16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Automatic Measurement of Visual Attention to Video Content using Deep Learning

Attila Schulc[1], Jeffrey F. Cohn[1,2], Jie Shen[3,4], and Maja Pantic[1,3,4]

[1]*Realeyes, Budapest, HU, attila.schulc@realeyesit.com*
[2]*Department of Psychology, University of Pittsburgh, Pittsburgh, USA, jeffcohn@pitt.edu*
[3]*Department of Computing, Imperial College London, London, UK, {jie.shen07, m.pantic}@imperial.ac.uk*
[4]*Samsung AI Center, Cambridge, UK, {jie1.shen, maja.pantic}@samsung.com*

Abstract

Advances in automated face analysis have made possible webcam-based assessment of viewer emotion during presentation of commercials and other video content. A key assumption of this technology is that viewer emotion is in response to the media. Is that assumption warranted? Because viewer attention is seldom assessed, emotional responses could result from other sources, such as talking to a friend, enjoying a meal, or attending to a pet. We developed a CNN-LSTM approach that detects attention and non-attention to commercials using webcam and mobile devices in settings of viewer's choice. Because cultural variation in viewer response is likely, we included participants from both Western and Eastern countries. Participants were 28,911 adults (ages 18 to 69 years) in Europe, USA, Russia, and China. A total of 15,543 sessions (ca. 6.5 million video frames) was analyzed. Accuracy was quantified using a variety of metrics. Our approach outperformed baseline and achieved moderate to high accuracy that approached that of human annotators.

1 Introduction

In advertising, commercials are intended to capture viewer attention, elicit specific emotions, and ultimately influence consumer behavior (i.e., sales). With advances in computer vision and machine learning, it has become possible to detect emotion during viewing of commercials, to do so in naturalistic settings in large numbers of participants, and from that to predict sales performance in the marketplace [1, 2]. The validity of measuring emotion in response to commercials depends in part on the assumption that emotion profiles are in response to commercials themselves. Is that assumption valid? How often do viewers lapse into inattention?

A key question in marketing or in any context in which information is presented is whether content captures the intended receiver's attention. Cognitive abilities are limited, and in a world filled with digital marketing noise, measuring attention in a quantifiable way is of utmost importance for display advertising. Attention is the ability to focus on content while suppressing focus on other stimuli. Attention is a gatekeeper for a successful ad campaign. Depending on the type of creative device, different ways exist for measuring attention with various success. In video advertising, view-ability metrics such as view-through rate, video-completion rate and average view rate are commonly used. While summary measures such as these are valuable, they often are subjective and fail to measure time-varying changes in attention over the course of a presentation.

To address the need for objective, time-varying measures of attention, advertisers are developing custom metrics that leverage widely available webcam-enabled devices. They seek to measure viewer behavior that is indicative of attention. Proposed behavioral indices of attention include head pose, eye gaze, facial expression, and other non-verbal behavior. Automatic measurement using computer vision and machine learning affords an inexpensive, objective, scalable, and unobtrusive way to measure viewer attention at video frame rate. We propose an efficient CNN-based approach to measure viewer attention to commercial media.

2 Existing Work

Tracking of attention has numerous practical applications, but the utilized features and the proposed models can be highly dependent on the actual use case. It is important to emphasize that attention is an umbrella term involving multiple behavior signals, and it is a key component of user engagement detection. For assessment of attention in social interaction [3, 4, 5], human-robot interaction [6, 7], virtual reality and gaming [8], autonomous vehicles and driver safety [9, 10], online learning [11, 12], market research [13, 14] and developmental psychopathology [15], head pose and eye gaze direction have been the primary features mapped to the hidden variable of attention. Head pose and gaze have the advantage of being readily measured in RGB video without use of dedicated eye tracking or other specialized equipment. Our approach requires no eye-trackers, infrared cameras or other specialized sensors, which are expensive and not widely available.

Our approach requires only an RGB webcamera placed in front of the user. Our approach goes beyond gaze, eye closure, and head motion [16, 17] to include comprehensive facial features.

In [18], visual distractions are detected by estimating the face pose of the driver, and sleepiness is inferred from yawning, eyebrow movement, and degree of eye closure. A recent work [19] on driver drowsiness detection uses a multi-granularity Convolutional Neural Network (CNN) on well-aligned facial patches and extracts facial representations. Representation are then fed to a Long Short Term Memory (LSTM) network to extract temporal information. Using temporal dependencies, the trained network can distinguish blinking and eye closure. In online education (e-learning) settings, monitoring students' attention and engagement level is necessary to improve learning quality. The EmotiW Challenge engagement task [20] yielded several novel approaches. In [21] head pose and eye gaze features have been extracted from each frame, then segment features have been created using a sliding window approach. The sequence of segments is processed with an LSTM network with a final average pooling layer. In common with [22], facial features as well as upper body posture features are extracted followed by a bidirectional LSTM with feed forward attention mechanism trained on the video level features. The authors also introduced hand-crafted features from body posture information to adjust the final predicted engagement level.

In the context of TV viewers, [14] trained a support vector machine for binary engagement classification on hand-crafted features using various temporal aggregation methods. Facial expression related features, along with head position and head size features, had the most discriminative power.

Because the facial area is a rich source of non-verbal information about attention and engagement, it makes sense to leverage well established modeling pipelines from facial emotion research (FER), which has received much attention in computer vision [23]. Recent work suggests that deep learning approaches significantly outperform hand-crafted ones and make possible training on far larger numbers of participants. We propose a hybrid deep learning (DL) approach that combines CNN for spatial features and LSTM for temporal features. Both modules are state of the art in FER systems. We evaluate performance in relation to a baseline method and manual human annotation.

3 Data Collection

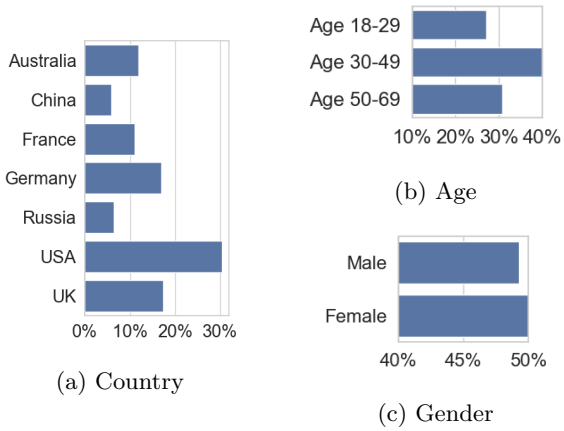
3.1 Participants

Participants were recruited online in the US, Europe, China, and Russia by panel providers (Figure 1) and ranged in age between 18 and 69 years. The proportion of men and women was comparable. Of the 28,911 participants, video from 15,543 was selected for analysis as described below.

3.2 Recordings

In a setting of their choosing, participants watched one or more commercials on a personal computer (86%) or mobile device (14%). Eighty percent of the participants watched a single commercial; the others watched two or more (157 commercials were used in total). Participants' faces were recorded by webcam or by mobile device. Because the settings were highly variable, variations in head pose, illumination, and occlusion were common. While these factors might have been controlled in a laboratory setting, in-the-wild data collection was preferred to minimize inhibition, increase participation, and allow for greater naturalness in participants' responses. All participants gave informed consent and received a small allowance (ca. $.25).

Participants' video was streamed to the cloud for processing. The average duration of the videos was 19.3s (std = 5.9s, min= 10s, max = 30s). Video was captured at a standard resolution of 640 x 480 pixels. Because of variation in internet connections, frame rate varied (mean = 17 fps, std = 8 fps). Frame rate was higher for mobile devices than for personal computers.



(a) Country (b) Age (c) Gender

Figure 1: Session statistics

3.3 Manual Annotation

Manual annotation of video-recorded behavior can be laborious and expensive. The annotator must make a judgment for each frame. A properly designed tool can make a sizable difference in the time and effort required. We developed a custom web-based tool that enables efficient crowd-sourced manual annotation of event onsets and offsets.

Crowd sourcing has the advantage of speed. A disadvantage is that workers are not experts. By providing preliminary training and aggregating annotations

from multiple annotators, however, high reliability can be obtained [24]. Annotators received instruction and practice with the annotation tool and were required to pass a competence test prior to start of annotation. For each video, 7 annotators provided ratings. Majority voting was used to label ground truth for classifier training and testing.

Annotators assigned each video frame to one of two categories:

(1) **Attention**: The participant attends to the screen as inferred from their head pose, gaze, eye closure, and posture. Slight changes in each modality may occur as participants respond to a commercial.

(2) **No attention**: The participant fails to look at the screen, turns their head away from the screen, closes their eyes, or becomes engaged in an unrelated activity (e.g., typing on or telephoning). In some cases the participant may look at the screen, but their attention is divided among several activities, such as talking, eating, or other tasks.

When level of attention could not be determined, they had the option of labeling the frame or frames as *bad quality*. Bad quality frames were characterized by inadequate illumination, self- or other occlusion, or inadequate camera orientation. The participant might be outside of the video frame or have only a portion of their face visible.

Treating the attention task as a binary classification problem (attention/no-attention) reduces task complexity. Annotators can proceed at a much higher speed, and they experience less fatigue than would otherwise be the case. Annotation time and cost thereby are reduced relative to a more complex annotation scheme.

An initial set of 29,000 sessions was selected for processing. To reduce the cost of annotation, a two-wave annotation strategy was employed. First, each video was annotated by two workers. If either worker labeled the minority target class (*"no attention"*), an additional 5 workers were requested, adding up to 7 judgments per each frame in total. Around 60% of the initial set went through a second round, and the results were further filtered based on the *"bad quality"* labels from the human annotators.

The cleaned ground truth consisted of 15,543 sessions, ca. 6.5 million frames. Based on the majority vote 20% of the frames were in the "no attention" category. Seventy one percent of the sessions included at least one *"no attention"* event in them. The distribution of sessions in Figure 2 shows the ratio of frames annotated as *"no attention"* out of all the frames in a session.

Following Rosenthal [25], the effective reliability of the aggregated annotations was quantified using the Spearman-Brown formula:

$$R_{SB} = \frac{nr}{1 + (n-1)r} \quad (1)$$

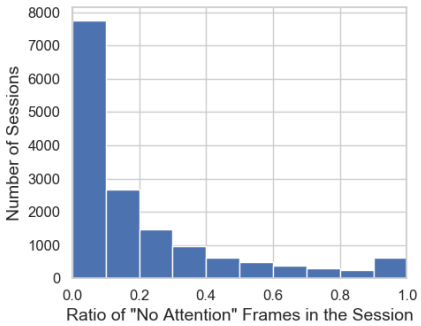


Figure 2: Distribution of the ratio of no attention frames out of all frames in a session

where R_{SB} = 'effective' reliability, n = number of judges (7), r = mean reliability among all n judges, calculated from all possible annotator pairs.

Table 1: Effective reliability of aggregated per-frame annotation

Metric	Attention	No attention	Bad quality
R_{SB}	0.918	0.899	0.905

4 Classifier training

The attention recognition task is formulated as a binary classification problem for each frame. First a CNN model was trained on the still images leveraging the spatial information in individual frames. Second, using the CNNs last layer representation, temporal sequences were generated for training a recurrent neural network. The models were implemented in Python using the Tensorflow and Keras libraries. All training was performed on a Titan X GPU video card.

4.1 Pre-Processing

Each video frame was pre-processed. First, the face was detected and then 49 fiduciary points were localized [16]. These two steps often were combined to increase processing speed. Given a high enough confidence in the localized 49 fiduciary points in a given frame, the position of the face in the following video frame was assumed and only face alignment was performed. Next, RGB images were converted to grayscale and cropped to the face region. Face images were resized to 50 x 50 dimensions for feeding into the first layer of the CNN. Pixel range then was normalized between 0 and 1. Face frontalization [26] typically is an important step in facial expression analysis. Because head orientation is a key indicator of *"no attention"*,

head orientation was not normalized. For the same reason image rotations were not applied during data augmentation.

4.2 Baseline method

We used head pose histogram features as a baseline approach to estimate attention. Use of head pose as a baseline was informed by current state of the art and by the neurophysiology of head and gaze coordination. Affectiva, for instance, uses head pitch and yaw as a proxy for attention [27]. And head and eye gaze are both coordinated by the medial longitudinal fasciculus and moderately correlated, [28, 29]. For our implementation of baseline method, the algorithm used the largest head pose variation within a sliding window of approximately 2 seconds with 1 second strides.

4.3 CNN for Modeling Attention

The CNN model consists of 4 convolutional layers and 4 fully connected layers ending with a softmax layer giving a probability distribution over the two possible classes. Both convolutional and dense layers interspersed with dropout and batch normalization layers were used. To mitigate over-fitting, data augmentation was used (random horizontal flips, zooming, and translations) during training to expose the network to a large variety of training samples.

Four approaches were explored for handling the imbalance in the dataset.

(1) Leave original class ratios, only reshuffle after each epoch.
(2) Give a higher class weight for the minority class, used for weighting the loss function (categorical cross entropy) during training.
(3) Use a custom balanced batch iterator.
(4) Subsample negative class to balance out the two classes.

Approach (1) and (3) proved most effective. The balanced batch iterator over-sampled the minority class, choosing samples randomly from both classes with replacement and applying on-the-fly random image augmentation on each image. Adam optimization was used with exponential learning rate decay and early stopping to avoid overfitting.

4.4 Learning Sequences with LSTM

To leverage temporal variations in the features extracted by the CNN module, long short-term memory (LSTM) is used to model the spatio-temporal dependencies of consecutive frames. After the CNN completes training, representations from the last dense layer are saved for each input frame. Fixed length sequences are generated from the representations for the LSTM training. A sequence length of 35 time-steps was chosen, which corresponds approximately to a 2s video segment given the average frame rate of 17 fps. In each sequence the label for the last frame is the target building up a sequence using the preceding 34 frames. In case the actual target frame is dropped due to bad facial alignment, the network makes no prediction for that frame. Removing bad frames from the queue, sequences are generated using only good quality frames. As in the CNN, the network ends with a softmax layer using Adam optimization with learning rate decay. LSTM training was done without balanced batch iteration, thus maintaining the original distribution of samples.

5 Results

Training and test splits were obtained using a random 80-20 split on the session level. Similarly training was partitioned further into training and validation sets for hyperparameter optimization. No participant appeared in both training, validation, and test sets, and the class ratios were the same across partitions (19.9% train, 21.0% validation, 20.1% test).

Because different metrics correspond to different aspects of agreement and are affected differently by class imbalance [30], we report several: Matthews Correlation Coefficient (MCC), balanced accuracy (BA), area under the Receiver Operating Characteristics Curve (ROC AUC), and *S score* (also known as free-marginal kappa) [31, 32].

S score is a chance-adjusted summary measure that estimates chance agreement by assuming each category is equally likely to be chosen at random. When applied to two annotators or methods (e.g., manual and automated annotation), it is calculated as (2), where n_{00} is the number of objects that both annotators or methods assign to the negative (non-attention) class and n_{11} is the number of objects that both annotators or methods assign to the positive class (attention). *S score* is relatively robust to class imbalance [30]. For each of the metrics, we present 95% confidence intervals to quantify precision of estimate.

$$S = \frac{(n_{00} + n_{11})/n - 1/2}{1 - 1/2} \tag{2}$$

Figure 3 shows an example of system output. The probability of attention is high when the participant is looking toward the screen and decreases when they look or turn their head in another direction, or engage in another activity. Short breaks in the signal happen when face tracking fails.

Table 2 shows classifier performance on the frame level when attention is the positive target class. Because annotators give only categorical scores, AUC for them is not meaningful.

For all metrics, our classifier outperformed baseline and approached the accuracy of manual annotators.

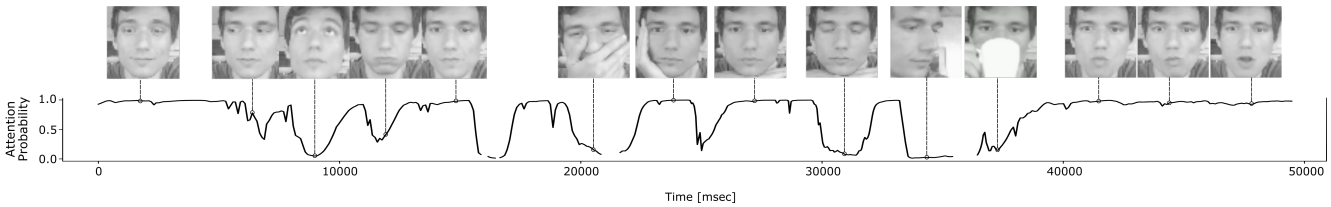


Figure 3: Example showing the predicted attention outputs of a user watching some video content on the screen

Table 2: Results with 95% confidence intervals

Model	MCC	S	BA	AUC
Baseline	.30 ± .03	.59 ± .02	.65 ± .01	.74 ± .0
CNN	.45 ± .02	.64 ± .02	.72 ± .01	.82 ± .0
CNN-LSTM	.49 ± .02	.68 ± .02	.75 ± .01	.85 ± .0
Human	.70 ± .08	.81 ± .08	.86 ± .06	-

Note. MCC is Matthews correlation coefficient, S free-marginal kappa, BA balanced accuracy, AUC area under ROC.

Table 3: Performance differences within subgroups with 95% confidence intervals

Test data grouped by	Partition	MCC	S	BA
Country	AUS	.46 ± .07	.68 ± .04	.73 ± .04
	CHN	.52 ± .07	.58 ± .06	.77 ± .03
	FRA	.45 ± .06	.66 ± .05	.73 ± .03
	DEU	.49 ± .05	.71 ± .03	.73 ± .03
	RUS	.45 ± .09	.66 ± .07	.74 ± .04
	GBR	.52 ± .05	.7 ± .04	.76 ± .02
	USA	.49 ± .04	.68 ± .03	.74 ± .02
Age	50-69	.46 ± .04	.69 ± .03	.72 ± .02
	30-49	.49 ± .04	.66 ± .03	.75 ± .02
	19-29	.48 ± .04	.64 ± .04	.75 ± .02
Gender	Female	.49 ± .03	.67 ± .03	.75 ± .02
	Male	.47 ± .03	.66 ± .03	.74 ± .02
Device	Desktop	.48 ± .02	.66 ± .02	.74 ± .01
	Mobile	.56 ± .06	.78 ± .04	.79 ± .03

As expected, agreement was lowest for MCC, which is attenuated by imbalance among classes [30].

Performance was consistent across differences in country, age, and gender (Table 3). Performance was higher on mobile devices. Faster frame rate and fewer occlusions on mobile likely contributed to this finding.

6 Discussion and Future Work

We developed an automated system that estimates attention to commercial media on webcam or mobile device. In contrast to state-of-the art deep networks for image classification (e.g., VGG, Inception, and ResNet), the system's deep network has a light-weight architecture. This light-weight architecture makes use possible on devices that have small computation capabilities or limited power, such as mobile phones.

In comparison with head-pose based measures of attention, our approach achieved much higher accuracy on multiple metrics and approached that of human annotators. Important for basic and applied research use, the approach proved robust to variation in age, gender, and country. Significant differences between these factors were absent or relatively minor.

A disadvantage of the deep learning approach is that contribution of specific features (e.g., head turn) cannot be quantified. We are exploring visualization methods to improve interpretability. While hand-crafted features and shallow-learning might be considered, the millions of video frames in 15,000 sessions exceed the capacity of shallow learning. In other work, we wish to consider the impact of variable sampling rate and sequence duration and the relation between attention, emotion expression, and product sales.

7 Conclusion

Our approach detected attention and non-attention with moderate to high accuracy on both personal computers and mobile devices. The approach was robust to differences in country, age, and gender. It extends the descriptive power of current approaches that estimate only emotion responsiveness or infer attention from head pose only. Because emotion may be in response to either target media (i.e., commercials) or distractors (e.g., another activity), the ability to capture attention as well as emotion represents a valuable extension of current approaches.

References

[1] G Szirtes, J Orozco, I Petrás, D Szolgay, Á Utasi, and JF Cohn. Behavioral cues help predict impact of advertising on future sales. *Image and Vision Computing*, 65:49–57, 2017.

[2] D McDuff, RE Kaliouby, E Kodra, and L Larguinet. Do emotions in advertising drive sales. In *Proceedings of ESOMAR Congress*, volume 1, pages 2–6, 2013.

[3] R Stiefelhagen, M Finke, J Yang, and A Waibel. From gaze to focus of attention. In *International Conference on Advances in Visual Information Systems*, pages 765–772. Springer, 1999.

[4] B Massé, S Ba, and R Horaud. Tracking gaze and visual focus of attention of people involved in social

interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[5] SO Ba and JM Odobez. Head pose tracking and focus of attention recognition algorithms in meeting rooms. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 345–357. Springer, 2006.

[6] S Sheikhi and JM Odobez. Recognizing the visual focus of attention for human robot interaction. In *International Workshop on Human Behavior Understanding*, pages 99–112. Springer, 2012.

[7] O Palinko, F Rea, G Sandini, and A Sciutti. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054, 2016.

[8] S Asteriadis, K Karpouzis, and SD Kollias. The importance of eye gaze and head pose to estimating levels of attention. In *VS-GAMES*, pages 186–191. IEEE Computer Society, 2011.

[9] L Fridman, J Lee, B Reimer, and T Victor. owland lizard: patterns of head pose and eye pose in driver gaze classification. *IET Computer Vision*, 10(4):308–314, 2016.

[10] M Ngxande, JR Tapamo, and Burke M. Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques. *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 156–161, 2017.

[11] A Gupta, R Jaiswal, S Adhikari, and V Balasubramanian. DAISEE: dataset for affective states in e-learning environments. *CoRR*, abs/1609.01885, 2016.

[12] T Robal, Y Zhao, C Lofi, and C Hauff. Webcam-based attention tracking in online learning: A feasibility study. In *23rd International Conference on Intelligent User Interfaces*, pages 189–197. ACM, 2018.

[13] Y Li, P Xu, D Lagun, and V Navalpakkam. Towards measuring and inferring user interest from gaze. In *WWW*, 2017.

[14] J Hernandez, Z Liu, G Hulten, D DeBarr, K Krum, and Z Zhang. Measuring the engagement level of tv viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[15] K Campbell, KLH Carpenter, J Hashemi, S Espinosa, S Marsan, JS Borg, Z Chang, Q Qiu, S Vermeer, E Adler, et al. Computer vision analysis captures atypical attention in toddlers with autism. *Autism*, 2018.

[16] X Xiong and F De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[17] LA Jeni, JF Cohn, and T Kanade. Dense 3d face alignment from 2d video for real-time use. *Image and Vision Computing*, 58:13–24, 2017.

[18] LM Bergasa, JM Buenaposada, J Nuevo, P Jimenez, and L Baumela. Analysing driver's attention level using computer vision. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 1149–1154. IEEE, 2008.

[19] J Lyu, Z Yuan, and D Chen. Long-term multi-granularity deep framework for driver drowsiness detection. *CoRR*, abs/1801.02325, 2018.

[20] A Dhall, A Kaur, R Goecke, and T Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *ICMI*, pages 653–656. ACM, 2018.

[21] A Mustafa, A Kaur, L Mehta, and A Dhall. Prediction and localization of student engagement in the wild. *CoRR*, abs/1804.00858, 2018.

[22] C Chang, C Zhang, L Chen, and Y Liu. An ensemble model using face and body tracking for engagement detection. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pages 616–622. ACM, 2018.

[23] BC Ko. A brief review of facial emotion recognition based on visual information. In *Sensors*, 2018.

[24] H Li and Q Liu. Cheaper and better: Selecting good workers for crowdsourcing. In *HCOMP*, 2015.

[25] R Rosenthal. Conducting judgment studies: Some methodological issues. In *Handbook of Nonverbal Behavior Research Methods in the Affective Sciences*, pages 199–236. NY: Oxford, 2005.

[26] T Hassner, S Harel, E Paz, and R Enbar. Effective face frontalization in unconstrained images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[27] How do you calculate the user attention using pitch roll and yawn of head? `http://discuss.affectiva.com/t/how-do-you-calculate-the-user-attention-using-pitch-roll-and-yawn-of-head/261`. Accessed: 2018-12-10.

[28] WM King, SG Lisberger, and AF Fuchs. Responses of fibers in medial longitudinal fasciculus (mlf) of alert monkeys during horizontal and vertical conjugate eye movements evoked by vestibular or visual stimuli. *Journal of neurophysiology*, 39(6):1135–1149, 1976.

[29] EM Klier, H Wang, and JD Crawford. Three-dimensional eye-head coordination is implemented downstream from the superior colliculus. *Journal of Neurophysiology*, 89(5):2839–2853, 2003.

[30] LA Jeni, JF Cohn, and F De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013.

[31] EM Bennett, R Alpert, and AC Goldstein. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308, 1954.

[32] RL Brennan and DJ Prediger. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699, 1981.