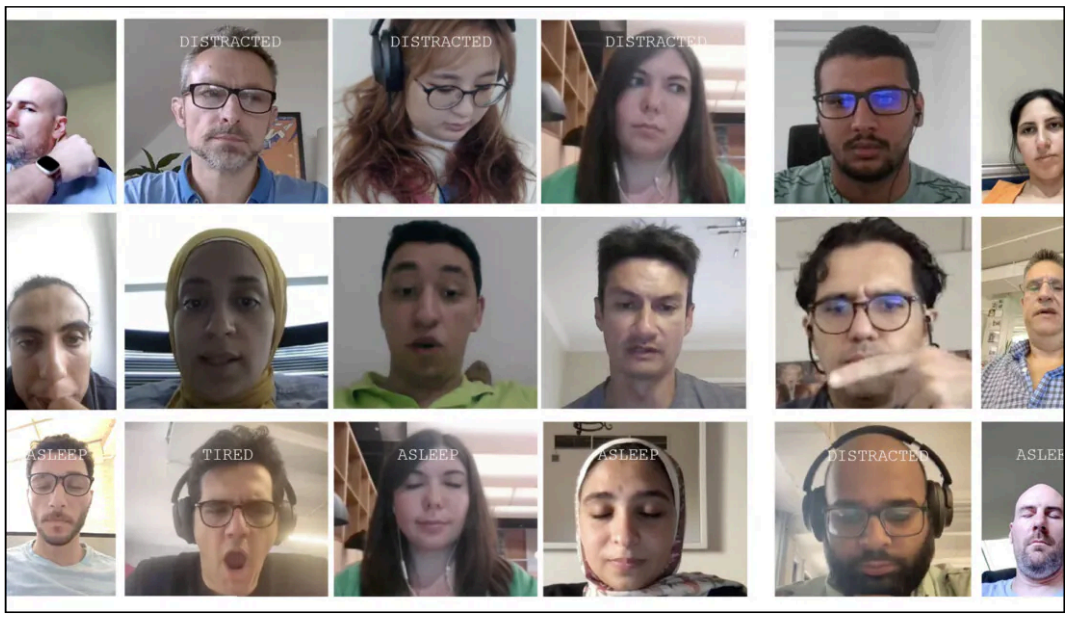
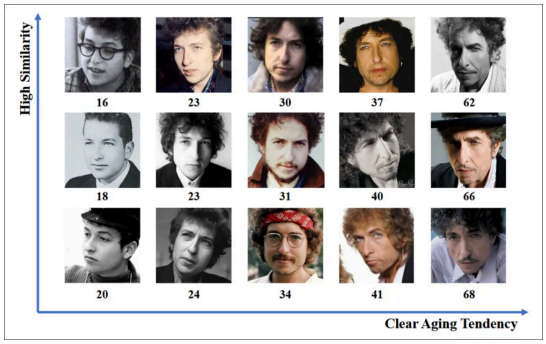
Looking for ‘Owls and Lizards’ in an Advertiser’s Audience

Originally published at https://www.unite.ai/looking-for-owls-and-lizards-in-an-advertisers-audience/



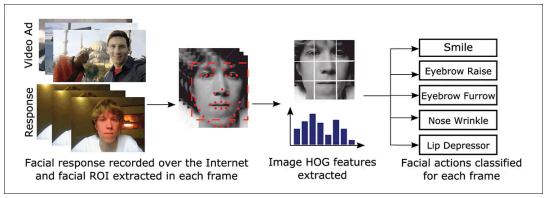
Since the online advertising sector is estimated to have spent $740.3 billion USD in 2023, it’s easy to understand why advertising companies invest considerable resources into this particular strand of computer vision research.

Though insular and protective, the industry occasionally publishes studies that hint at more advanced proprietary work in facial and eye-gaze recognition – including age recognition, central to demographic analytics statistics:



Estimating age in an in-the-wild advertising context is of interest to advertisers who may be targeting a particular age demographic. In this experimental example of automatic facial age estimation, the age of performer Bob Dylan is tracked across the years. Source: https://arxiv.org/pdf/1906.03625

These studies, which seldom appear in public repositories such as Arxiv, use legitimately-recruited participants as the basis for AI-driven analysis that aims to determine to what extent, and in what way, the viewer is engaging with an advertisement.

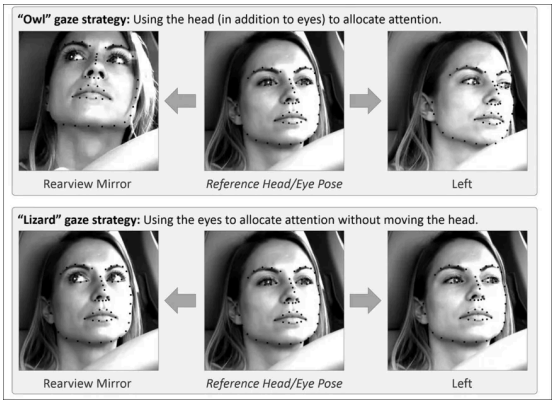


Dlib’s Histogram of Oriented Gradients (HoG) is often used in facial estimation systems. Source: https://www.computer.org/csdl/journal/ta/2017/02/07475863/13rRUNvyarN

Animal Instinct

In this regard, naturally, the advertising industry is interested in determining false positives (occasions where an analytical system misinterprets a subject’s actions), and in establishing clear criteria for when the person watching their commercials is not fully engaging with the content.

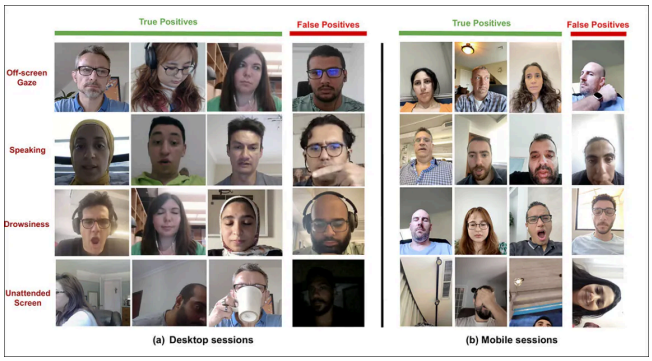
As far as screen-based advertising is concerned, studies tend to focus on two problems across two environments. The environments are 'desktop' or 'mobile', each of which has particular characteristics that need bespoke tracking solutions; and the problems – from the advertiser's standpoint – are represented by *owl behavior and lizard behavior* – the tendency of viewers to not pay full attention to an ad that is in front of them.



Examples of 'Owl' and 'Lizard' behavior in a subject of an advertising research project. Source: https://arxiv.org/pdf/1508.04028

If you're looking *away* from the intended advertisement with your whole head, this is 'owl' behavior; if your head pose is static but your eyes are *wandering away* from the screen, this is 'lizard' behavior. In terms of analytics and testing of new advertisements under controlled conditions, these are essential actions for a system to be able to capture.

A new paper from SmartEye's Affectiva acquisition addresses these issues, offering an architecture that leverages several existing frameworks to provide a combined and concatenated feature set across all the requisite conditions and possible reactions – and to be able to tell if a viewer is bored, engaged, or in some way remote from content that the advertiser wishes them to watch.



Examples of true and false positives detected by the new attention system for various distraction signals, shown separately for desktop and mobile devices. Source: https://arxiv.org/pdf/2504.06237

The authors state*:

> *'Limited research has delved into monitoring attention during online ads. While these studies focused on estimating head pose or gaze direction to identify instances of diverted gaze, they disregard critical parameters such as device type (desktop or mobile), camera placement relative to the screen, and screen size. These factors significantly influence attention detection.*
>
> *'In this paper, we propose an architecture for attention detection that encompasses detecting various distractors, including both the owl and lizard behavior of gazing off-screen, speaking, drowsiness (through yawning and prolonged eye closure), and leaving screen unattended.*
>
> *'Unlike previous approaches, our method integrates device-specific features such as device type, camera placement, screen size (for desktops), and camera orientation (for mobile devices) with the raw gaze estimation to enhance attention detection accuracy.'*

The new work is titled *Monitoring Viewer Attention During Online Ads*, and comes from four researchers at Affectiva.

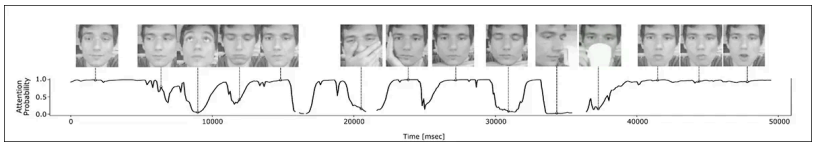
Method and Data

Largely due to the secrecy and closed-source nature of such systems, the new paper does not compare the authors' approach directly with rivals, but rather presents its findings exclusively as ablation studies; neither does the paper adhere in general to the usual format of Computer Vision literature. Therefore, we'll take a look at the research as it is presented.

The authors emphasize that only a limited number of studies have addressed attention detection specifically in the context of online ads. In the AFFDEX SDK, which offers real-time multi-face recognition, attention is inferred solely from head pose, with participants labeled inattentive if their head angle passes a defined threshold.

An example from the AFFDEX SDK, an Affectiva system which relies on head pose as an indicator of attention. Source: https://www.youtube.com/watch?v=c2CWb5jHmbY

In the 2019 collaboration *Automatic Measurement of Visual Attention to Video Content using Deep Learning*, a dataset of around 28,000 participants was annotated for various inattentive behaviors, including *gazing away*, *closing eyes*, or engaging in *unrelated activities*, and a CNN-LSTM model trained to detect attention from facial appearance over time.

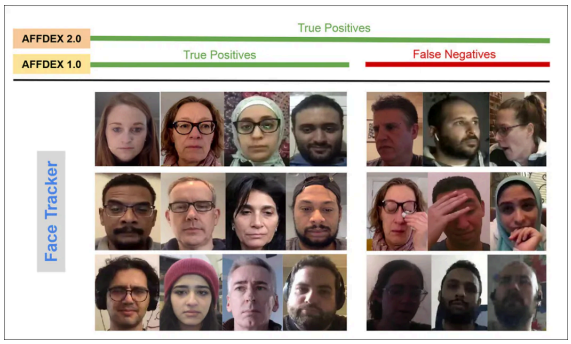


From the 2019 paper, an example illustrating predicted attention states for a viewer watching video content. Source: https://www.jeffcohn.net/wp-content/uploads/2

However, the authors observe, these earlier efforts did not account for device-specific factors, such as whether the participant was using a desktop or mobile device; nor did they consider screen size or camera placement. Additionally, the AFFDEX system focuses only on identifying gaze diversion, and omits other sources of distraction, while the 2019 work attempts to detect a broader set of behaviors – but its use of a single shallow CNN may, the paper states, have been inadequate for this task.

The authors observe that some of the most popular research in this line is not optimized for ad testing, which has different needs compared to domains such as driving or education – where camera placement and calibration are usually fixed in advance, relying instead on uncalibrated setups, and operating within the limited gaze range of desktop and mobile devices.

Therefore they have devised an architecture for detecting viewer attention during online ads, leveraging two commercial toolkits: AFFDEX 2.0 and SmartEye SDK.



Examples of facial analysis from AFFDEX 2.0. Source: https://arxiv.org/pdf/2202.12059

These prior works extract low-level features such as facial expressions, head pose, and gaze direction. These features are then processed to produce higher-level indicators, including gaze position on the screen; yawning; and speaking.

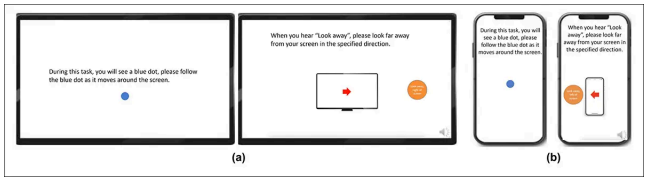
The system identifies four distraction types: *off-screen gaze*; *drowsiness*,; *speaking*; and *unattended screens*. It also adjusts gaze analysis according to whether the viewer is on a desktop or mobile device.

Datasets: Gaze

The authors used four datasets to power and evaluate the attention-detection system: three focusing individually on gaze behavior, speaking, and yawning; and a fourth drawn from real-world ad-testing sessions containing a mixture of distraction types.

Due to the specific requirements of the work, custom datasets were created for each of these categories. All the datasets curated were sourced from a proprietary repository featuring millions of recorded sessions of participants watching ads in home or workplace environments, using a web-based setup, with informed consent – and due to the limitations of those consent agreements, the authors state that the datasets for the new work cannot be made publicly available.

To construct the *gaze* dataset, participants were asked to follow a moving dot across various points on the screen, including its edges, and then to look away from the screen in four directions (up, down, left, and right) with the sequence repeated three times. In this way, the relationship between capture and coverage was established:



Screenshots showing the gaze video stimulus on (a) desktop and (b) mobile devices. The first and third frames display instructions to follow a moving dot, while the second and fourth prompt participants to look away from the screen.

The moving-dot segments were labeled as *attentive*, and the off-screen segments as *inattentive*, producing a labeled dataset of both positive and negative examples.

Each video lasted approximately 160 seconds, with separate versions created for desktop and mobile platforms, each with resolutions of 1920×1080 and 608×1080, respectively.

A total of 609 videos were collected, comprising 322 desktop and 287 mobile recordings. Labels were applied automatically based on the video content, and the dataset split into 158 training samples and 451 for testing.

Datasets: Speaking

In this context, one of the criteria defining 'inattention' is when a person speaks for *longer than one second* (which case could be a momentary comment, or even a cough).

Since the controlled environment does not record or analyze audio, speech is inferred by observing inner movement of estimated facial landmarks. Therefore to detect *speaking* without audio, the authors created a dataset based entirely on visual input, drawn from their internal repository, and divided into two parts: the first of these contained approximately 5,500 videos, each manually labeled by three annotators as either speaking or not speaking (of these, 4,400 were used for training and validation, and 1,100 for testing).

The second comprised 16,000 sessions automatically labeled based on session type: 10,500 feature participants silently watching ads, and 5,500 show participants expressing opinions about brands.

Datasets: Yawning

While some 'yawning' datasets exist, including YawDD and Driver Fatigue, the authors assert that none are suitable for ad-testing scenarios, since they either feature *simulated* yawns or else contain facial contortions that could be confused with *fear,* or other, non-yawning actions.

Therefore the authors used 735 videos from their internal collection, choosing sessions likely to contain a *jaw drop* lasting more than one second. Each video was manually labeled by three annotators as either showing *active* or *inactive yawning*. Only 2.6 percent of frames contained active yawns, underscoring the class imbalance, and the dataset was split into 670 training videos and 65 for testing.

Datasets: Distraction

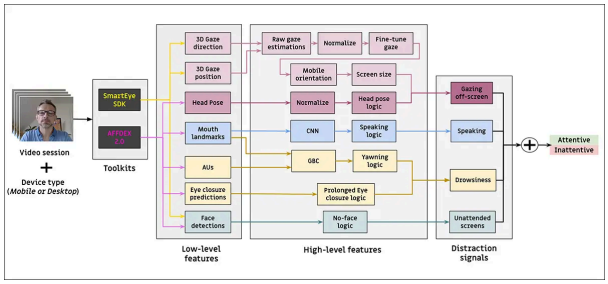
The *distraction* dataset was also drawn from the authors' ad-testing repository, where participants had viewed actual advertisements with no assigned tasks. A total of 520 sessions (193 on mobile and 327 on desktop environments) were randomly selected and manually labeled by three annotators as either *attentive* or *inattentive*.

Inattentive behavior included *off-screen gaze*, *speaking*, *drowsiness*, and *unattended screens*. The sessions span diverse regions across the world, with desktop recordings more common, due to flexible webcam placement.

Attention Models

The proposed attention model processes low-level visual features, namely facial expressions; head pose; and gaze direction – extracted through the aforementioned AFFDEX 2.0 and SmartEye SDK.

These are then converted into high-level indicators, with each distractor handled by a separate binary classifier trained on its own dataset for independent optimization and evaluation.



Schema for the proposed monitoring system.

The *gaze* model determines whether the viewer is looking at or away from the screen using normalized gaze coordinates, with separate calibration for desktop and mobile devices. Aiding this process is a linear Support Vector Machine (SVM), trained on spatial and temporal features, which incorporates a memory window to smooth rapid gaze shifts.

To detect *speaking without audio*, the system used cropped mouth regions and a 3D-CNN trained on both conversational and non-conversational video segments. Labels were assigned based on session type, with temporal smoothing reducing the false positives that can result from brief mouth movements.

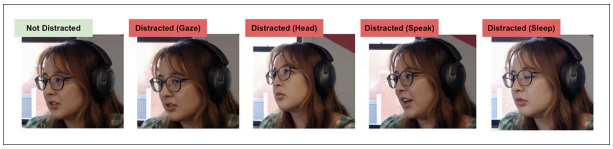
Yawning was detected using full-face image crops, to capture broader facial motion, with a 3D-CNN trained on manually labeled frames (though the task was complicated by yawning's low frequency in natural viewing, and by its similarity to other expressions).

Screen abandonment was identified through the absence of a face or extreme head pose, with predictions made by a decision tree.

Final attention status was determined using a fixed rule: if any module detected inattention, the viewer was marked *inattentive* – an approach prioritizing sensitivity, and tuned separately for desktop and mobile contexts.

Tests

As mentioned earlier, the tests follow an ablative method, where components are removed and the effect on the outcome noted.



Different categories of perceived inattention identified in the study.

The gaze model identified off-screen behavior through three key steps: normalizing raw gaze estimates, fine-tuning the output, and estimating screen size for desktop devices.

To understand the importance of each component, the authors removed them individually and evaluated performance on 226 desktop and 225 mobile videos drawn from two datasets. Results, measured by G-mean and F1 scores, are shown below:

Gaze model	Real Distraction Dataset				Gazing Dataset			
	Desktop		Mobile		Desktop		Mobile	
	G-mean	F1	G-mean	F1	G-mean	F1	G-mean	F1
w/o normalization	0.678	0.330	0.795	0.560	0.656	0.523	0.740	0.595
w/o fine-tuning model	0.639	0.443	0.802	0.651	0.659	0.557	0.751	0.645
w/o screen size detection (desktop only)	0.726	0.479	–	–	0.682	0.557	–	–
Eye gaze (full model)	**0.740**	**0.504**	**0.807**	**0.628**	**0.685**	**0.563**	**0.761**	**0.647**

Results indicating the performance of the full gaze model, alongside versions with individual processing steps removed.

In every case, performance declined when a step was omitted. Normalization proved especially valuable on desktops, where camera placement varies more than on mobile devices.

The study also assessed how visual features predicted mobile camera orientation: face location, head pose, and eye gaze scored 0.75, 0.74, and 0.60, while their combination reached 0.91, highlighting – the authors state – the advantage of integrating multiple cues.

The *speaking* model, trained on vertical lip distance, achieved a ROC-AUC of 0.97 on the manually labeled test set, and 0.96 on the larger automatically labeled dataset, indicating consistent performance across both.

The *yawning* model reached a ROC-AUC of 96.6 percent using mouth aspect ratio alone, which improved to 97.5 percent when combined with action unit predictions from AFFDEX 2.0.

The unattended-screen model classified moments as *inattentive* when both AFFDEX 2.0 and SmartEye failed to detect a face for more than one second. To assess the validity of this, the authors manually annotated all such no-face events in the *real distraction* dataset, identifying the underlying cause of each activation. Ambiguous cases (such as camera obstruction or video distortion) were excluded from the analysis.

As shown in the results table below, only 27 percent of 'no-face' activations were due to users physically leaving the screen.

Reason	Reason percentage	Valid activation (%)?
Gazing off-screen w/ extreme angle	31.5%	True (81%)
Leaving screen unattended	27.4%	
Face largely occluded by hand/object	13.7%	
Excessive participant movement	8.2%	
Significantly cropped or darkened face	16.4%	False (19%)
Rotated face	2.7%	

Diverse obtained reasons why a face was not found, in certain instances.

The paper states:

> *'Despite unattended screens constituted only 27% of the instances triggering the no-face signal, it was activated for other reasons indicative of inattention, such as participants gazing off-screen with an extreme angle, doing excessive movement, or occluding their face significantly with an object/hand.'*

In the last of the quantitative tests, the authors evaluated how progressively adding different distraction signals – off-screen gaze (via gaze and head pose), drowsiness, speaking, and unattended screens – affected the overall performance of their attention model.

Testing was carried out on two datasets: the *real distraction* dataset and a test subset of the *gaze* dataset. G-mean and F1 scores were used to measure performance (although drowsiness and speaking were excluded from the gaze dataset analysis, due to their limited relevance in this context)s.

As shown below, attention detection improved consistently as more distraction types were added, with *off-screen gaze*, the most common distractor, providing the strongest baseline.

Model/Behaviour	Real Distraction Dataset				Gazing Dataset			
	Desktop devices		Mobile devices		Desktop devices		Mobile devices	
	G-mean	F1	G-mean	F1	G-mean	F1	G-mean	F1
AFFDEX 1.0 [22]	0.350	0.170	0.522	0.417	0.423	0.294	0.483	0.362
Gazing off-screen (head model)	0.570	0.384	0.532	0.420	0.589	0.475	0.566	0.461
+ Gazing off-screen (gaze model)	0.788	0.531	0.797	0.650	0.706	0.588	0.756	0.650
+ Drowsiness	0.789	0.533	0.802	0.655	-	-	-	-
+ Speaking	0.804	0.548	0.834	0.687	-	-	-	-
+ Unattended screen (all distractors)	**0.829**	**0.585**	**0.876**	**0.768**	**0.717**	**0.601**	**0.773**	**0.666**

The effect of adding diverse distraction signals to the architecture.

Of these results, the paper states:

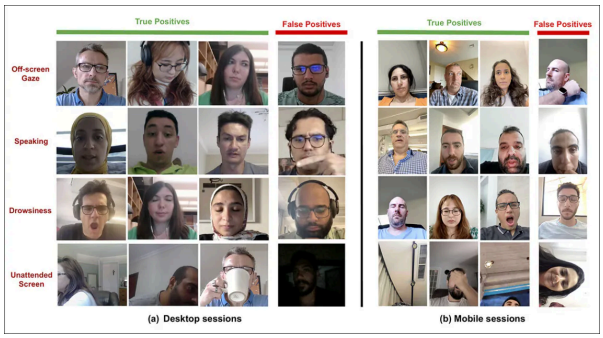
'From the results, we can first conclude that the integration of all distraction signals contributes to enhanced attention detection.

> *'Second, the improvement in attention detection is consistent across both desktop and mobile devices. Third, the mobile sessions in the real dataset show significant head movements when gazing away, which are easily detected, leading to higher performance for mobile devices compared to desktops. Fourth, adding the drowsiness signal has relatively slight improvement compared to other signals, as it's usually rare to happen.*
>
> *'Finally, the unattended-screen signal has relatively larger improvement on mobile devices compared to desktops, as mobile devices can be easily left unattended.'*

The authors also compared their model to AFFDEX 1.0, a prior system used in ad testing – and even the current model's head-based gaze detection outperformed AFFDEX 1.0 across both device types:

> *'This improvement is a result of incorporating head movements in both the yaw and pitch directions, as well as normalizing the head pose to account for minor changes. The pronounced head movements in the real mobile dataset have caused our head model to perform similarly to AFFDEX 1.0.'*

The authors close the paper with a (perhaps rather perfunctory) qualitative test round, shown below.



Sample outputs from the attention model across desktop and mobile devices, with each row presenting examples of true and false positives for different distraction types.

The authors state:

> *'The results indicate that our model effectively detects various distractors in uncontrolled settings. However, it may occasionally produce false positives in certain edge cases, such as severe head tilting while maintaining gaze on the screen, some mouth occlusions, excessively blurry eyes, or heavily darkened facial images. '*

Conclusion

While the results represent a measured but meaningful advance over prior work, the deeper value of the study lies in the glimpse it offers into the persistent drive to access the viewer's internal state. Although the data was gathered with consent, the methodology points toward future frameworks that could extend beyond structured, market-research settings.

This rather paranoid conclusion is only bolstered by the cloistered, constrained, and jealously protected nature of this particular strand of research.

* *My conversion of the authors' inline citations into hyperlinks.*

First published Wednesday, April 9, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai