistent physiological and perceptual abilities (Bordwell 1997, 158–272). Thus, the allure of cinema has evolutionary roots, even though there are other culturally conditioned reasons for films to appeal to viewers. Nevertheless, edited videos significantly impact our visual cognitive experience, leading to different gaze behaviors, eye movements, and memory creations than those generated when perceiving the real world or continuous video (Tatler et al. 2011). Films reorganize and select actions and events with an economy, legibility, and coherence that outperform naturally perceived events (Carroll 1996, 86; Tatler et al. 2011). This aspect of cognitive construction, built partly from saliency management, may also explain the general degree of engagement generated by films (Bordwell and Thompson 2008, 158–272).

Filmmakers resort to mechanisms of saliency assignment through framing, lighting, image composition, movement, shot-by-shot relationship, color palette, actors' sight direction, optical focus, makeup, costumes, editing, and sound to control the viewer's perception of the film (Bordwell 1997, 1–11). Thus, moving images can be analyzed through saliency as a structural categorization of the information conveyed in their images and sounds through the mentioned elements. Finally, what David Bordwell defines as film style can also be considered as attention guidance tools (ibid., 4).

Bordwell divides film style into: mise-en-scène (everything that appears inside the frame), cinematography (similar to Eisenstein's concept of *mise-en-cadre*) (Eisenstein 1957, 168), editing, and sound (Bordwell and Thompson 2008, 4). He explores the idea that a film director's skill resides, in part, to persuade viewers to attend to certain picture sections at specific times. Such appeals to attentional processes may lead to the dubious conclusion that films force the audience to attend to only one part of the frame, transforming the viewer into an entity with no agency. However, the observer can ignore the image's pull by focusing on other areas (Bordwell 1997, 163–

164). Saliency nudges the viewers' gaze more than forcing it toward a particular element. However, there is no univocal manner to guide attention; there are countless ways to do it through film-style elements, probably as many as there are filmmakers (ibid., 267).

Previous Saliency Mapping Research

Renowned Soviet director and film theorist Sergei Eisenstein studied saliency by analyzing paintings and his own films through visual estimation before technological advancements were available to perform such tests. Specifically, on the book *Film Sense*, a collection of writings posthumously published by Jay Leyda, Eisenstein analyzes how the audience would perceive saliency in his films. He posed the rhetorical question, "Have we any right to claim that our film-frames also gauge the eye's movement over a determined path?" (Eisenstein 1957, 194).

Reflecting on his movie *Alexander Nevsky* (Sergei Eisenstein, 1938), in which the musical score was composed by Sergei Prokofiev, Eisenstein noted that the sequence *Battle on Ice* was structured to produce a visual movement that went from left to right throughout the twelve shots matching the melody line of the unfolding scene (Eisenstein 1957, 194). According to Eisenstein's analysis, the primary visual saliency area should be located on the left in the first shots, while in shots five, six, and seven, an element would be found on the right part of the frame that would attract the eye (ibid., 199).

In 2014, Tim J. Smith empirically tested Eisenstein's assessments through eye-position tracking in silent and sound versions of the film. Among his findings were that the mute and audio versions of the test had very slight differences regarding regions of the image that attracted the viewers, suggesting that image is preponderant in audio visual media narrative (Smith 2014, 85–105).

Early Cinema Staging

Films emerged among a tradition of visual arts with composition rules that can be traced back to the Renaissance (Canudo 1993, 13–18). Eisenstein explained in detail depth staging, elucidating how foreground and background interactions might generate a type of montage within the shot. He further defined mise-en-scène as the organizing principle of the dramatic and affective material that would be transformed into cinematic by framing and editing. He also said that *mise-en-cadre*, or framing, could be used with staging to generate a continual gesture that heightened the drama through the use of compositional lines leading attention from one point of interest to the other (Bordwell 1997, 218).

In an early cinematography manual, Louis Lumière advises managing saliency in visual composition by prioritizing an object, which would grab the viewer's attention and then guide their eye to the rest of the frame areas. Unfortunately, Lumière does not further explain how to achieve such eye guidance to the rest of the frame (Chardère 1987, 102). Accordingly, some of Lumière's early films have a visual composition structure that prioritizes certain elements; most likely a contribution of the diverse camera operators that the Lumière brothers hired to shoot footage around the world with their new invention (Aubert and Seguin 2015). Consequently, placing the critical parts in the geometrical center of the composition can be considered as the most elementary saliency guidance strategy.

Arroseur et Arrosé

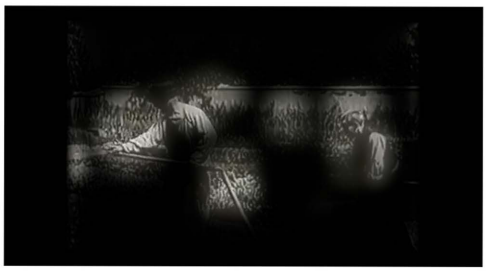
Probably due to box office success, the Lumière brothers produced three versions of the short film *Arroseur et arrosé* (Lumière Brothers, 1895 [first version] and 1897 [second version] and 1897 [third version]) (Aubert and Seguin 2015). In this case, we will consider only the first and third versions from 1895 and 1897, respectively. Even though those short films were produced within three years, notable changes in framing and composition structure appear in these. The first was shot with the early film mise-en-scène practice of frontality, also known as clothesline staging. In contrast, the latter deploys a depth staging scheme. This can be an example of how filmmakers quickly experimented with basic schemas for guiding attention through frontward movement rather than planimetric layout. In addition, deep staging provided more options for driving mise-en-scène prominence by shifting attention from one part of the image to another (Bordwell 1997, 173–174).

The results in Figure 3 show that in comparison to the eye-position maps created from the first version of *Arroseur*, produced in 1895, the saliency maps created with the CENDVS method score a correlation coefficient of 0.1, which is a weak level of correlation. Similarly, the saliency maps generated using the saliency mapping method TASED displayed a correlation coefficient of 0.1, again a weak level of correlation. In contrast, the second version of the film, dated from 1897, display correlation coefficients between eye-position maps and CENDVS and TASED saliency maps of 0.4 (moderate correlation) and 0.6 (strong correlation), respectively. Finally, it is interesting to observe that there is a correspondence between the primary events depicted in the short films and the chart representation of the correlation coefficient, which appears in the form of peaks within the area of strong correlation as shown in Figure 3.

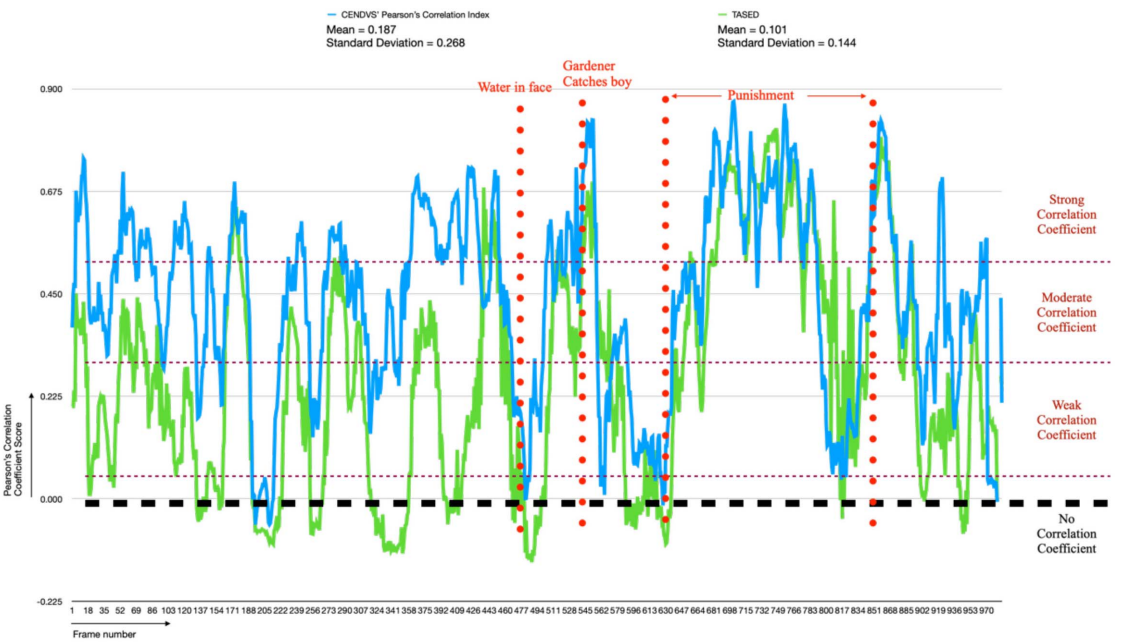
Original

CENDVS

TASED



Eye Tracking



L' ARROSEUR ARROSE 1895

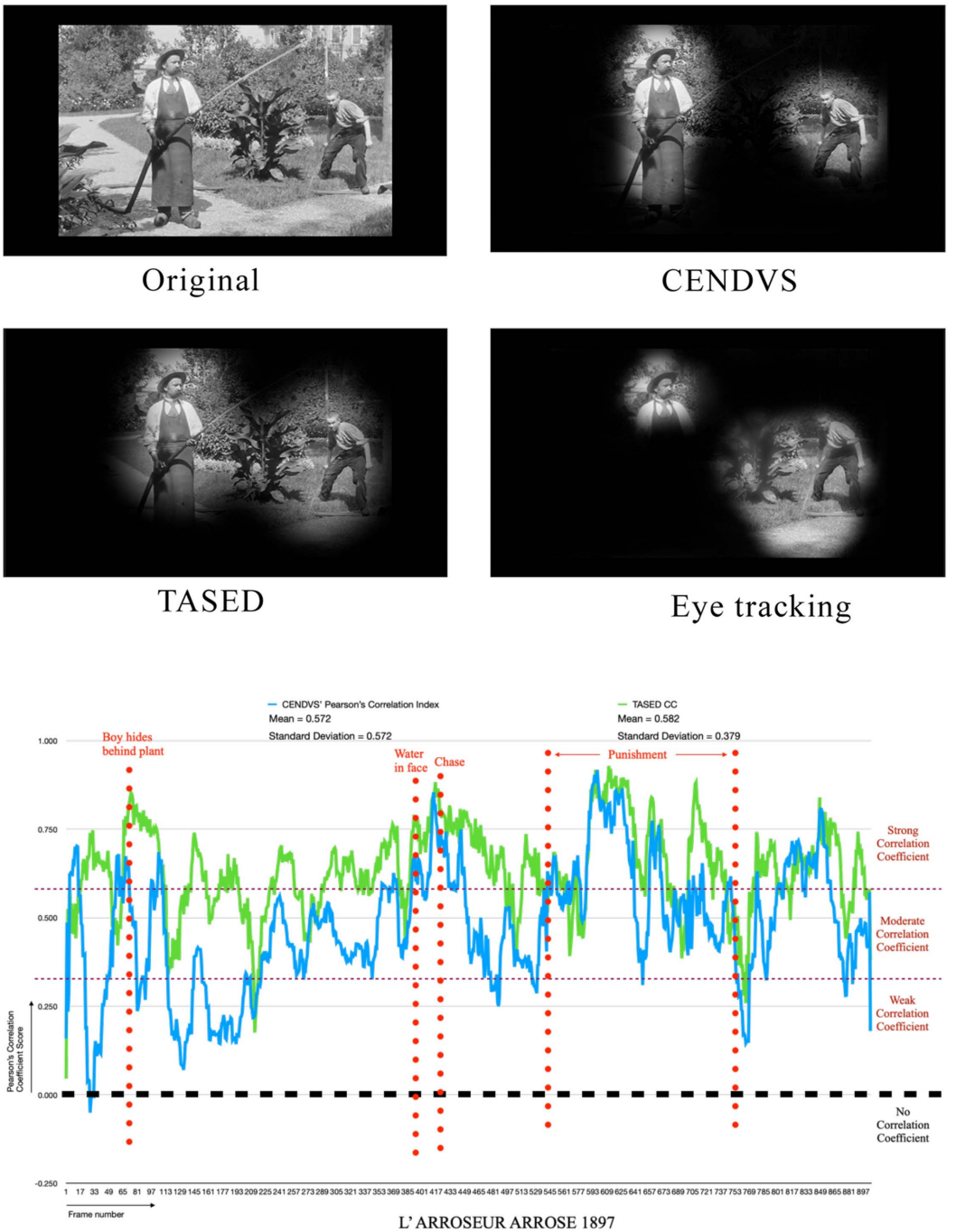


Figure 3. Arroseur et arrosé *(Lumière Brothers, 1895 and 1897). The chart representations of the two Arroseur versions correlation coefficient support the claims stemming from visual evaluations; the latter version (1897) scored higher than the earlier one (1895). However, it is worth noting that the crucial events of the film appear as peaks on the chart.*

Tom, Tom, the Piper's Son

Other early films employed an opposite saliency guidance strategy by disguising the essential elements. For example, in his book *History of Film Style,* Bordwell points out that the film *Tom, Tom, the Piper's Son* (American Mutoscope & Biograph, 1905) lacks a visual composition structure that guides the audience's attention. In this example, the mise-en-scène is so tumultuous and unstructured that it is challenging to appreciate a pig's robbery, the pivotal action of the scene. Furthermore, all the character's figures are presented with similar shapes while a clown juggles in the center area of the frame, he drops one of his balls at the precise moment when the pig is stolen, and the clown blocks the theft view by picking up his ball. The scene coincident actions point toward an intentional obtrusive staging design which appears as such in the quantitative correlation chart, as shown in Figure 4. Furthermore, Pearson's coefficient correlation score for CENDVS was determined to be 0.4, a moderate level of correlation. In contrast, the saliency model TASED, which incorporates movement, exhibited a correlation score of 0.7, indicating a significant level of relationship. Both datasets revealed a standard deviation ranging from 1.5 to 2.

From Staging to Montage

As mentioned before, frame composition is one way to guide attention and is crucial in filmmaking. Additionally, editing, the process of selecting, cutting, and assembling distinct portions of film or video, is unique to cinema. André Bazin developed an essentialist-dualist stance dividing filmmakers into plastic and montage directors. According to the author, the first group tend to stage their films in depth, while the other use cutting between shots to tell their stories. This classification goes as far as identifying plastic editing with European art cinema and shot-cutting with American films. From these categories, and even though saliency plays a vital role in all of them, editing is the one dedicated to highlight specific details of a scene by breaking it down into several shots. Here camera positions are changed between shots to prioritize points of view according to narrative goals. In such cases, when a film is composed of more than one shot, the resulting meaning of the assembled shots is different from the one that each shot has by itself (Bazin 1967, 24–25).

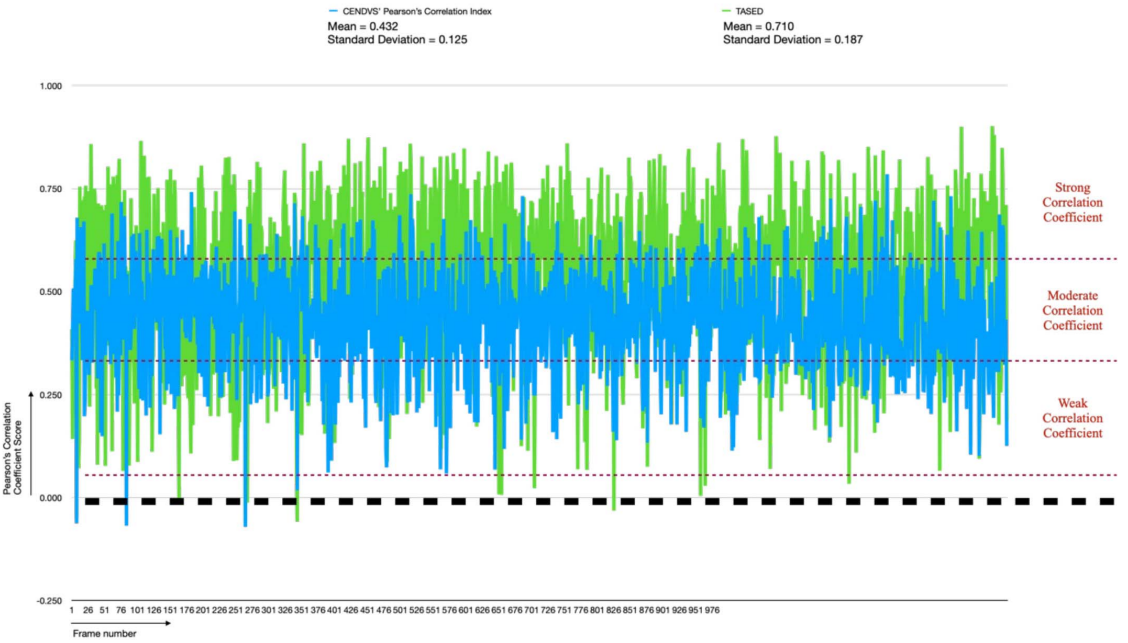
Another author, Ben Brewster (1990), further distinguished between the European and American styles. He described the European style as deep staging with long-lasting shots and the American style with shallow staging and fast cutting. Furthermore, Brewster adds two categories: the Scandinavian style, with depth staging and light-prompted cues, and the Vitagraph style, with depth staging, low camera positions, and a focus on the surroundings as an agent for narrative development (Brewster 1990, 45–55).

Original

CNDVS

Eye Mapping

TASED



Tom, Tom, the Piper's Son

Figure 4. Tom, Tom, the Piper's Son *(American Mutoscope & Biograph, 1905). The copy used of this film flickers probably due to the quality of cinematography capture of the early twentieth century, this contributes considerably to the noise observed in the correlation statistics.*

The Student of Prague

An example of the European, or plastic staging, approach is the German film *The Student of Prague* (Rye, 1913). This film is recognized as an early German Expressionism work where the staging style mirrors the psyche of the character, and it is used as a motif throughout the film to direct the viewers' gaze. Our case of study is the second shot of the film, which showcases an impressive demonstration of saliency management through depth staging.

In Figure 5 we show that the results of this shot's analysis with saliency mapping methods—CENDVS, TASED, and viewer-sourced eye-position maps—produced a moderate correlation value of 0.4 when compared. These values were closely grouped around the mean with a standard deviation of 0.2. Initially, the attention regions oscillate between the people in the background. Afterwards, attention fluxes among background students; here CENDVS maps a student raising his wine glass, which is also noticed by human viewers. TASED maps a background student taking his cap off, which CENDVS doesn't, and, in the case of human viewers, the attention gravitates toward the Prague student, who enters the scene parallel to the camera trajectory. The saliency mapping methods' correlation with the eye-position mapping of this action appear in the chart as two peaks. TASED crests are within the moderate range of correlation, with a coefficient score of 0.6, near the strong correlation threshold. On the other hand, CENDVS scores a significant level of correlation, with coefficient values ranging between 0.7 and 1, a strong-to-perfect correlation range.

Later on, as he walks diagonally toward the camera, increasing in size and visually separating himself from the backdrop, the Prague student becomes the frame's most prominent element until two students approach him. In this moment, the quantitative results of the saliency mapping correlation between the automated saliency mapping methods and eye-position mapping chart a second peak in the graph within the strong correlation range. Two additional peaks appear in the chart inside this strong correlation range when the protagonist begins and finishes the action of sitting down as seen in Figure *5*.

Analytic Montage as Saliency Guidance

Compared to depth staging and visual composition, continuity editing is a more effective tool to guide viewers' attention since it is based on selecting what viewers can or cannot see, leaving any irrelevant narrative elements out. Thus, one of the theoretical explanations for films going from one shot to multiple shots is saliency guidance. Footage selection and camera placement were practices used to prioritize and discriminate elements inside a shot, display a selection of the most salient moments of a scene, and create a new form of visual narrative construction.

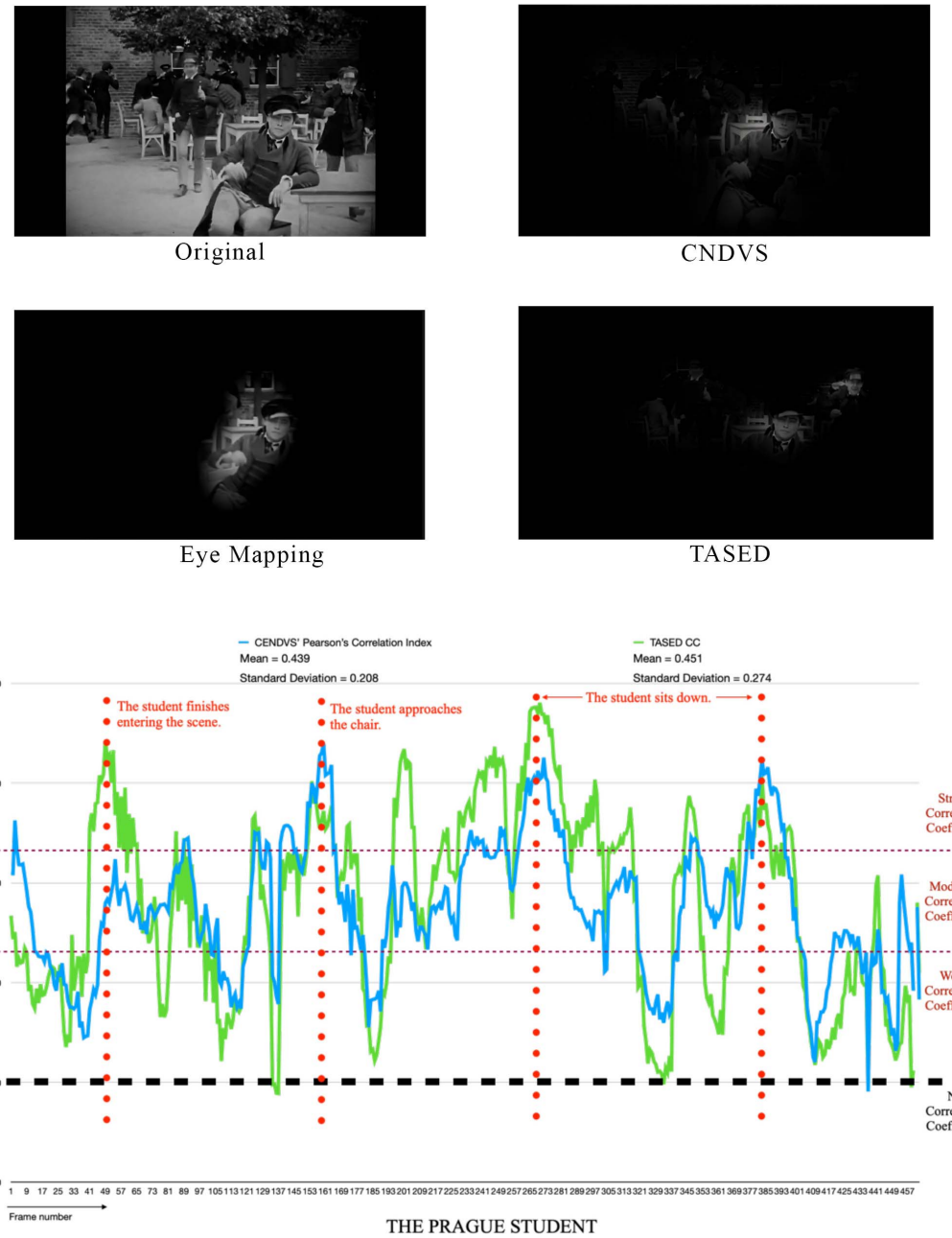


Figure 5. The Student of Prague *(Rye, 1913). The correlation coefficient score peaks correlate with the beginning or conclusion of the principal character movements during the mise-en-scène.*

The British author Stephen Bottomore argues that film editing practices arose with the intention of showing the most salient portions of a scene. This goal was achieved by stopping the camera and running it only during the moments considered relevant (Elsaesser and Bottomore 1990, 104–

Footage selection and camera placement were practices used to prioritize and discriminate elements inside a shot, display a selection of the most salient moments of a scene, and create a new form of visual narrative construction.

113). Bottomore quotes a 1915 cinematography manual: "It is more than usual to expose only the principal features or most striking portions by stopping the handle when one of these has passed and starting it again when the next (moment) appears" (Jones 1915, 23). Early newsreels such as *Taking President McKinley's Body* (Edison, 1901) illustrate this early editing practice.

Taking President McKinley's Body

In this case, *Taking President McKinley's Body*'s visual analysis of the maps generated by the two saliency mapping models and eye-position mapping showed that the coffin was the focus of the viewers' fixations during fifty-one seconds, and three frames out of fifty-three seconds, and seven frames where the coffin appears in the film. In contrast, CENDVS followed the casket entirely during fifty-three seconds and seven frames, although saliency is also registered in other areas. TASED maps the coffin mostly at the time when the coffin is at the center of the frame, probably due to center bias. The film's second shot is remarkable because a wall fills most of the frame, however, viewers followed the casket as registered on the resulting eye-position map, probably due to the flowers on top of it, a different texture from the rest of the image. Finally, it is relevant to note that one of the distinctions between AI-based models and eye tracking is the absence of the delay in saliency mapping between shots, which is a trait of natural gaze behavior.

The Pearson's correlation coefficient between the saliency mapping models and eye-position mapping of *Taking President McKinley's Body* had a mean score of 0.6 for CENDVS and 0.407 for TASED, a strong and moderate correlation score. Both models scored a low standard deviation of around 0.2. The data suggests that through the practice of dividing the film into shots, filmmakers were able to prioritize certain elements, creating a more precise attention guidance process.

Editing was developed by filmmakers with the objective of nudging the viewer's attention by presenting distinct and relevant aspects of a scene. This practice was quickly adopted for two reasons: first, continuity editing-based films attracted more viewers, while dividing the scene into shorter and easier to shoot fragments led to a faster and less expensive production process (Bordwell 1997, 198–199). Finally, the commercial success of multi-shot films can be attributed to the more explicit saliency indicated by cutting and assembling different shots, allowing an easier and more engaging experience.

***The General* (Buster Keaton, 1926)**

Depth staging did not disappear; it remained a practice in addition to editing and montage. Thanks to depth staging style, filmmakers had the potential to guide the viewer's attention through the frame with the aid of visual composition. On the other hand, editing allowed the selection of specific segments of footage granting a more precise saliency selection. One of the earliest examples of this moving image narrative style was directed by Buster Keaton, who employed editing and depth staging to create *The General*, an enduring masterpiece of cinema.

In Figure 6, it can be seen that when compared with each other, the three sources yielded similar results. Particularly TASED produced dynamic maps that resembled the results of eye-positioning. This similarity is noteworthy as it aligns with the fact that the human gaze is inherently dynamic, even when it appears to be fixed. Furthermore, *The General*'s saliency maps are clearly delimited, probably because, as film style further developed, more salient imagery was possible to create. In the studied scene, a combination of analytical editing and depth staging creates a gag where the protagonist arrives at his fiancé's house and knocks at her door, only to discover that she has been following him the entire time. This sequence begins with a full long shot, followed by a full shot of the protagonist's fiancée, whose dress fabric creates a stark contrast with the background; here, the maps show a strongly defined saliency area. The mentioned shots appear as peaks in the chart visualization of Pearson's correlation coefficient of the three methods: the two automated saliency mapping models and the eye-position map. Both AI-based saliency mapping methods relate to eye-position mapping between the upper range of moderate correlation and strong correlation. The next shot is a wide shot of Buster and two children following him; the saliency in the frame is now dispersed in four locations. By the end of the shot, Buster Keaton's saliency mapping has completely disappeared. In the next shot, the characters approach the house's entrance in depth staging style. This moment's correlation coefficient is charted as a peak for both saliency mapping methods. Finally, the lady waits until Buster Keaton knocks on the door and notices her, at this point the gag is solved. Buster Keaton holds his hat in embarrassment, and the girl crosses the scene, inviting the group into her house. Here a group of three peaks in the upper range of strong correlation appear in coincidence with those three actions. Considering the entire scene, both saliency mapping methods scored a moderate correlation coefficient mean of 0.5 for CENDVS and 0.4 for TASED, with a deviation of 0.1 and 0.2, respectively.

More testing is needed to assure that the correlation coefficient score between different saliency prediction models is a point of reference to quantify the prominence of a specific region and that the resulting charts can be a quantitative representation of a narrative structure.

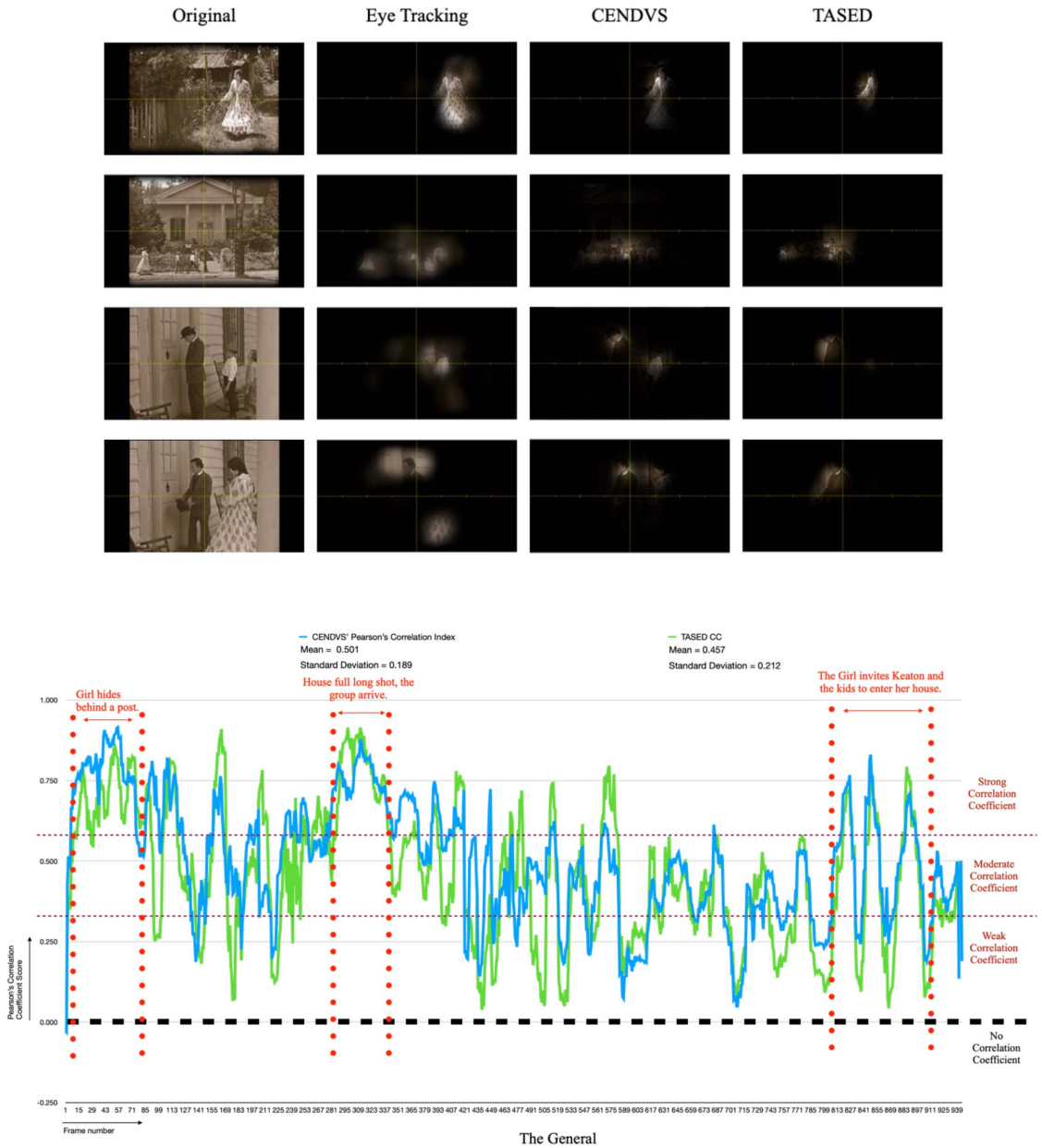


Figure 6. The General *(Buster Keaton, 1926) is an example of saliency guidance through mise-en-scène, mise-en-cadre, and editing. The correlation coefficient coincidence between the two saliency mapping models and eye-position mapping is notable.*

Discussion

Cinema's popularity stems from its ability to capitalize on our innate capability of saliency detection. Since its early days, film has relied on pictorial composition to guide the viewers' gaze while tapping into innate gaze skills. Later, film

editing was added as a tool that saved resources during the film's production and, at the same time, resulted in higher box office income. However, it's crucial to keep in mind that saliency cannot fully account for film style and narration.

Considering the results of this study, saliency mapping models such as CENDVS and TASED have the potential to be a moderate-to-strong precision image analysis tool. More testing is needed to assure that the correlation coefficient score between different saliency prediction models is a point of reference to quantify the prominence of a specific region and that the resulting charts can be a quantitative representation of a narrative structure. Such an approach has the potential to be implemented as a saliency mapping method within a shot, between shots, and even by charting the saliency correlation coefficient scores and their relation to the general narrative structure of a feature film. However, it is always crucial to consider that the results are just an informed conjecture of human gaze behavior. Finally, a further empirical application is selecting footage according to its saliency correlation coefficient during the editing process of a video and contrasting these results with a video edited without saliency mapping methods.

Since its early days, film has relied on pictorial composition to guide the viewers' gaze while tapping into innate gaze skills. Later, film editing was added as a tool that saved resources during the film's production and, at the same time, resulted in higher box office income.

Automated saliency mapping models are trained to predict, on average, what would be considered as important by a human viewer. These models are trained by analyzing data sets made up of images and their respective maps of eye movement. Thus, this process implies a cultural categorization that is passed from the viewer, who exists inside a cultural context, to the automated saliency mapping model. However, it is important to consider that saliency detection is only one of the stages of human vision, preceded by an encoding stage and followed by a decoding stage where figure recognition takes place. Thus, saliency mapping models are not capable of making meaning.

Acknowledgments

This work was supported in part by NEH Grant AKB-279509-21. This article was also possible thanks to the Comexus-Fulbright graduate studies grant program and the Mexico's government organization—CONACYT grant program for foreign graduate studies. Finally, we would like to thank Dr. Ana Hedberg Olenina for her guidance and support through the writing of this article.

Lein de Leon Yong, a film editor from Mexico City who has contributed to fiction series for Netflix and Amazon Prime Video, is currently pursuing PhD

degree at Arizona State University's media arts and sciences program. She earned a master of arts in art history with focus in film studies from Mexico National University in 2018, for which she conducted part of her research at the University of Texas at Austin. She recently published a collective science fiction micro novel, Nova Aera.

Suren Jayasuriya is an assistant professor at Arizona State University, in the School of Arts, Media and Engineering (AME) and Electrical, Computer and Energy Engineering (ECEE). Before this, he was a postdoctoral fellow at the Robotics Institute at Carnegie Mellon University. Suren received his PhD in ECE at Cornell University in January 2017 and graduated from the University of Pittsburgh in 2012 with a BS in mathematics (with departmental honors) and a BA in philosophy. His research interests range from computational cameras, computer vision and graphics, machine learning, sensors, STEAM education, and philosophy.

Note

[1] All the examples mentioned in this paper can be viewed at https://bit.ly/cine_saliency. "

References

American Academy of Ophthalmology. 2017. "Fovea." https://www.aao.org/eye-health/anatomy/fovea (accessed January 2023.).

Aubert, Michelle, and Jean-Claude Seguin. 2015. "Catalogue Lumière." *Conception Multimédia de l'Ecole d'Arts Appliqués de la Chaux-de-Fonds* [in French]. https://catalogue-lumiere.com/?s=Arrose (accessed 13 April 2023).

Bazin, André. (1967) 2005. *What Is Cinema? vol.1*. Ed. Dudley Andrew, trans. Hugh Gray. 3rd ed. Reprint. Los Angeles: University of California Press.

Bland, J. Martin, and Douglas G. Altman. 1996. "Statistics Notes: Measurement Error." *The BMJ* 312. https://doi.org/10.1136/bmj.312.7047.1654.

Bordwell, David. 1997. *On the History of Film Style*. Cambridge, MA: Harvard University Press.

Bordwell, David, and Kristin Thompson. 2008. *Film Art, an Introduction*, 8th ed. New York: McGraw Hill.

Bottomore, Stephen. 1990. "Shots in the Dark—The Real Origins of Film Editing." In *Early Cinema: Space, Frame, Narrative*, ed. Thomas Elsaesser, 104–113. London: British Film Institute.

Brewster, Ben. 1990. "Deep Staging in French Films 1900–1914." In *Early Cinema: Space, Frame, and Narrative*, ed. Thomas Elsaesser, 45–56. London: British Film Institute.

Canudo, Ricciotto. 1993. "I. El cine es un arte [Film is Art]." *Textos y manifiestos del cine. Estética. Escuelas. Movimientos. Disciplinas*. Innovaciones [Cinema's Texts and Manifests. Aesthetic. Schools. Movements. Disciplines], eds. Joaquim Romaguerai Ramió and Homero Alsina Thevenet, 13–18. Madrid: Cátedra.

Carroll, Noël. 1996. *Theorizing the Moving Image*. 1st ed. Reprint. Cambridge: Cambridge University Press.

Chardère, Bernard. 1987. *Lumières Sur Lumière*. Reprint. Lyon: Presses Universitaires de Lyon.

Eisenstein, Sergei. 1957. *The Film Sense*, trans. and ed. Jay Leyda. Reprint. London: Faber.

Itti, Laurent, and Christof Koch. 2000. "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention." *Vision Research* 40 (10–12): 1489–1506. https://doi.org/10.1016/s0042-6989(99)00163-7.

Itti, Laurent, Cristof Koch, and Ernest Niebur. 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.

Jones, Bernard E. 1915. *The Cinematograph Book*. London: Cassell and Company.

Kent State University. 2017. "Pearson Correlation." https://libguides.library.kent.edu/SPSS/PearsonCorr#cite_cohen (accessed July 26, 2023.).

Kirch, Wilhelm, ed. 2008. "Pearson's Correlation Coefficient." *Encyclopedia of Public Health*, 1090–1091, Dresden Germany: Springer Science + Business Media B.V. https://doi.org/10.1007/978-1-4020-5614-7_2569.

Kroner, Alexander, Mario Senden, Kurt Driessens, and Rainer Goebel. 2020. "Contextual Encoder–Decoder Network for Visual Saliency Prediction." Neural Networks 129: 261–270. https://doi.org/10.48550/arXiv.1902.06634.

Martinez-Conde, Susana, and Stephen L. Macknik. 2013. "Microsaccades." In *The Oxford Handbook of Eye Movements*, eds. Simon P. Liversedge, Iain D. Gilchrist, Stafan Everling. Oxford: Oxford University Press.

Min, Kyle, and Jason Corso. 2019. "Tased-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv.2019.00248.

Neisser, Ulric. 1964. "Visual Search." *Scientific American* 210 (6): 94–103. https://www.jstor.org/stable/10.2307/24931530.

Schütt, Heiko H., Lars O. Rothkegel, Hans A. Trukenbrod, Ralf Engbert, and Felix A. Wichmann. 2019. "Disentangling Bottom-up versus Top-down and Low-Level versus High-Level Influences on Eye Movements over Time." *Journal of Vision* 19, no. 3. https://doi.org/10.1167/19.3.1.

Sharma, Tarun. 2017. "Saliency Metrics" Github Repository. https://github.com/tarunsharma1/saliency_metrics/blob/master/salience_metrics.py.

Smith, Tim J. 2014. "Audiovisual Correspondences in Sergei Eisenstein's Alexander Nevsky." In *Cognitive Media Theory*, ed. Ted Nannicelli and Paul Taberham, 85–105. New York: Routledge, Taylor & Francis Group.

Tatler, Benjamin W., Mary Hayhoe, Michael F. Land, and Dana H. Ballard. 2011. "Eye Guidance in Natural Vision: Reinterpreting Salience." *Journal of Vision* 11, no. 5. https://doi.org/10.1167/11.5.5.

Tseng, Po-He, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. 2009. "Quantifying Center Bias of Observers in Free Viewing of Dynamic Natural Scenes." *Journal of Vision* 9, no. 7. https://doi.org/10.1167/9.7.4.

Veale, Richard, Ziad M. Hafed, and Masatoshi Yoshida. 2017. "How Is Visual Salience Computed In The Brain? Insights From Behaviour, Neurobiology And Modelling" in *Philosophical Transactions Of The Royal Society B* 372, no. 1714. https://doi.org/10.1098/rstb.2016.0113.

Wisiecka, Katarzyna, Krzysztof Krejtz, Izabela Krejtz, Damian Sromek, Adam Cellary, Beata Lewandowska, and Andrew Duchowski. 2022. "Comparison of Webcam and Remote Eye Tracking." *Etra '22: 2022 Symposium on Eye Tracking Research and Applications*, Article No. 32, 1–7. https://doi.org/10.1145/3517031.3529615.