Decoupling Light Reflex from Pupillary Dilation to Measure Emotional Arousal in Videos

Pallavi Raiturkar
pallaviraiturkar@ufl.edu
University of Florida

Andrea Kleinsmith
andreak@umbc.edu
University of Maryland, Baltimore County

Andreas Keil
akeil@ufl.edu
University of Florida

Arunava Banerjee
arunava@mail.ufl.edu
University of Florida

Eakta Jain
ejain@cise.ufl.edu
University of Florida

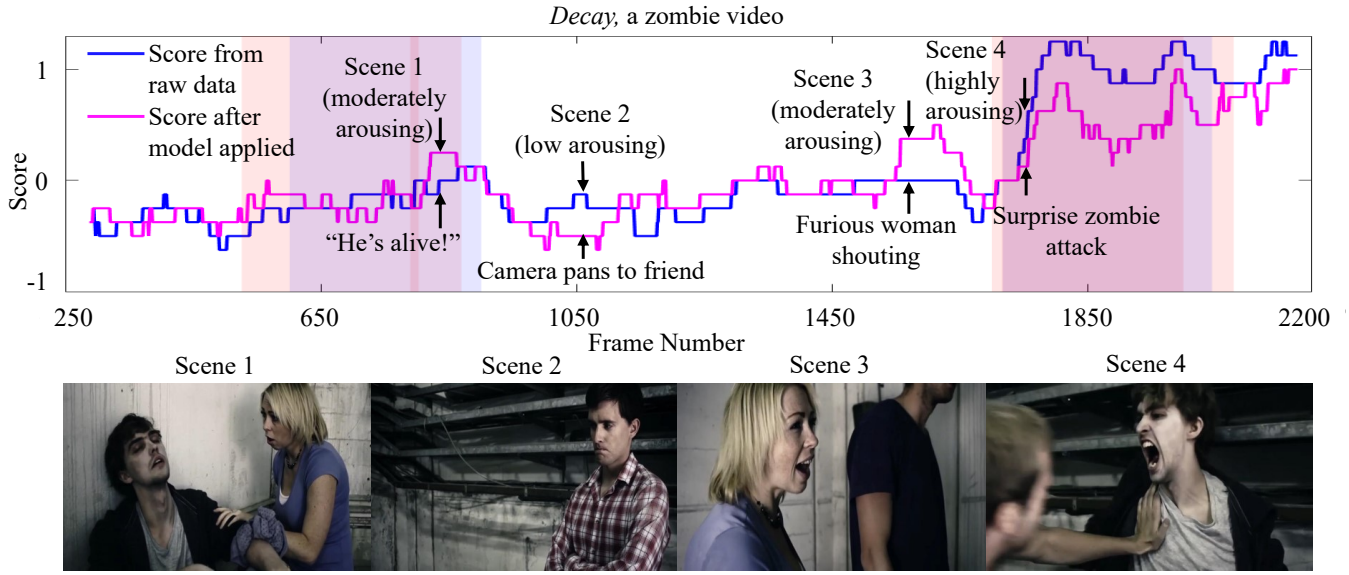

Figure 1: *Measured pupillary diameter has been previously used as an index of arousal, or, exciting-ness in videos. In this paper, we consider whether it is possible to factor out the impact of pupillary light reflex on an exciting-ness score computed from pupil diameter data of viewers watching a video. We model the light reflex as a linear function of the grayscale intensity in the foveal neighborhood of a viewer's gaze point. Top: The score computed from raw pupil measurements is shown in dark blue, while the score computed after our light reflex model has been applied is shown in lighter pink. The colored regions denote the portions reported as "exciting" by three independent coders. Bottom: Representative frames from each of the four scenes where our model has made a significant difference.*

Abstract

Predicting the exciting portions of a video is a widely relevant problem because of applications such as video summarization, searching for similar videos, and recommending videos to users. Researchers have proposed the use of physiological indices such as pupillary dilation as a measure of emotional arousal. The key problem with using the pupil to measure emotional arousal is accounting for pupillary response to brightness changes. We propose a linear model of pupillary light reflex to predict the pupil diameter of a viewer based only on incident light intensity. The residual between the measured pupillary diameter and the model prediction is attributed to the emotional arousal corresponding to that scene. We evaluate the effectiveness of this method of factoring out pupillary light reflex for the particular application of video summarization. The residual is converted into an exciting-ness score for each frame of a video. We show results on a variety of videos, and compare against ground truth as reported by three independent coders.

Keywords: Eyetracking, Linear Model, Video Understanding

Concepts: •**Computing methodologies** → **Video summarization; Perception;** •**Information systems** → *Multimedia and multimodal retrieval;*

SAP '16, July 22 - 23, 2016, Anaheim, CA, USA
ISBN: 978-1-4503-4383-1/16/07
DOI: http://dx.doi.org/10.1145/2931002.2931009

1 Introduction

Predicting the regions of high intensity, or, peak emotional arousal in videos is a widely relevant problem because of applications such as video summarization, and video search and recommendation. Computational solutions to this problem have sought to use computer vision to identify regions that are most different from the rest of the video, or, score high on factors such as motion in the scene, or, contain important objects, and repeated occurrences that indicate the importance of this scene to the storyline of the video [Truong and Venkatesh 2007; Money and Agius 2008; Feng et al. 2012; Lu and Grauman 2012; Gygli et al. 2014].

Finding an algorithmic solution to the problem of predicting regions of high emotional arousal remains hard because the components

that go into creating an emotional response in a viewer are complex. Film-makers have identified devices as varied as color palette, theatrical setting, sound effects, actors' expressions and body language, and of course, dialogue and setup to be just a few of the factors that go into creating the desired *mis-en-scene*. Though there have been past efforts to find computational features that quantify these concepts [Rasheed and Shah 2002], we are still far from a good model of the emotional content of a scene.

Previous work has shown that viewers' reactions to videos are measurable via physiological indices such as heart rate variation, galvanic skin response, and pupillary dilation [Lang and Cuthbert 1997; Picard et al. 2001]. Thus, data recorded from viewers could be used as a proxy for computationally extracted features. Indeed, researchers in affective computing have proposed approaches to summarizing videos based on a variety of physiological indices that are correlated with emotional arousal. Among these, pupillary dilation stands out because it can be collected simultaneously with gaze position via an eyetracker, thus, allowing us to know both where a viewer was looking, and what her arousal level was at that time. However, it is an index that is impacted by brightness changes. This behavior is known as the pupillary light reflex [Loewenfeld 1958].

In this paper, we propose a first order method to factor out light reflex related diameter changes from measured pupillary data while viewers watched videos. We model the brightness induced pupil diameter change as a linear function of the grayscale intensity in the foveal neighborhood around the gaze position. The regression parameters are learned from a pre-capture calibration procedure, and the response time is modeled by a lag parameter. Assuming that light reflex and emotional arousal superimpose linearly to achieve the measured pupil diameter values, we hypothesize that the residual between the measured data and the model prediction is an index of emotional arousal. We compute an exciting-ness score for each frame of a video using eyetracking data collected from multiple viewers as they watched a video (see Figure 1). We show results on a variety of videos, and compare with ground truth in the form of self-reports by three independent coders.

2 Related Work

The main contributions of this paper are a model of pupillary light reflex, a calibration method to fit the parameters of this model, and a method to factor out light reflex related pupillary diameter changes when using the pupil as an index of emotional arousal in videos. This work is based on insights from psychophysiology, affective computing, and implicit methods for video annotation. In this section, we present a discussion of related work in these areas.

Physiological indices of emotional arousal An extensive body of work in the human psychophysiology of emotional arousal has explored data recorded from bodily, behavioral, and physiological systems during emotional challenge [Lang 1979]. This literature has converged to show that multiple measures such as heart rate, skin conductance, endocrine changes, etc. are useful in describing an observers multifaceted affective response [Lang and Cuthbert 1997; Picard et al. 2001]. These autonomic indices of emotional arousal have been studied in the context of visual stimuli such as photographs [Lang et al. 1993], and film clips [Kolodyazhniy et al. 2011]. Pupillary dilation in particular has been studied in various contexts ranging from opthalmology [Loewenfeld and Lowenstein 1993], to attention [Hoeks and Levelt 1993], and cognitive load [Hess and Polt 1964; Palinko et al. 2010], etc. Our focus is on pupillary diameter as a measure of interest, or, engagement in visual stimuli. Recently, Bradley and colleagues [2008; 2009] showed that pupil dilation exhibits a similar pattern to skin conductance: following a brief pupillary constriction associated with the light reflex, pupil diameter shows significantly greater dilation when viewing pleasant or unpleasant, compared to neutral, pictures. This effect is observable even on repeat trials, compared to skin response, which suffers from habituation.

These findings suggest that pupillary diameter has the potential to serve as the basis of an exciting-ness score on a video: when multiple viewers exhibit an increase in pupillary diameter, it would indicate that those portions of the video are more engaging. However, it is an index that is impacted by brightness changes, in addition to the emotional state of the viewer. This behavior is known as the pupillary light reflex [Loewenfeld 1958; Ellis 1981; Watson and Yellott 2012].Researchers have previously proposed models of pupil diameter change in response to light of varying intensity (for a review, see [Watson and Yellott 2012]) which are approximately linear in the luminance range spanned by a video playing on an LCD monitor. Our work is different from previous work in that previous models assume that the response of the pupil is instantaneous, but, in the context of videos, it is necessary to account for the time that it takes for the pupil to respond to a change in brightness, which we model as a lag parameter Δ.

Predicting the interesting/arousing regions in a video We focus on the line of work that uses data to implicitly annotate each frame of a video with its interestingness, for example, using similarity to highly rated photographs [Liu et al. 2009], and playback logs [Yu et al. 2003]. Concurrent with advances in personal wearable sensors, researchers are looking towards biosignals such as galvanic skin response, blood volume pulse, and facial EMG, to obtain an affective annotation of videos [Arapakis et al. 2009; Soleymani et al. 2014]. In particular, Katti and colleagues [2011] have proposed a method to convert pupil diameter measurements from multiple viewers into a score that represents the exciting-ness of a video. A key problem with this approach is that pupillary diameter is an index that is impacted by brightness changes, in addition to the emotional state of the viewer. Our work addresses this problem by proposing a model of light reflex that can be applied to factor out brightness induced changes. We show the application of our model to the method of Katti and colleagues [2011] for several videos.

3 Linear Model of Pupillary Light Reflex

One of the primary functions of the pupil is to regulate the amount of light entering the eye. It does this by constricting when faced with a brighter light intensity, and by dilating at lower intensity levels. At the same time, pupillary dilation also reflects emotional arousal. Research in psychophysiology has accounted for this dual function of the pupil by controlling for stimuli brightness, usually by ensuring that the average brightness across conditions is the same. Alternately, data collected on image stimuli is recorded for a long duration of time, and the initial few seconds are typically discarded during data analysis. Because the majority of this body of work has focused on images, which can be presented for as long as needed, these controls have worked. In the case of videos, however, each frame is presented for a duration dictated by the frame rate of the original video, and the brightness induced pupillary diameter changes cannot be similarly controlled for. Additionally, for real-world applications that use pupillary diameter as an index of arousal, what is needed is the measurement on a particular image or video frame rather than averages over all instances in a category.

We propose a first order method to factor out brightness induced diameter changes in the pupil by modeling the relationship between pupillary diameter and incident light intensity as the simplest possible linear model. Let the pupil diameter d at time instant t be proportional to the grayscale image intensity I in the foveal neighborhood of the gaze position. Because prior literature reports a tempo-

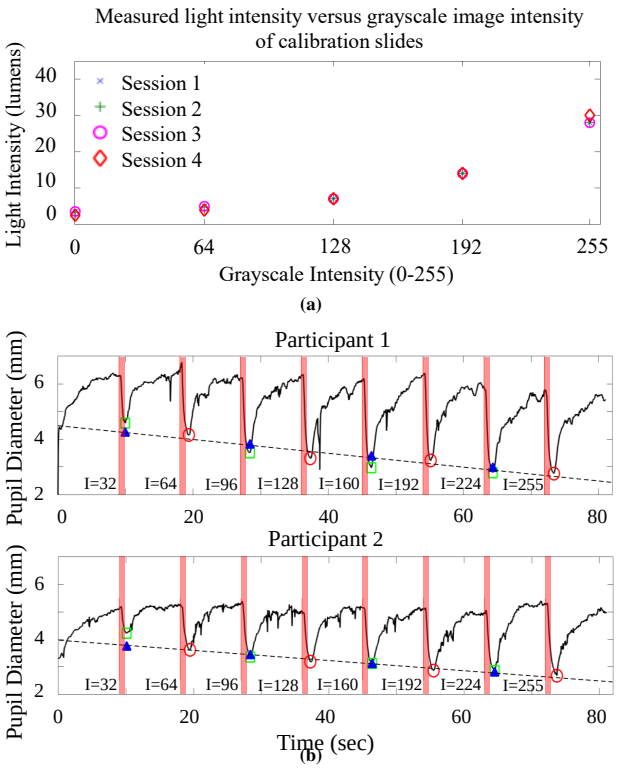

Figure 2: *(a) As a check, we measured the light intensity with a light meter as the calibration slides were presented to the participant. (b) The pupil diameter for two participants for the calibration procedure (blinks not removed). The red markers show the pupil diameter values that were used to fit the regression parameters. The green markers are the measured pupil diameters at test intensities, while the blue markers are the values predicted by the linear model.*

ral delay between intensity change and pupillary response [Bradley et al. 2008], we account for this delay in the form of a lag parameter Δ. Then,

$$d(t) = d_0 + k * I(t - \Delta), \quad (1)$$

where the parameters k and d_0 are learned in a least squares sense for each individual in a calibration procedure described below. Once these parameters have been fit to each individual, the model predicts what the measured pupillary diameter would be if the only underlying factor was brightness change.

Because our ultimate goal is to deploy pupillary response as an index of emotional arousal in real world settings, we make a further assumption. We assume that when pupillary measurements are recorded from viewers watching videos, the diameter changes due to light intensity, and emotional arousal sumperimpose linearly to achieve the actual recorded values. Once the regression parameters have been fit to each individual via the calibration procedure, the model prediction can be subtracted from the pupillary diameter values recorded on each video frame, and the residual would be proportional to that individual's emotional arousal.

Calibration procedure: Each participant is presented with a set of slides of increasing grayscale intensities. The slides are uniformly colored, with a small white fixation cross to center the participant's gaze. Previous literature has shown that it takes about one second for the pupil to constrict in response to brightness change [Bradley et al. 2008], and thus, each calibration slide is displayed for one second, followed by a black slide (grayscale intensity = 0) for eight seconds. As a check, we measured the incident light intensity via a Sekonic light meter at the head position of a test participant for each slide over multiple sessions (Figure 2a).

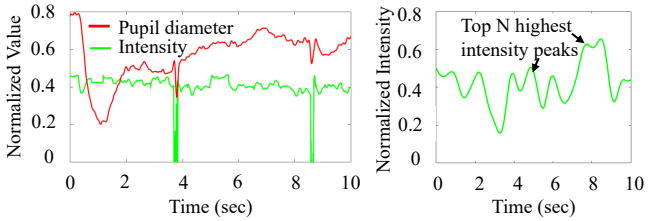

Figure 3: *Left: Pupil diameter for one participant (red). Average intensity in the foveal neighborhood (approx 2° visual angle) around the gaze position (green). Right: Illustration of high intensity peaks, which form ground truth test data for Section 5.*

Table 1: *We selected seven videos from YouTube as stimuli for the experiment. Details are provided below.*

Video	Theme	Duration(s)	Resolution(px)
Decay	Drama/Action	91	960 x 480
Bear	Adventure	84	1280 x 720
GoPro	Underwater	66	540 x 360
Lion	Wildlife	38	1280 x 720
Rhino	Wildlife	41	1280 x 720
Volcano	Explosion	63	1280 x 720
KidBreak	Entertainment	93	640 x 360

Predictive accuracy: Figure 2b illustrates the recorded data for two participants. We observe the pupil constrict to regulate the amount of light entering the eye. The minimum pupillary diameter for each calibration slide is observed to be inversely proportional to the grayscale intensity of the slide. As can be seen from this graph, our model does not incorporate the dynamics of the pupillary light response.We only model the maximum constriction, i.e., the minimum achieved diameter. The red circular markers (grayscale intensity $I = 64, 128, 192, 255$) show the pupil diameter values that were used to fit the regression parameters. The dotted line illustrates the fit for each individual. The green square markers are the measured pupil diameters at test intensities $I = 32, 96, 160, 224$, while the blue triangular markers are the values predicted by the linear model. We find that the model is a better fit at higher intensity levels than at lower intensity levels. The mean errors at the training points for the two participants were 0.1306mm (σ = 0.0974) and 0.0643mm (σ = 0.0033), respectively. The mean errors at the test points were 0.2938mm (σ = 0.1131) and 0.1665mm (σ = 0.2263), respectively.

4 Data Collection: Pupillary diameter measured on short videos

Eyetracking data was collected on videos that were long enough to contain a plot, but had a clear eliciting event, that is, a scene that could be expected to generate emotional arousal in a viewer. We selected seven video clips from YouTube that contained elements of surprise or drama (Table 1). The videos were displayed at their original resolution, centered on the screen, with a black background. The videos were shown in randomized order, and each video was preceded by a two second filler slide containing a white fixation cross centered on a black background. Prior to the stimuli videos, participants went through a pupillary light response calibration.

Participants were recruited in accordance with an IRB approved protocol from the university community, and compensated monetarily. In total, we recorded 10 viewers (5 male) using a remote eye tracker (SMI RED-m, 120 Hz). Participants were seated at a distance of approximately 64 cm from the screen (1680 × 1050 resolution, 18in×11in). A visual angle of 1 degree is approximately 37

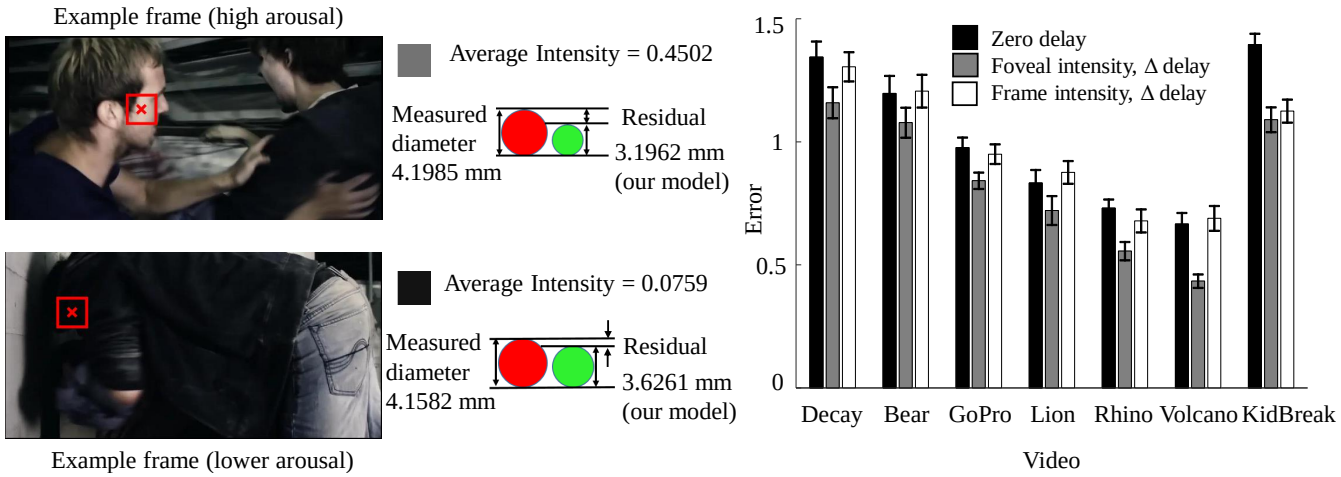

Figure 4: *Left: Illustration of measured diameter and model prediction on a brighter region and a darker region. Right: Mean predictive error on 10 brightest points in a video for all seven videos is shown. Error bars mark the standard error of the mean.*

pixels at these settings. After each video, participants were asked if they had seen the video before, and the responses were logged. Data collected consisted of gaze position, corresponding time stamp, and pupil diameter.

Samples where the gaze position was outside the screen were discarded. The recorded data was low-pass filtered (fifth order Butterworth) to remove noise. The cut off frequency was set to 10 Hz based on previous literature [Klingner et al. 2008]. We ensured that the input to the low-pass filter was uniformly sampled by reconstructing missing samples using linear interpolation. The first twelve seconds of data was removed to account for the initial light reflex due to the black fixation cross slide shown prior to the movie. Since the pupil is dilated before the start of the movie, the sudden increase in intensity caused the pupil diameter of all participants to have a sharp dip in pupil size, at the beginning of the video (Figure 3 (Left)), as it adjusts to the video's higher intensity. Because our method relies on finding the mean pupil diameter across the video, we remove this portion of the data.

5 How well does the model of light reflex apply to video data?

Although Figure 2b tests the predictive accuracy of the fitted linear model, it is a simplified case. In particular, it does not incorporate the effect of considering only the foveal neighborhood versus the entire video frame, or, the effect of considering the temporal lag between stimulus onset, and associated response. We evaluate the utility of both of these choices in our model. We consider the top 10 brightest regions across the duration of a video. We assume that in these regions, pupillary constriction is solely governed by light reflex. Therefore, in these 10 regions, measured pupil diameter is considered to be ground truth for the purpose of evaluating our model. Figure 3 (Right) illustrates how test points were selected. The green curve shows an excerpt from one of our videos. The vertical axis marks the average grayscale intensity in the foveal neighborhood around the gaze position of a participant watching this video. The ten highest intensity points in this graph are identified as ground truth regions.

In the next two subsections, we evaluate the utility of each of the two choices: temporal lag, and foveal neighborhood. When evaluating temporal lag, we incorporate the foveal neighborhood in our computations. When evaluating the effect of foveal neighborhood, we keep the temporal lag constant.

5.1 Utility of temporal lag

We compare the prediction of our model when it incorporates a temporal lag term Δ, against the case where this term is set to zero. The error in the zero lag case is shown in Figure 4 (Right) as black bars. This error is the average absolute difference between the model prediction and the ground truth value, averaged across all 10 ground truth points, and all participants (as explained previously).

We now describe how we estimate the temporal lag parameter for each individual participant on a given video. While this parameter could be computed from the calibration data (Figure 2b), we explicitly tune the lag to the particular video that we are operating on. Let us denote the identified top 10 highest intensity points in the course of the video as "brightness events". In the 10 test regions, each brightness event is followed by a minimum point in the pupillary diameter curve immediately following this event. Let us call this minimum point a "constriction event". The value of Δ that best fits the temporal distance between the brightness event and the constriction event is the temporal lag parameter we select. We hand-tune Δ for each participant and each video, though this procedure could be reasonably done automatically in the future.

Using this value of Δ, the error between the model prediction and ground truth is computed. The average error across all participants on a given video is shown in gray in Figure 4 (Right). For all videos, the gray bars are shorter than the black bars, showing that the prediction error is reduced when the model incorporates temporal lag.

5.2 Utility of foveal neighborhood

Here we compare the prediction of our model when it incorporates foveal neighborhood while computing incident intensity versus the case where the model simply uses the average intensity over the entire video frame. The error in the entire frame case is shown in Figure 4 (Right) as white bars. This error is the average absolute difference between the model prediction and the ground truth value, averaged across all 10 ground truth points, and all participants.

The foveal region is defined as the 2° visual angle around the gaze point. In our experimental setup, this corresponds to a circular region approximately 74 pixels in diameter centered at the gaze point. For simplicity, we take the foveal neighborhood to be a square of size 74×74 pixels centered at the gaze point. The input intensity to our model is the average grayscale intensity in this square region. For example, in Figure 4 (Left), the average intensities in the foveal neighborhood in the two selected frames are shown.

The test points are the same as in the temporal lag evaluation. Note that for the purpose of evaluating the utility of foveal neighborhood, we keep the temporal lag unchanged between the white and gray bars in Figure 4 (Right). The results indicate that prediction error is reduced when the model considers the region around when the viewer was actually looking, rather than merely averaging the intensity of the entire video frame. This suggests that incorporating foveal neighborhood in our model is a reasonable choice.

6 Does the residual provide a more accurate index of emotional arousal?

In the previous sections, we have described a linear model of brightness induced pupillary diameter change. We have tested its predictive accuracy in a controlled test that mimics the calibration procedure. In order to apply this model to factor out the effect of light response from pupil diameter data recorded on real world videos, we tested the performance of the model at points where we expected the light reflex to dominate other sources of pupillary dilation, i.e., points in the video where the grayscale intensity of the attended pixels is brightest. Now, we evaluate the model from an application point of view. The intuition is the following: because pupillary diameter is known to increase when the viewer is aroused, does factoring out light reflex change the regions of the video that are tagged as being exciting to viewers?

Katti and colleagues [2011] have previously proposed a binning based method to convert pupillary diameter measurements from multiple viewers into a per-frame emotional arousal score. We computed an arousal score for our data using their method, including the same parameter values reported by them: for a given video, each participant is considered in turn. The mean and standard deviation of measured pupil diameters across the entire duration of the video are computed. Then, the unsigned distance of the pupil diameter to the mean (in units of standard deviation) is computed for each frame of the video. This distance is smoothed via a moving average filter (window size = 4 seconds), and then, is binned in steps of 1. The binned values are a per-frame score of emotional arousal for this participant. We evaluate the impact of our model of pupillary light reflex by subtracting the model prediction for each frame from the measured pupil diameter value. The underlying assumption is that light reflex and emotional arousal superimpose linearly, and factors such as age, cognitive effort, etc. are negligible in this setting.

Figure 5 shows the measured pupil diameter and residual, and the corresponding binned values for one participant, and we see that the binned values which index the emotional arousal of the viewer, or, alternately, the exciting-ness of the video, are different in the two cases. Figure 6 shows the scores computed with light reflex factored out as the lighter green line, and the scores without this as the darker red line for participants on the video *Decay*. The regions marked with a black bounding box illustrate the impact of removing light response using the proposed model. We observe that the model performs as expected intuitively: for a darker scene, the model predicts that light reflex will cause the viewer's pupil to dilate, and thus, the portion of pupillary dilation that can be attributed to emotional arousal is less. This is why the green curve is below the red curve in the second dotted rectangle in Figure 6.

Figure 1 shows the average score across all participants for the video *Decay*, after applying the moving average filter. The dark blue line represents the score computed from the raw pupil diameter measurements, while the lighter pink line represents the score computed after our model of light reflex has been applied. In Scene 1 and Scene 3, the residual after factoring out light response causes an increased deviation from the mean, resulting in a high score. In particular, in the region marked Scene 3, the agitated woman is

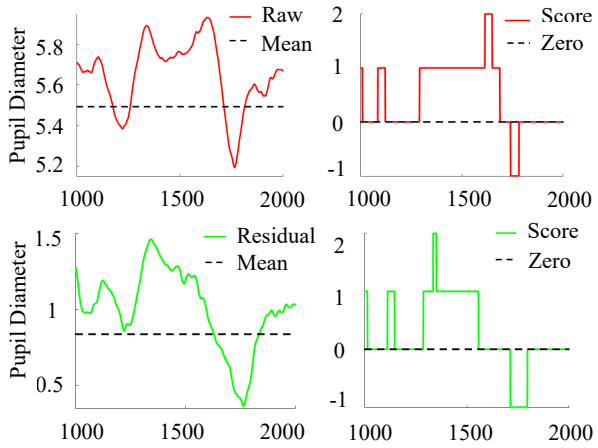

Figure 5: *Left: The red line represents the raw pupil diameter, and the green line represents the residual between raw and predicted pupil diameter for one participant, for the video Decay. Right: The red line represents the binned vector computed from raw data. The green line shows the binned vector computed from data after our model is applied.*

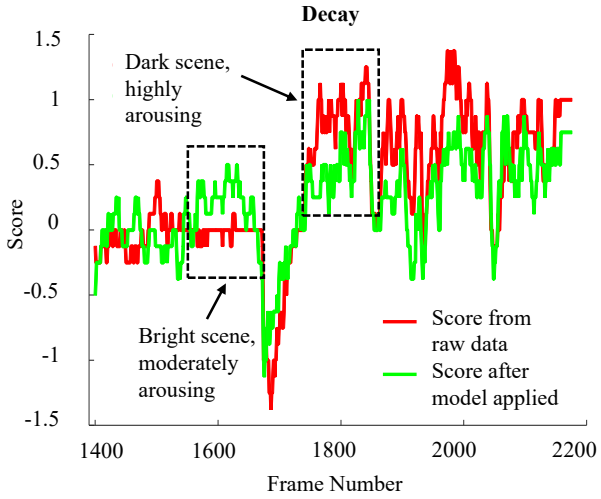

Figure 6: *The red line represents the binned vector computed from for raw data. The green line represents the binned vector computed after our model is applied. Results are across all participants, for the video Decay.*

chiding her friend for wanting to leave behind an injured member of their party. The brightness of the scene caused an index based on raw pupillary response to incorrectly score it low on arousal. Our method of factoring out light response fixed this. Towards the end of the video excerpt (Scene 4), the injured friend turns into a zombie and attacks everybody. This scene is relatively darker, and the model correctly subtracted out the portion of pupillary dilation that can be attributed to the low grayscale intensity of the frames, but the rising behavior due to emotional arousal has been retained.

Annotations by independent coders We asked three independent coders to annotate our seven example videos by marking out the start and end points of two regions where the viewer would "sit up and take notice". They were asked to rank these regions in order of intensity. These regions are shown as colored rectangles in Figure 7, where red marks the region reported as most exciting, and blue marks the region reported second. We observed that the consistency across coders depends on the content of the video. For example, the coders are much more consistent for *Decay* (Figure 1),

as compared to *Lion* (Figure 7c), and *Kid Break* (Figure 7f). A possible reason for the differing consistency is that the most exciting events in *Decay* and *Lion* are quite evident (lion growls, and zombie attacks). In contrast, in *Kid Break*, one of the coders marked the set up as exciting, while another marked the actual event, while another marked the replay of the actual event. Despite this variability, we show these annotations to illustrate where self-reports matched, and where they differed from a physiologically-based index.

Figure 7a corresponds to *Bear Attack*, a first person video by a man cycling away from a charging bear. Most of the video shows the trail and the trees zipping past. In this example, video cues, such as, optical flow and color histogramming, which have previously been used to predict exciting-ness, would not work well. Thus, this video is a candidate where a physiological data based score would be very useful. The dark blue curve shows that pupillary data collected from viewers indeed captures the rise in exciting-ness as the cyclist's path is blocked by a fallen tree. The lighter pink curve shows that incorporating our model into the score computation corrects for light response in regions that are incorrectly scored as more (and less) exciting than perhaps they should have been (third dotted rectangle). In the first dotted rectangle, viewers are looking at the relatively darker trees, and our model removes the spike, but this is a region between two consecutive bear sightings and perhaps for this reason was marked by one of the coders as an exciting region. Thus, we have marked it a false negative.

Figure 7b shows the results for the video *GoPro*, where a GoPro camera is mounted under water. A fish approaches the camera, and repeatedly attacks it. The first strike at the camera is captured by the computed score both for the raw data, and after our model is applied. Figure 7c is a video of a lion sitting by the side of the road while visitors at a wildlife sanctuary photograph it. The lion suddenly gets up and growls. The dotted rectangle shows the spike in exciting-ness retained after our model is applied.Figure 7d shows the results for a video where a rhino charges towards a car. The rhino initially walks around, and then, positions itself right in front of the car. Around the first dotted rectangle, the rhino charges towards the car, and the people in the car are screaming. The exciting-ness score in this region (driven by the participants' pupillary diameters increasing) is slightly accentuated after the removal of light response. Around frames 1000-1060, the people in the car are trying to escape from the rhino, and the camera is shaken a lot, mostly pointing towards the sun. Incorporating our model causes the score to increase in this region.

Figure 7e shows the results for the video *Volcano*. Here, removal of the light response allows us to capture the first explosion of the volcano (first dotted rectangle). Soon after the explosion, a man warns against a shock about to happen, and a loud sound is heard. Most of the spikes in this region are either consistent with the score from raw pupil diameter, or have been heightened (second and third dotted rectangles). Figure 7f corresponds to a YouTube video of a young child cracking a wine glass by making a high pitched noise. Around the first dotted rectangle, the glass actually breaks. This spike in exciting-ness is retained even after light reflex is factored out. We observe that the model appropriately removed the components of pupil diameter increase caused by light reflex in the region marked by the second dotted rectangle (frames 1540-1780). These frames correspond to a scene where the child shows his mother the broken glass, i.e., the main event of this video has already occurred.

7 Conclusion and Discussion

Pupil diameter measurements are a useful physiological index of emotional arousal as they can be collected concurrently with eye-tracking data, thus, enabling a rich implicit annotation of images and videos: where are people looking, and what are they engaged by. A problem with this measurement is that it is impacted by brightness changes, known as the pupillary light reflex, in addition to the emotional state of the viewer. In situations where the impact of brightness can be factored out by averaging responses across multiple images or videos in a category, the impact of pupillary light reflex on can be controlled for. However, for this index to be applicable in real world applications, such as, measuring arousal while viewers watch movies, and sports games, or, when students interact with tutoring systems, or, when trainees use simulators, then, we need a way to factor out the impact of light reflex.

A key assumption in the proposed model is that the pupil's response to brightness changes is linear. This assumption only holds for a limited range of incident light intensities; there is a physical maximum and minimum diameter for a person's pupil, and so the response must become non-linear at some point. The linear assumption is adequate for our application (videos), but does not cover the full range of pupillary behavior [Watson and Yellott 2012]. A useful direction of future work would be to propose and evaluate more sophisticated non-linear models in the context of videos, similar to the models proposed by Pamplona and colleagues [2009]. Another assumption is that the pupillary constriction measured in a "one-second-on" calibration procedure adequately models the brightness induced pupillary constriction when a video is playing (typically at 30fps). We also assume that the response of the pupil, at each time instant, is independent of the previous light intensity that it was exposed to. Future work could address this assumption with a more sophisticated calibration procedure which randomizes the duration and order of presentation of the fixed intensity slides, and a model that accounts for history.

Though environmental lighting likely affects pupillary light response, in addition to incident light intensity, in our experiment, ambient lighting conditions were kept constant between the calibration procedure and data collection. Similarly, factors such as age, accommodation, and cognitive activity were assumed to be constant between calibration and data collection. Previous research has shown that increased cognitive effort is correlated with a constriction of the pupil [Marshall 2002; Bailey and Iqbal 2008]. The interaction between cognitive constriction, and emotional dilation is not well understood, and most research assumes that only one of these two factors is at play at a given time. Because our videos were simple to understand, and the viewers were asked to watch them without an explicit task, we have assumed that their cognitive effort remained constant throughout the data collection, and would thus be factored out with the baseline.

Previous research has pointed out that the pupil looks elliptical to the eyetracker when viewed at off-axis positions, and this could lead to errors in measuring diameter [Mathur et al. 2013]. We did a sanity check where we measured pupillary diameter on a 30fps video which displayed a white fixation cross on a gray background (grayscale value = 128) at five different positions on the screen, for two seconds each. We observed that the average pupil diameter at each fixation cross location was approximately the same. This is a preliminary finding however, and it would be a valuable experiment to repeat this for multiple participants' data.

Our example videos are short, and are centered around one, or, two clearly exciting events. As part of evaluation, we asked independent coders to mark out the start and end points of the two most exciting regions of each video. However, we found that while there was agreement amongst the coders on the first most exciting event, there was large variance in the second most exciting event. This could be because the videos were too short to contain a second, equally clearly exciting event. An informative next step would be to work with longer videos with a more complex narrative.

It would be also interesting to cross-validate the score computed from pupillary measurements with a different physiological index, such as, heart rate variability, or, galvanic skin response. Bradley and colleagues [2008] measured pupil diameter, heart rate and galvanic skin response simultaneously on emotionally arousing pictures. Their results show that pupil's response during affective picture viewing reflects emotional arousal associated with increased sympathetic activity. They also observed greater skin conductance and greater cardiac deceleration during affective picture viewing. However, the timescales of the two physiological indices were not equivalent. In future, it may be interesting to reverify these results for emotionally arousing videos.

References

ARAPAKIS, I., KONSTAS, I., AND JOSE, J. M. 2009. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *ACM International Conference on Multimedia (MM)*, 461–470.

BAILEY, B. P., AND IQBAL, S. T. 2008. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI) 14*, 4, 21.

BRADLEY, M. M., MICCOLI, L., ESCRIG, M. A., AND LANG, P. J. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology 45*, 4, 602–607.

BRADLEY, M. M. 2009. Natural selective attention: orienting and emotion. *Psychophysiology 46*, 1–11.

ELLIS, C. J. 1981. The pupillary light reflex in normal subjects. *British Journal of Ophthalmology 65*, 11, 754–759.

FENG, S., LEI, Z., YI, D., AND LI, S. 2012. Online content-aware video condensation. In *Computer Vision and Pattern Recognition (CVPR)*.

GYGLI, M., GRABNER, H., RIEMENSCHNEIDER, H., AND VAN GOOL, L. 2014. Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*.

HESS, E. H., AND POLT, J. M. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science 143*, 3611, 1190–1192.

HOEKS, B., AND LEVELT, W. J. 1993. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers 25*, 1, 16–26.

KATTI, H., YADATI, K., KANKANHALLI, M., AND TAT-SENG, C. 2011. Affective video summarization and story board generation using pupillary dilation and eye gaze. In *IEEE International Symposium on Multimedia*.

KLINGNER, J., KUMAR, R., AND HANRAHAN, P. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ACM, New York, NY, USA, ETRA '08, 69–72.

KOLODYAZHNIY, V., KREIBIG, S. D., GROSS, J. J., ROTH, W. T., AND WILHELM, F. H. 2011. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology 48*, 7, 908–922.

LANG, P.J., B. M., AND CUTHBERT, B., 1997. Motivated attention: Affect, activation, and action.

LANG, P., GREENWALD, M., BRADLEY, M., AND HAMM, A. 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology 30*, 261–273.

LANG, P. J. 1979. A bioinformational theory of emotional imagery. *Psychophysiology 16*, 495–512.

LIU, F., NIU, Y., AND GLEICHER, M. 2009. Using web images for measuring video frame interestingness. In *Twenty-first International Joint Conference on Artificial Intelligence (IJCAI 2009)*.

LOEWENFELD, I. E., AND LOWENSTEIN, O. 1993. *The pupil: anatomy, physiology, and clinical applications*, vol. 2. Wiley-Blackwell.

LOEWENFELD, I. E. 1958. Mechanisms of reflex dilatation of the pupil. *Documenta Ophthalmologica 12*, 1, 185–448.

LU, Z., AND GRAUMAN, K. 2012. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition (CVPR)*.

MARSHALL, S. P. 2002. The index of cognitive activity: Measuring cognitive workload. In *Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on*, IEEE, 7–5.

MATHUR, A., GEHRMANN, J., AND ATCHISON, D. A. 2013. Pupil shape as viewed along the horizontal visual field. *Journal of vision 13*, 6, 3–3.

MONEY, A. G., AND AGIUS, H. 2008. Video summarization: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation 19*, 2, 121–143.

PALINKO, O., KUN, A. L., SHYROKOV, A., AND HEEMAN, P. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Symposium on Eye Tracking Research & Applications (ETRA)*, 141–144.

PAMPLONA, V. F., OLIVEIRA, M. M., AND BARANOSKI, G. V. 2009. Photorealistic models for pupil light reflex and iridal pattern deformation. *ACM Transactions on Graphics (TOG) 28*, 4, 106.

PICARD, R., VYZAS, E., AND HEALEY, J. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23*, 10.

RASHEED, Z., AND SHAH, M. 2002. Movie genre classification by exploiting audio-visual features of previews. In *International Conference on Pattern Recognition*, vol. 2, 1086–1089.

SOLEYMANI, M., LARSON, M., PUN, T., AND HANJALIC, A. 2014. Corpus development for affective video indexing. *IEEE Transactions on Multimedia 16*, 4, 1075–1089.

TRUONG, B. T., AND VENKATESH, S. 2007. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 3*, 1.

WATSON, A. B., AND YELLOTT, J. I. 2012. A unified formula for light-adapted pupil size. *Journal of Vision 12*, 10, 12–12.

YU, B., MA, W.-Y., NAHRSTEDT, K., AND ZHANG, H.-J. 2003. Video summarization based on user log enhanced link analysis. In *ACM International Conference on Multimedia*.

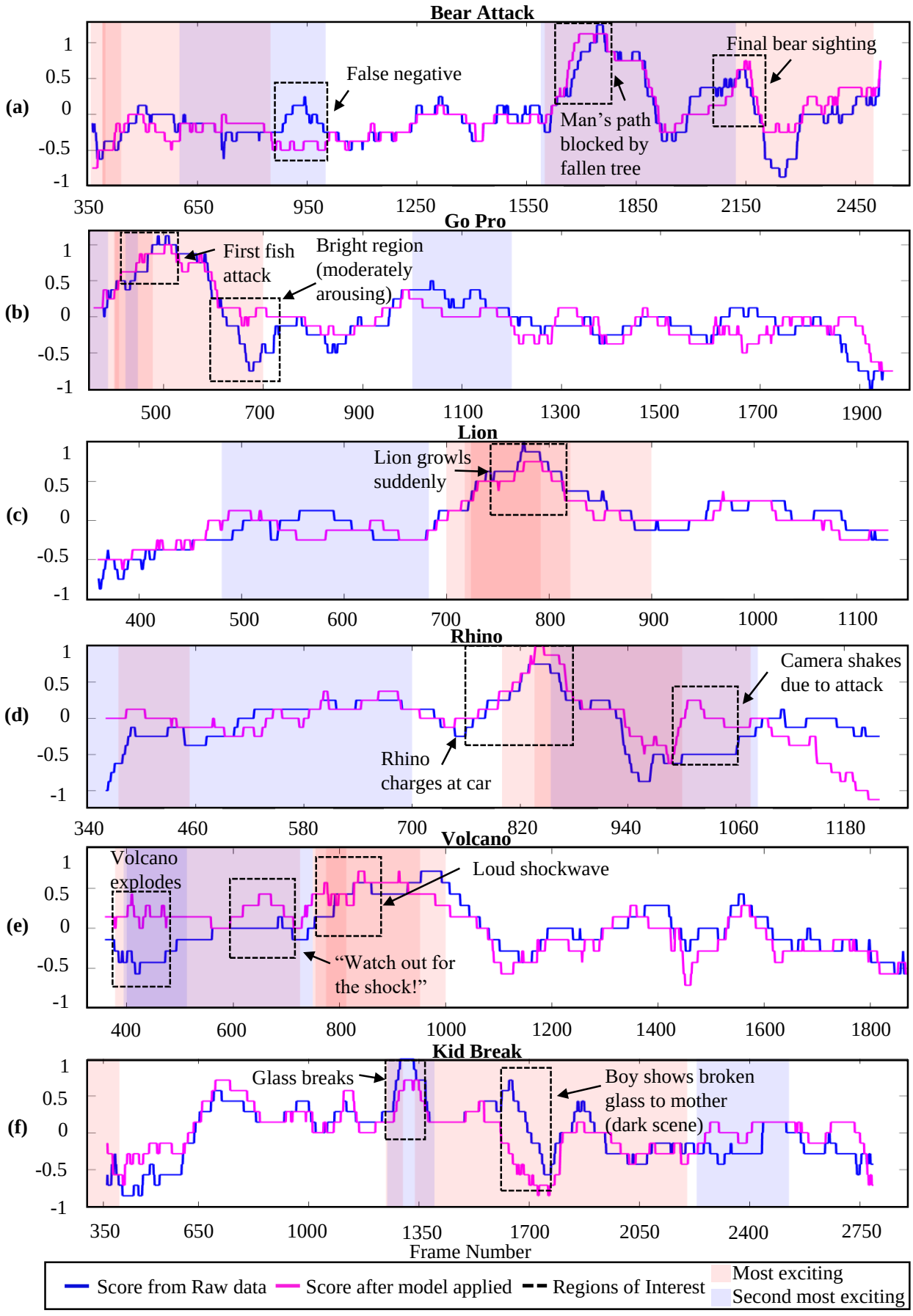

Figure 7: *We show results on a variety of example videos.*