University of Antwerp
Faculty of Business and Economics

Implicit train-free calibration for video-based eye-tracking

V. Mygdalis and N. Dens
Department of Marketing, University of Antwerp

EURASIP
EUROPEAN ASSOCIATION FOR SIGNAL PROCESSING

Introduction

- ***Eye-tracking* or *gaze estimation*** refers to the problem of determining ***"where"*** **a (human) subject is looking at on a given specific time moment.**
- **Eye-tracking has attracted the interest of a wide range of disciplines**
  - ranging from signal processing, neuroscience, computer vision, machine learning, due to its important applications in psychology, medicine, sports, virtual reality, robotics, education, and marketing, to name a few.
- **Many different solution variants exist depending on the target application and its specifications:**
  - infrared cornea reflection
  - event camera-based,
  - eye-tracking glasses
  - **video-only eye-tracking**, which is the main focus of this paper.

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Eye-tracking in marketing

- Eye-tracking is an invaluable tool for extracting metrics (e.g., fixation time to some stimuli) related to cognitive emotional responses, such as attention.
- Controlled marketing experiments are expensive, time consuming and in many cases unnatural.
- Equipment used in marketing experiments is typically too much over-specified for the task at hand (detecting saccades is rarely if ever used).
- These cost limitations leads to developing relatively small datasets with few participants, thus gaining limited insights.
- Video-based eye-tracking can offer a valuable alternative, being implemented on mobile devices or webcams, however, it suffers for the standard challenges of appearance-based eye-tracking.

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

The video-based gaze-tracking problem

- **Consists of several steps, each accumulating error:**
  - Camera calibration (estimating intrinsic and extrincic camera parameters)
  - Face detection (using haar cascades or Deep Learning)
  - Optional steps:
    - Eye detection
    - Fiducials
    - Estimation of the 3D eye-position
  - **Estimation of the gaze vector (3d or 2d after canceling translation and scaling factors)**
  - Intersection of the gaze vector with the computer screen (2D pixel coordinates)

CONVISE project has received funding from the European Union’s HE research and innovation programme under the g.a.n. 101103256

Challenges in appearance-based gaze-estimation

- **Multiple sources of input variance:**
  - **capture conditions, that can be controlled** by stringent experimental settings, well-defined camera specifications, experiment locations, illumination conditions, subject distances/angles from the sensor, and/or even employing head/chin rests wherever possible.
  - **variance related to individual test subject appearance**, such as their physical characteristics (e.g., age, gender, skin/eye/face color/dimensions, and ophthalmic health conditions)
    - That can only be controlled with calibration….

University of Antwerp | Faculty of Business and Economics
EUSIPCO 2024
CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Problem statement

- **Assume a dataset of $\mathcal{D} = \{\mathbf{I}_i, \boldsymbol{y}_i, p_i\}, i = 1, \dots, N$ items, where**
  - $\mathbf{I} \in \mathcal{R}^{\boldsymbol{H}\times\boldsymbol{W}\times\boldsymbol{C}}$ **are facial/eye cropped images**
  - $\mathbf{y} \in \mathcal{R}^{\mathbf{2}}$ **is the normalized gaze direction (angular coordinates)**
  - $p \in \{1, \dots, P\}$ is an index for each participant
- **Gaze estimation is formulated as a regression problem:**
  - $\boldsymbol{g}(\mathbf{I}, \boldsymbol{\theta}) \mapsto \mathcal{X}, \boldsymbol{x} \in \mathcal{R}^{\boldsymbol{D}}$, a mapping function from images to features
  - $\widehat{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b})$ where $\boldsymbol{W}, \boldsymbol{b}$ are the Weight and bias of a linear operation
- **The linear operation is trained simultaneously with the mapping function using an appropriate loss function:**
  - Typically the L1 loss $L = |\widehat{\boldsymbol{y}} - \boldsymbol{y}|$.

University of Antwerp | Faculty of Business and Economics
EUSIPCO 2024
CONVISE project has received funding from the European Union’s HE research and innovation programme under the g.a.n. 101103256

Uncalibrated vs uncalibrated gaze estimation

- **In the uncalibrated gaze estimation $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b})$ is the same for all subjects, not updated in the test phase.**
- **In the calibrated gaze estimation $\boldsymbol{f}_p(\boldsymbol{x}; \boldsymbol{W}_p, \boldsymbol{b}_p)$, a different linear operation is learned for each subject/participant:**
  - Support Vector Regression.
  - Few shot learning e.g., Model Agnostic Meta-Learning
  - SVR works better with many calibration points, Few-shot learning with fewer, it is not trivial to decide which to employ.
- **Calibrated gaze estimation works significantly better than uncalibrated, but it requires ground truth annotations.**
  - Additional time, prone to subject co-operation issues, limitations in applicability.
  - Many works address these limitations by employing un-supervised learning of the linear operation.

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Contribution

- **We attempt to control between subject variances by proposing an architecture that supports *implicit – train-free* system calibration for each individual subject.**
  - Unlike existing approaches, the proposed methodology requires ***no model re-training/finetuning*** and ***no annotations*** at all.
- **The proposed architecture provides considerable performance gains when compared to its respective uncalibrated baseline (which remains a fair comparison).**
- **Besides performance gains, the proposed method offers practical implications**
  - minimize individual researcher calibration efforts and potential calibration errors, reducing the time spent by human subjects during each experimental session.

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Method properties

- **Implicit calibration is achieved by employing merely per participant (facial) images**
  - in a novel calibration-aware neural architecture that learns to operate with comparative/attentive information between the test participant image and the proposed ***calibration anchors***.
  - **Calibration anchors** are features extracted by using representative images for a given train/test subject.
  - The derived features of the test images are combined with the calibration anchors using an attention mechanism.
- **Neither the linear operation nor the neural architecture are trained/finetuned during the test phase.**

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Differences with state of the art

- **SoA focuses on deriving more efficient architectures for the feature extraction phase (i.e., faster, stronger):**
  - $\boldsymbol{x} = g(\boldsymbol{I}; \boldsymbol{\theta})$
- **Or better ways to train or derive the regression part:**
  - $\boldsymbol{f}_p(\boldsymbol{x}; \boldsymbol{W}_p, \boldsymbol{b}_p)$,
- **Our focus is to train $g()$ and $f()$ only once from the train dataset, and then adapt by different inputs:**
  - $\mathbf{x} = g(\boldsymbol{I}; \boldsymbol{\theta})$, features of input images
  - $\mathbf{z} = g(\boldsymbol{S}, p; \boldsymbol{\theta})$, features of “support” images of participant p_i
  - $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{z}; \mathbf{W}, \mathbf{b})$

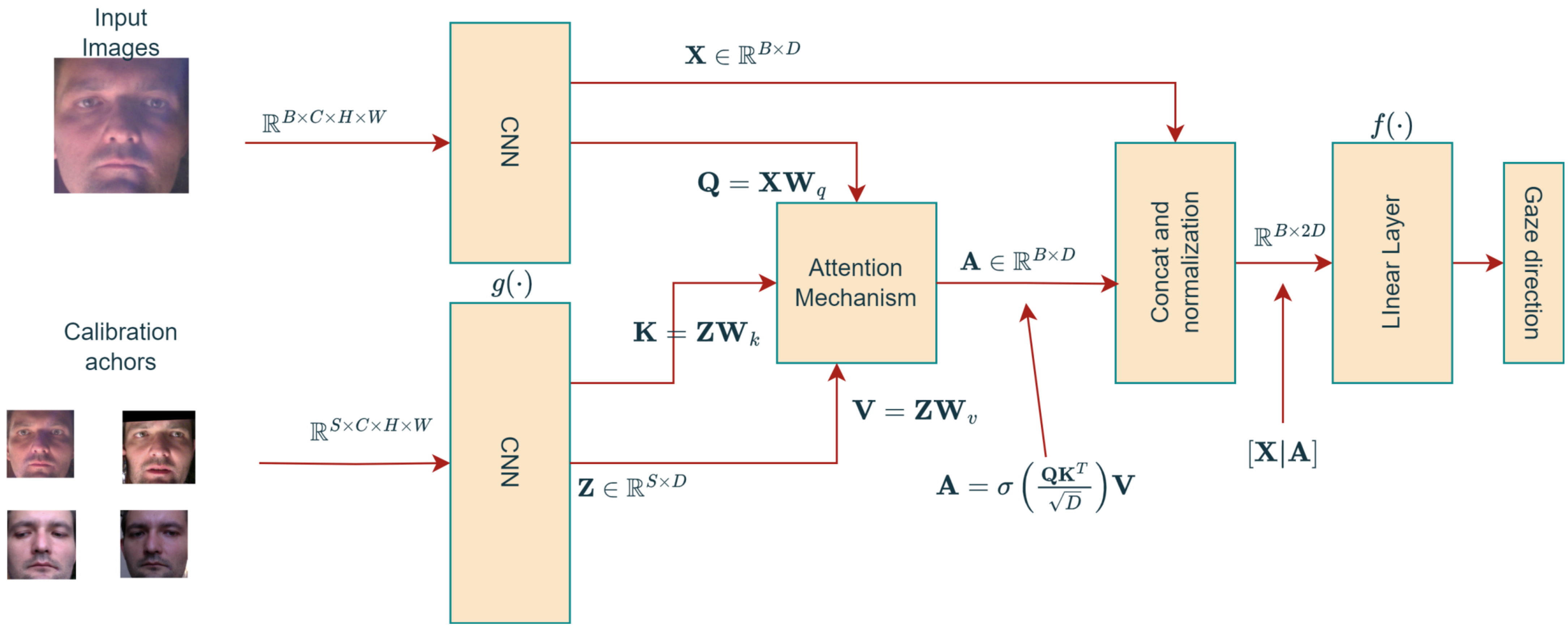
University of Antwerp | Faculty of Business and Economics
EUSIPCO 2024
CONVISE project has received funding from the European Union’s HE research and innovation programme under the g.a.n. 101103256

How calibration anchors are obtained

- **During training:**
  - Recall $\mathcal{D} = \{\mathbf{I}_i, \boldsymbol{y}_i, p_i\}, i = 1, \dots, N$ a gaze estimation dataset of P subjects.
  - We break this dataset into subsets for each subject: $\mathcal{D}_p = \{\mathbf{I}_i, \boldsymbol{y}_i\}, i = 1, \dots, N_p$.
  - We define support gazes as the intersection between (roughly) similar gazes directions between subjects:
    - $\mathcal{Y}_s = \{\mathcal{Y}_1, \cap \cdots \cap Y_P\}$
  - Our support set consists of images corresponding to the subset of roughly similar gazes (bottom looking, top looking, left looking, etc.).
- **At deployment stage, the support set is obtained by using random images for the test participant, and ordered based on detected gaze direction.**
- **No ground truth neither model re-training with these data is needed.**

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Overall architecture



CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Experimental results

TABLE I
ANGULAR ERRORS OF COMPETING METHODS IN MPIIFACEGAZE DATASET

Method	p00	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10	p11	p12	p13	p14	Average
Baseline	2.31	4.77	5.2	6.29	3.76	4.33	3.03	4.62	4.69	5.06	5.84	6.02	4.6	4.24	7.38	4.81
Proposed Method	**2.19**	**4.73**	**3.09**	5.84	3.77	**4.18**	3.08	**4.33**	**4.65**	**3.64**	**5.25**	5.64	**3.82**	4.33	**5.79**	**4.29**
Proposed Architecture*	2.48	6.42	3.3	**5.6**	**3.63**	4.24	**2.91**	4.38	4.68	5.08	7.25	**5.25**	4.06	**3.93**	6.02	4.61
Baseline + FT (5 epochs)	2.31	4.23	4.28	6.06	3.84	4.25	2.98	4.54	4.81	4.58	4.8	5	4.15	4.17	6.93	4.46
Baseline + FT (10 epochs)	2.64	4.16	3.21	5.94	4.25	4.37	3.2	4.54	5.11	4.43	3.85	3.91	3.95	4.3	6.67	4.30
Baseline + FT (15 epochs)	3.31	4.66	3.34	6.31	4.95	4.9	3.71	4.78	5.53	4.87	3.99	3.83	4.37	4.74	7.11	4.69
Baseline + SVR	2.9	3.28	3.21	4.93	**3.35**	**4.09**	4.45	4.9	6.57	3.62	4.08	**2.99**	4.44	5.5	**4.55**	4.19
Proposed + SVR	2.4	**3.15**	**2.11**	**4.14**	3.99	5.08	**2.81**	4.41	4.7	**3.45**	**2.75**	3.97	**3.32**	**3.89**	6.26	**3.76**

TABLE II
ANGULAR ERRORS OF COMPETING METHODS IN MPIIGAZE DATASET

Method	Resnet	Lenet
Baseline	5.76	6.33
Baseline + FT	5.55	5.91
Baseline + SVR	5.50	6.21
Proposed	**5.06**	**5.73**

University of Antwerp
Faculty of Business and Economics
EUSIPCO 2024
CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Significance analysis

- **Friedman's statistical tests:**
  - MPIIFacegaze dataset ($p = 0.0201$)
  - MPIIGaze-Resnet ($p = 0.0201$)
  - MPIIGaze-Lenet ($p = 0.0008$)
- **Wilcoxon signed-rank tests:**
  - MPIIFacegaze dataset ($p = 0.0025$)
  - MPIIGaze-Resnet (p-value =0.0071)
  - MPIIGaze-Lenet (p-value = 0.0001).

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Explanation of the results and limitations

- The architecture itself although introducing some additional parameters and complexity, does not improve the results significantly, most performance comes when anchors are appropriately selected.

- The involvement of support samples normalizes some variance related to subject appearance, thus potentially leading to improved generalization on the feature extraction and regression operations.

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Limitations

- Results have only been tested on datasets where translation and rotation factors had been cancelled.
- We examined datasets with limited variation in illumination conditions and distances from the screen.
- Further improvements could be expected when the method is applied on raw images (realistic case) or conditions of wider variance.
- Neverthelss, optimizing the number of support samples or estimating the influence of each support sample are non-trivial problems.

CONVISE project has received funding from the European Union’s HE research and innovation programme under the g.a.n. 101103256

Conclusions and future work

- **A gaze estimation method that implicitly learns to operate for different subjects was described. Important and statistically significant differences were observed between the proposed and competing methods.**
- **This architecture is promising for other regression problems that present individual subject particularities (e.g., human 3D pose estimation).**
- **Future work:**
  - further polishing of the architecture and training procedure and additional comparisons in other datasets and more baselines.
  - Ablations with different types and number of support samples (during both the training and the test phase)

CONVISE project has received funding from the European Union's HE research and innovation programme under the g.a.n. 101103256

Thank you!

Q & A

Vasileios.mygdalis@uantwerpen.be

CONVISE project has received funding from the European Union’s HE research and innovation programme under the g.a.n. 101103256