A Deep Learning Approach for Robust Head Pose Independent Eye Movements Recognition from Videos

Rémy Siegfried
Idiap Research Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
remy.siegfried@idiap.ch

Yu Yu
Idiap Research Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
yu.yu@idiap.ch

Jean-Marc Odobez
Idiap Research Institute
Martigny, Switzerland
EPFL
Lausanne, Switzerland
odobez@idiap.ch

ABSTRACT

Recognizing eye movements is important for gaze behavior understanding like in human communication analysis (human-human or robot interactions) or for diagnosis (medical, reading impairments). In this paper, we address this task using remote RGB-D sensors to analyze people behaving in natural conditions. This is very challenging given that such sensors have a normal sampling rate of 30 Hz and provide low-resolution eye images (typically 36x60 pixels), and natural scenarios introduce many variabilities in illumination, shadows, head pose, and dynamics. Hence gaze signals one can extract in these conditions have lower precision compared to dedicated IR eye trackers, rendering previous methods less appropriate for the task. To tackle these challenges, we propose a deep learning method that directly processes the eye image video streams to classify them into fixation, saccade, and blink classes, and allows to distinguish irrelevant noise (illumination, low-resolution artifact, inaccurate eye alignment, difficult eye shapes) from true eye motion signals. Experiments on natural 4-party interactions demonstrate the benefit of our approach compared to previous methods, including deep learning models applied to gaze outputs.

CCS CONCEPTS

• **Computing methodologies → Tracking**; **Neural networks**; *Activity recognition and understanding.*

KEYWORDS

eye movements, saccade, blink, remote sensors, video processing, convolutional neural network.

ACM Reference Format:
Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. 2019. A Deep Learning Approach for Robust Head Pose Independent Eye Movements Recognition from Videos. In *2019 Symposium on Eye Tracking Research and Applications (ETRA '19), June 25–28, 2019, Denver , CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3314111.3319844

ETRA '19, June 25–28, 2019, Denver , CO, USA

ACM ISBN 978-1-4503-6709-7/19/06…$15.00
https://doi.org/10.1145/3314111.3319844

1 INTRODUCTION

By providing access to the attention of people or even to their intention and mind, gaze tracking finds application in many domains ranging from advertisement, human behavior and communication analysis [3], human-robot/computer interaction [14], psychological studies [19], or medical diagnosis [11]. However, beyond the sheer instantaneous estimation of gaze direction, gaze analytics can often benefit from the recognition of the actual eye movements (fixations, saccades, blinks, ...). They provide not only a good way to denoise the gaze signal and therefore improve attention inference but also a better characterization of the eye activities useful for behavior understanding.

In this paper, we address the recognition of eye movements from videos with a normal sampling rate (30 Hz) and low-resolution eye images (36x60 pixels). Previous approaches instead mainly relied on common infrared-based sensors like Eyelink 1000, iView X or Tobi TX300, but they are rather expensive, often require calibration and can restrain user movements (head pose, headbox size), limit their applicability to screen-based tasks, or can be quite invasive (need to wear goggles). These conditions might not be a problem for applications like medical exams or neurological investigation. However, they render difficult the application of eye movements recognition for gaze analytics at large scales in fields like driving assistance, conversational agents or sociological studies, where we want users to act naturally without head-mounted devices, constrained head pose or the need for user-specific calibration.

Computer vision technologies are best suited for such applications, as they adapt to cheaper sensors and allow a larger variety of head pose and the monitoring of larger spaces. They have their own drawbacks: eye images have lower resolution, sampling rates are limited by the sensors used, natural conditions and novel tasks introduce higher variabilities (e.g. head pose, illuminations, and dynamics of eye movements). Although promising works have been achieved in particular thanks to the use of deep learning techniques [15, 18, 21, 27, 29, 30], the extracted gaze signals remain noisier and less reliable than with dedicated IR eye trackers, making previous methods for eye movement recognition less suitable. Hence, new methods are required but, to the best of our knowledge, none has been proposed to recognize eye movements from standard videos or video-based gaze sensors.

The novel method we propose detects eye movements from the streams of eye images and head pose information, as presented in Fig. 1. Processing the raw signal allows to leverage computer vision and machine learning techniques to distinguish nuisance elements like low-quality data, illumination factors, eye shape variations,

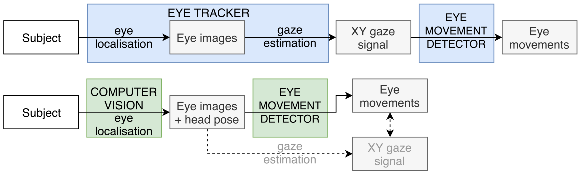

Figure 1: Common workflow when using eye tracker system (top) and proposed workflow (bottom).

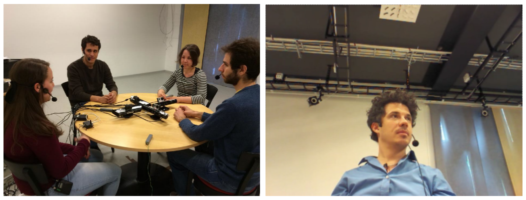

Figure 2: KTH-Idiap dataset [20]. Recording setup (left) and example of video (right).

or bad eye alignments which might be responsible for noisy and unstable gaze outputs, from the information (e.g. pupil motion) useful for the classification of eye movements. We evaluate this method on a dataset consisting of videos from a four-party meeting recorded by Kinect sensors (RGB-D sensor, VGA, 30 Hz recording), as presented in Fig. 2. The nature of the recorded signals forces us to focus on macro movements like fixation and saccade, as post-saccadic oscillations are too subtle to be observed in this kind of data. Our approach also allows to directly detect blink in addition to fixation and saccade, which might also be useful for behavior analysis (e.g. to evaluate light comfort in office space, fatigue state while driving, cognitive load, etc.). In comparison, blinks are often removed manually [1, 22, 24] or by eye trackers, as they lead to data loss [17], but we do not have access to such a filtering method.

The remainder of this paper is organized as follow. Section 2 presents existing work on eye movements detection. Then, we present our proposed method in Section 3, the proposed evaluation protocol in Section 4 and the obtained results with our method and some baselines in Section 5.

2 RELATED WORK

Hand-crafted detection algorithms. The common approach for eye movements recognition is to define one or several features to perform rule-based classification (review in [2]). Starting with basic detectors that use velocity or dispersion to detect fixations [23], more and more complex features and rules were used to improve performances [9, 25] or detect more eye movements [16, 17]. The main limitations are the need to carefully design features and rules with the risk to limit the method’s application to a specific problem and the assumption of high-quality data [28].

Machine learning approaches have been proposed, e.g. using a Random Forest applied to 14 features [28]. Also, a Convolutional Neural Network (CNN) was shown effective to classify fixations, saccades and smooth pursuits in a dataset containing free viewing stimuli experiments [10]. In [6], authors obtained human-level performances with a U-Net inspired network that takes 179 frames equally distributed in the past and the future (corresponding to 180-360 ms time window). However, these methods use clean data recorded at high sampling rate by powerful, but invasive, sensors.

Low sampling rate signals were shown to be significantly more difficult to work with [28]. Some methods still achieved good performances using extracted gaze from mobile eye trackers data, at 30 fps [1, 24]. However, their estimation of gaze mostly relies on eye images with higher resolution than remote sensors can provide.

Importantly, none of the works above use eye images as input, which differs from the method we propose.

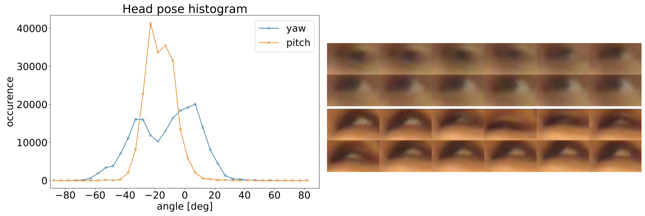

Figure 3: Histogram of estimated head pose (left) and examples of eye image sequences (right).

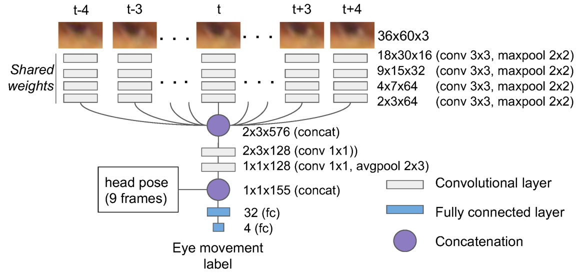

Figure 4: Network architecture of Eye Movement Detector.

3 METHOD

The workflow is presented in Fig. 1, and comprises two main steps. The first one performs an accurate head tracking followed by head frontalization and eye image cropping. The frontalization reduces the variability of the eye appearance due to the head pose orientation. The second step consists of eye movement recognition.

Eye image extraction. Though facial landmark detection [4, 5, 12] has achieved remarkable progress, their performances are sensitive to large head poses variations and occlusions. In this paper, we track the head pose based on the color and depth images, relying on RGB-D sensors along with the Headfusion method [26] which relies on the automatic fitting of both a 3D Morphable Model (3DMM) of the face and a 3D raw representation of the head. This makes the method more robust to the large head poses variations encountered in the used dataset (see Fig. 3a for a histogram). We then rectify the head mesh to a frontal pose as described in [8], and then crop the image of the eye which was closest to the camera (which is less distorted) based on eye landmarks detected with the Dlib library [13]. Sample eye image sequences are shown in Fig. 3b.

Eye movement recognition. The eye movement recognition is based on the processing of eye images sequences (color only). Since the eye activity we target can be distinguished from eye observations over fixed time windows [6], we adopt a temporal sliding window approach relying on a plain neural network consisting of convolutional and fully connected layers, rather than using explicit temporal models like recurrent architectures.

In our model, the label of a frame is estimated by taking as input several images, by processing them first individually to extract relevant and abstract information about the gaze (iris position), concatenate them in a common part to process the sequence of these features and perform the final classification. In the individual part, the feature extraction is the same for all eye images, i.e. the weights are shared. In the common part, the sequence of head poses is injected in the network to improve the results, as eye appearance changes can be due to head pose variations instead of eye movements, like when fixation occur along with head gesture motion.

The architecture is presented in Fig. 4. In our design, the network takes 9 consecutive frames (from $t-4$ to $t+4$) with dimension 36x60 as input and predicts the label of the frame t. Each eye frame is processed by 4 convolutional layers for feature extraction. Then the extracted features of the 9 frames are stacked and further processed by 2 convolutional layers processing along the temporal dimension. The estimated head pose (rotation angles of yaw, pitch, and roll) is introduced at this stage by concatenating the features from the previous layers with the rotation angles of the 9 frames. The result is forwarded to 2 fully connected layers for making final prediction. In this work, eye movement detection is modeled as a classification task and we use a cross-entropy loss for training the network.

This model was developed by investigating and comparing several architectures. Among others, we saw that using information from the past only highly decrease the performance. Also, adding more than 4 frames in the past and future does not improve the results. Our hypothesis is that 4 frames in the future (i.e. 130 ms) are enough for the network to decide if a variation in the eye appearance is due to a blink (recover the same appearance in the future), to a saccade (appearance changes in the future) or to some noise.

4 EXPERIMENTAL SETUP

4.1 Data and annotations

We used the video recordings from the KTH-Idiap database [20]. It consists of five four-party meetings (Fig. 2) in which people discuss naturally alternating monologues, dialogues, and animated discussions. This makes our task challenging because of the highly dynamic nature of the interaction. Indeed, participants are not only looking passively but are actively moving and performing head gestures, facial expressions and social gaze to communicate. Participants were recorded using Kinect sensors (VGA, 30 fps) placed on the table at around 0.8 meters from each participant.

All 20 videos were frame by frame annotated in 5 classes: fixation, blink during a fixation (fix-blink), saccade, blink during a saccade (sac-blink) and unknown. The distinction of blinks happening during fixation and saccade is important, as the eyes behavior is quite different in both cases. As annotations are time-consuming, in each video we took eleven 30 second long segments at regular time intervals for annotation, which represents overall 110 minutes of data. Also, we annotated one out of three frames, as for event detection we care more about annotating a large number of them rather than to precisely segment them. We ended up with a total of 65'000 annotated frames (fix:52'900, sac:8300, fix-blink:2400, sac-blink:1400), with an overrepresentation of fixation (81%),

4.2 Baseline methods

Lacking direct comparison, we used as baseline methods which use as inoput an XY gaze signal instead of eye images. Experiments were made on the same dataset, extracting gaze direction from color eye images using a multi-level HoG SVR [8] trained on a separate dataset. This method was shown to deliver state-of-the-art performance on low-resolution images involving large head rotation. From our experience, it is more reactive to eye motion compared to deep neural networks (although the latter perform better overall), which is important for the task at hand here. Note that it does not explicitly detect blinks, but it usually generates a down-up pattern on the Y-axis.

Dispersion-Threshold Identification (I-DT) [23]. This classic method distinguishes fixations and saccades by measuring dispersion in a moving time window. Parameters were trained to maximize the Cohen's kappa measure.

Naive Segmented Linear Regression (NSLR-HMM) [22]. This is a two-steps method. The first step consists of a segmented linear regression allowing to denoise and segment the signal. Then, an HMM classifies the obtained segments into fixation, saccade, smooth pursuit, and post-saccadic oscillations (PSO) classes. We used the implementation and the trained model provided by the author (https://gitlab.com/nslr/) and considered smooth-pursuit and PSO as fixation to allow comparison with our method.

Convolutional Neural Network (FFT-CNN) [10]. This method pre-processes the XY gaze signal with a Fast Fourier Transform and then use a CNN to predict the probability that the signal is a fixation, saccade or smooth pursuit. We implemented this method and trained it with the same protocol than for our method.

4.3 Experimental protocol

Training and evaluation protocol. We used the "leave one subject out" protocol, ignoring frames labeled as "unknown". To get balanced classes for training we applied down- and up-sampling on the 19 training videos to obtain 20'000 frames for each class, including frames with interpolated labels. However, we used neither interpolation nor resampling on the left out video used for testing. We used 10 epochs and mini-batch processing.

Performance measure. We used the Cohen's kappa [7], which measures agreement for classifications: $\kappa = \frac{p_o - p_e}{1-p_e}$, where p_o is the observed agreement probability and p_e the probability of random agreement. Usually, the agreement is considered as weak if $\kappa > 0.2$, as moderate if $\kappa > 0.4$ and as strong if $\kappa > 0.6$. It is more robust than classical accuracy as it takes into account the probability of random agreement, which is especially important in unbalanced datasets like this one. The Cohen's kappa was computed on each video, to highlight the variance across different subjects.

Classification tasks. Comparing methods that do not extract all the same eye movements is not possible without defining a way to combine and/or ignore some labels. We defined the following tasks:

- **4 classes.** 4 classes: fixation, fix-blink, saccade and sac-blink;

Method	Task	κ mean	κ std
our	4 classes	.536	.093
our	fix-sac-blink	.552	.091
our	blink-others	.671	.104
our	fix-sac	.501	.130
FFT-CNN [10]	4 classes	.417	.062
FFT-CNN [10]	fix-sac-blink	.431	.059
FFT-CNN [10]	blink-others	.297	.064
FFT-CNN [10]	fix-sac	.480	.122
I-DT [23]	fix-sac	.306	.095
NSLR-HMM [22]	fix-sac	.369	.065

Table 1: Evaluation of methods.

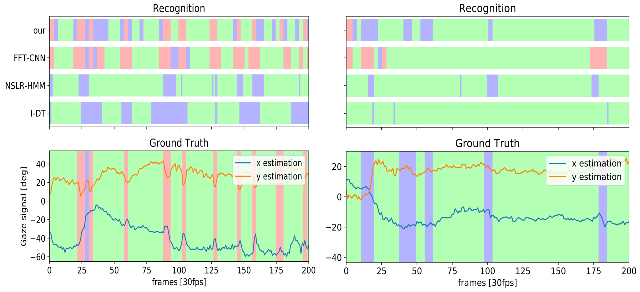

Figure 5: Qualitative comparison over 2 segments. Colors represent labels (green: fixation, blue: saccade, pink: blink), while curves represent the 3D gaze direction signal.

- **fix-sac-blink.** 3 classes: fixation, saccade and blink (fix-blink and sac-blink are merged in blink);
- **blink-others.** 2 classes: blink (i.e. fix-blink and sac-blink) and others, which is the combination of fixation and saccade;
- **fix-sac.** 2 classes: fixation and saccade. Fix-blink and sac-blink frames in the ground truth are ignored for evaluation and frames recognized as fix-blink/sac-blink are considered as fixation/saccade respectively.

5 RESULTS

Method evaluation. Results are reported in Tab. 1. Our method achieves an overall moderate agreement with the ground truth, which is a good result given the difficulty of the tasks. It seems particularly suited to detect blink (blink-others task) but saccade recognition is more challenging (fix-sac task). Note that the standard deviations show that performances variate across subjects.

Comparison with baselines. In Tab. 1, one can notice the overall low scores of all methods, highlighting again the difficulty of the task. The FFT-CNN method reaches lower performance than our method, although it is not significant for the fix-sac one. It struggles to distinguish saccades from blinks, which is consistent with the intuition that blinks are better handled using eye images than the gaze signal. Per-class accuracy validates this result: FFT-CNN reaches 86% accuracy on fixation but only 33% for saccades (versus 81% and 77% for our method), often confusing saccade and blinks. It shows that (1) deep learning seems an appropriate approach as FFT-CNN and our method beat the two other baselines and (2) that using eye images helps to detect blinks, while not decreasing saccade detection performances.

Qualitative results. The left side of Fig. 5 presents an example in which all classifiers detect most of the blinks, although I-DT and NSLR-HMM are predicting saccades for blinks. I-DT tends to merge successive blinks and deep learning methods tend to mistakenly predict saccades before and after blinks. Here, NSLR-HMM performs well.

In the example on the right of Fig. 5 we can see that I-DT and FFT-CNN struggle to detect saccades. NSLR-HMM is already better but still misses two events. Those events correspond to small saccadic movements which are difficult to distinguish from noise in the gaze signal. It shows that using eye images helps to recognize subtle saccades that would be mixed with noise in the gaze signal.

Discussions and limitations. Our method relies on future information which creates a delay of about 130 ms between frame acquisition and eye movement estimation. Some applications will suffer from this, but instant reactivity is not always needed, like for off-line analysis, global statistics computation (blink rate) and low-frequency behavior estimation (attention, conversation regime).

All the proposed baselines rely on the same gaze estimation method, chosen because it tends to react consistently to eye movements. It would be interesting to make experiments with more recent methods to see if those baseline methods can be improved.

Regarding performance, most errors of our method consist in predicting saccades instead of fixations around blinks. It might be interesting to check whether a temporal method, like HMM or LSTM, could help to better learn the label transition statistics and feature dynamics. However, the few tests we made using LSTM were not conclusive, showing that it is not straight forward.

6 CONCLUSION

In this paper, we proposed a method based on computer vision and deep learning in order to detect fixation, saccade, and blink in natural interaction video recorded with remote sensors. We show that deep learning approaches outperform classical methods for saccade detection when facing noisy data coming from computer vision methods instead of dedicated IR eye tracker sensors. Also, our method outperforms another deep learning approach on blink detection task, using eye images instead of gaze XY signal.

One limitation of our work is the precision of the annotations, as the sampling rate of the sensor used to record the data is relatively slow compared to the events we want to detect. Also, we compared our approach with baseline methods that were not designed for the exact same task, in term of detected eye movements or data quality. That shows an advantage of deep learning methods, which can be retrained on a different set of labels for direct comparison.

Finally, the overall performances obtained on the presented dataset show that detecting eye movements in low-sampling rate data acquired with remote sensors in natural conditions remains a challenging task, although it is of high interest in many fields.

ACKNOWLEDGMENTS

This research has been supported by the European Union Horizon 2020 research and innovation programme (grant agreement no. 688147, MuMMER project). Also, it uses the KTH-Idiap database made available by KTH, Sweden.

REFERENCES

[1] N. Anantrasirichai, Iain D. Gilchrist, and David R. Bull. 2016. Fixation identification for low-sample-rate mobile eye trackers. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3126–3130. https://doi.org/10.1109/ICIP.2016.7532935

[2] Richard Andersson, Linnea Larsson, Kenneth Holmqvist, Martin Stridh, and Marcus Nystrom. 2017. One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods* 49, 2 (2017), 616–637. https://doi.org/10.3758/s13428-016-0738-9

[3] S. Ba and J-M. Odobez. 2008. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*. Las-Vegas.

[4] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops (ICCVW '13)*. IEEE Computer Society, Washington, DC, USA, 354–361. https://doi.org/10.1109/ICCVW.2013.54

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *FG*. IEEE Computer Society, 59–66.

[6] Marie E. Bellet, Joachim Bellet, Hendrikje Nienborg, Ziad M. Hafed, and Philipp Berens. 2018. Human-level saccade detection performance using deep neural networks. https://doi.org/10.1101/359018

[7] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[8] Kenneth A. Funes-Mora and Jean-Marc Odobez. 2016. Gaze Estimation in the 3D Space Using RGB-D Sensors. *International Journal of Computer Vision* 118 (2016), 194–216.

[9] Roy S. Hessels, Diederick C. Niehorster, Chantal Kemner, and Ignace T. C. Hooge. 2017. Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods* 49, 5 (2017), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1

[10] Sabrina Hoppe and Andreas Bulling. 2016. End-to-End Eye Movement Detection Using Convolutional Neural Networks. arXiv:1609.02452

[11] Linda Isaac, Janna N. Vrijsen, Mike Rinck, Anne Speckens, and Eni S. Becker. 2014. Shorter gaze duration for happy faces in current but not remitted depression: Evidence from eye movements. *Psychiatry Research* 218, 1-2 (2014), 79–86.

[12] Vahid Kazemi and Josephine Sullivan. 2014. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*. IEEE Computer Society, Washington, DC, USA, 1867–1874. https://doi.org/10.1109/CVPR.2014.241

[13] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[14] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez. 2011. Engagement-based Multi-party Dialog with a Humanoid Robot. In *SIGDIAL Conference*.

[15] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, and Harini Kannan. 2016. Eye Tracking for Everyone. *IEEE Conference on Computer Vision and Pattern Recognition* (2016), 2176–2184. https://doi.org/10.1109/CVPR.2016.239 arXiv:arXiv:1606.05814v1

[16] Linnéa Larsson, Marcus Nystrom, Richard Andersson, and Martin Stridh. 2015. Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control* 18 (2015), 145–152. https://doi.org/10.1016/j.bspc.2014.12.008

[17] Linnea Larsson, Marcus Nystrom, and Martin Stridh. 2013. Detection of Saccades and Postsaccadic Oscillations in the Presence of Smooth Pursuit. *IEEE Transactions on Biomedical Engineering* 60, 9 (2013), 2484–2493. https://doi.org/10.1109/TBME.2013.2258918

[18] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. 2018. A Differential Approach for Gaze Estimation with Calibration. In *29th British Machine Vision Conference*.

[19] Skanda Muralidhar, Remy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. 2018. Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills?. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 121–126. https://doi.org/10.1145/3282894.3282925

[20] Catharine Oertel, Kenneth A Funes, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. 2014. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. 27–32.

[21] Seonwook Park, Adrian Spurr, and Otmar Hilliges. 2018. Deep Pictorial Gaze Estimation. (September 2018).

[22] Jami Pekkanen and Otto Lappi. 2017. A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific Reports* 7, 1 (2017), 17726. https://doi.org/10.1038/s41598-017-17983-x

[23] Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 71–78. https://doi.org/10.1145/355017.355028

[24] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. 2016. Bayesian Identification of Fixations, Saccades, and Smooth Pursuits. In *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research & Applications*. ACM, 163–170. https://doi.org/10.1145/2857491.2857512

[25] Giacomo Veneri, Pietro Piu, Francesca Rosini, Pamela Federighi, Antonio Federico, and Alessandra Rufa. 2011. Automatic eye fixations identification based on analysis of variance and covariance. *Pattern Recognition Letters* 32, 13 (2011), 1588–1593. https://doi.org/10.1016/j.patrec.2011.06.012

[26] Yu Yu, Kenneth Funes-Mora, and Jean-Marc Odobez. 2017. Robust and Accurate 3D Head Pose Estimation through 3DMM and Online Head Model Reconstruction. In *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 711–718.

[27] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2018. Deep Multitask Gaze Estimation with a Constrained Landmark-Gaze Model. In *European Conference on Computer Vision Workshop*.

[28] Raimondas Zemblys, Diederick C. Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. 2016. Using machine learning to detect events in eye-tracking data. *Behavior Research Methods* 50, 1 (2016), 160–181. https://doi.org/10.3758/s13428-017-0860-3

[29] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2015. Appearance-based gaze estimation in the wild. (2015), 4511–4520.

[30] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2016. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. (2016). arXiv:1611.08860 http://arxiv.org/abs/1611.08860