Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues

C. Palmero[1,2], J. Selva[1], M.A. Bagheri[3,4], and S. Escalera[1,2]

[1]Universitat de Barcelona, [2]Computer Vision Center, [3]University of Calgary, [4]University of Larestan

Universitat de Barcelona
CVC Computer Vision Center
HuPBA

Motivation

Remote 3D gaze estimation **without user calibration** is still an open issue.
State of the art: (deep) **appearance-based** approaches...

- □ Do not consider global structure explicitly.
- □ Mainly evaluated on HCI scenarios with restricted head pose and gaze direction.
  - Not suitable for general everyday settings.
- □ Only use static eye region appearance as input, but:
  - Gaze behavior is **not** static.
  - Whole-face images encode more head pose and illumination-specific information [4].

Fig 1. The Wollaston effect [3], "*the exact same set of eyes may appear to be looking in different directions due to the surrounding facial cues*".

Proposed Approach

- □ **Subject** and **head pose-independent** multi-modal recurrent CNN for 3D gaze regression with remote calibrated RGB cameras.
- □ The **sequential information of eye and head movements** is leveraged by combining static appearance and shape features on consecutive frames.
- □ **Face landmarks used as global shape cues** encode geometric constraints.

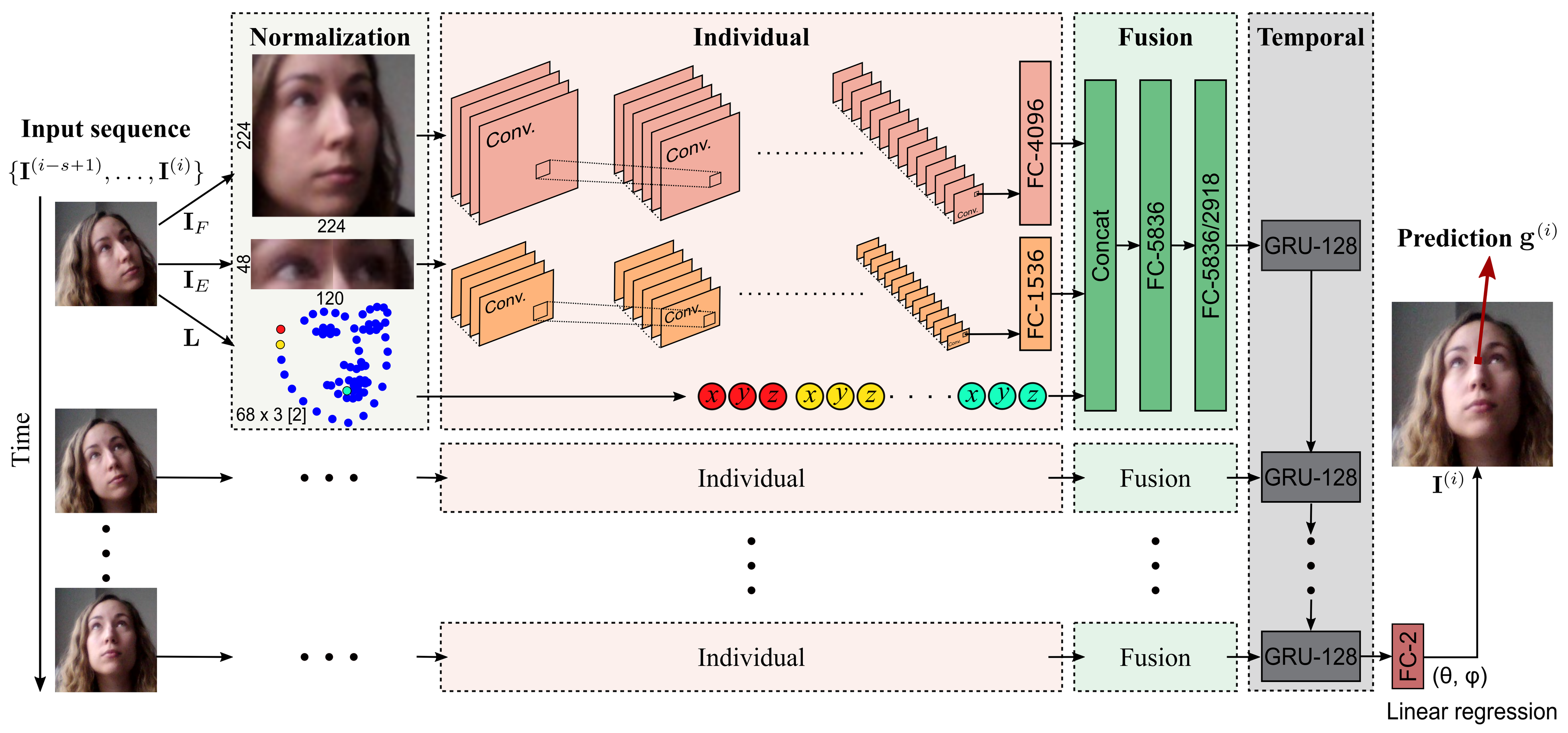

Fig 2. Pipeline overview. VGG-16 as base network for conv. blocks. Dropout between Fusion FCs as regularization.

3D Space Normalization

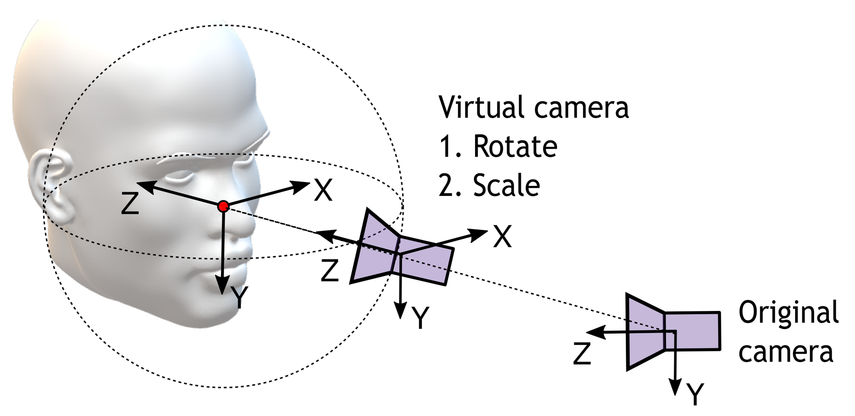

Fig 3. 3D space normalization process.

- □ Reduces the appearance variability.
- □ Makes the model invariant to intrinsic camera parameters.
- □ Gaze vector is **rotated** according to virtual camera transformation.

Stage-wise Training

1. Train **Static** model end-to-end on each individual frame:
   - □ *Individual* and *Fusion* modules and final regression layer.
   - □ Convolutional blocks pre-trained with VGGFace dataset.
   - □ FCs trained from scratch.
2. Train **Temporal** model:
   - □ Re-arrange training data to build input sequences of s = 4 frames.
   - □ Extract features of each sequence frame from frozen *Individual* module.
   - □ Fine-tune *Fusion* layers.
   - □ Train *Temporal* module and final regression layer from scratch.

Loss - average Euclidean distance.

Experimental Evaluation

EYEDIAP dataset: 3-minute VGA videos with 2 lighting conditions.
2 scenarios:

- □ *CS* - Continuous **Screen** target
  - 14 subjects.
  - All head poses: 5-fold CV.
- □ *FT* - **Floating ball** Target
  - 16 subjects.
  - All head poses: 4-fold CV.
  - *Static* and *moving* head poses separately: leave-one-out CV.

Pre-processing:

- □ Filter inconsistent data.
- □ Apply **data augmentation**.

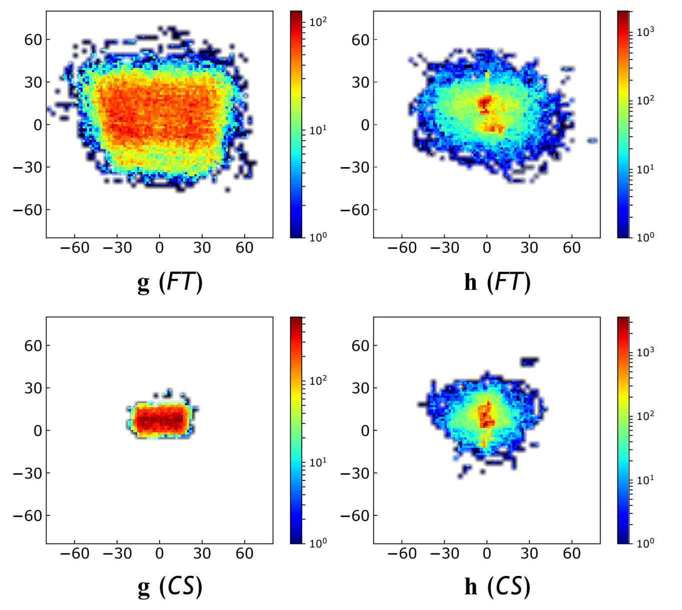

Fig 4. Ground-truth eye gaze **g** and head orientation **h** distribution on the filtered EYEDIAP dataset [1], in terms of x- and y- angles.

Static Modalities

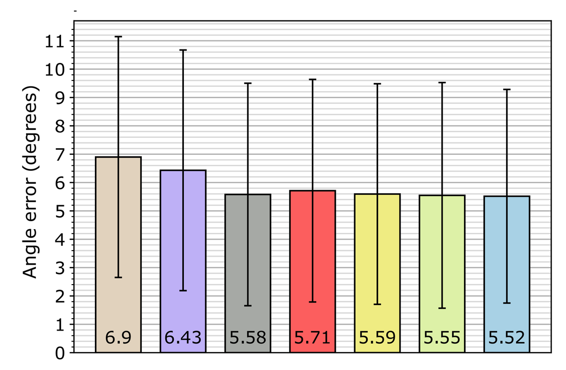

Fig 5. Performance evaluation of the *Static* network using different input modalities (*O - Not normalized, N -Normalized, F- Face, E - Eyes, L - 3D Landmarks)* and size of fusion layers on the *FT* scenario.

Static vs Temporal

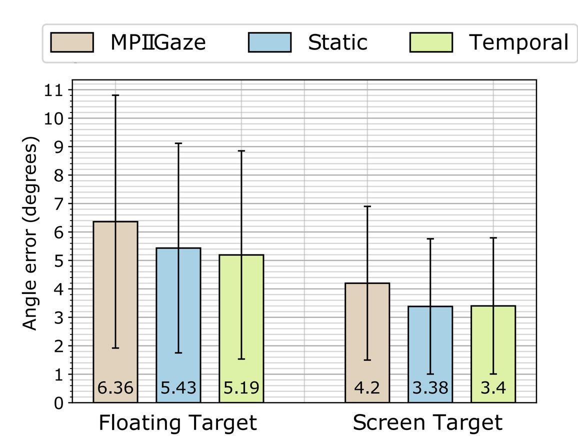

Fig 6. Comparison among state-of-the-art method MPIIGaze [4] and the *Static* and *Temporal* versions of our network.

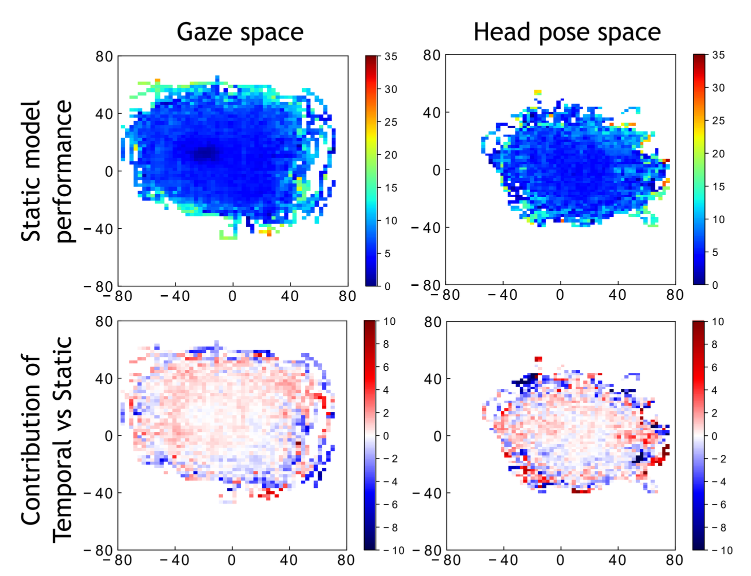

Fig 7. Angular error distribution on the *FT* scenario, in terms of x- and y- angles.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg.
Head	23.5	22.1	20.3	23.6	23.2	23.2	23.6	21.2	26.7	23.6	23.1	24.4	23.3	24.0	24.5	22.8	23.3
PR-ALR[1]	12.3	12.0	12.4	11.3	15.5	12.9	17.9	11.8	17.3	13.4	13.4	14.3	15.2	13.6	14.4	14.6	13.9
MPIIGaze	5.3	5.1	5.7	4.7	7.3	15.1	10.8	5.7	9.9	7.1	5.0	5.7	7.4	3.8	**4.8**	5.5	6.8
Static	**3.9**	**4.1**	**4.2**	**3.9**	**6.0**	**6.4**	7.2	**3.6**	**7.1**	**5.0**	5.7	6.7	**3.9**	4.7	5.1	**4.2**	**5.1**
Temporal	4.0	4.9	4.3	4.1	6.1	6.5	**6.6**	3.9	7.8	6.1	**4.7**	**5.6**	4.7	**3.5**	5.9	4.6	5.2
Head	19.3	14.2	16.4	19.9	16.8	21.9	16.1	24.2	20.3	19.9	18.8	22.3	18.1	14.9	16.2	19.3	18.7
MPIIGaze	7.6	6.2	5.7	8.7	10.1	12.0	12.2	6.1	8.3	5.9	6.1	6.2	7.4	4.7	4.4	6.0	7.3
Static	**5.8**	5.7	**4.4**	**7.5**	6.7	8.8	**11.6**	5.5	8.3	5.5	5.2	6.3	**5.3**	**3.9**	**4.3**	**5.6**	6.3
Temporal	6.1	**5.6**	4.5	**7.5**	**6.4**	**8.2**	12.0	**5.0**	**7.5**	**5.4**	**5.0**	**5.8**	6.6	4.0	4.5	5.8	**6.2**

Table 1. Gaze angular error comparison for *static* (top half) and *moving* (bottom half) head pose for each subject in the *FT* scenario.

Conclusions

- □ The approach combines face and eye appearance, facial landmarks and temporal information, and is tested on a wide range of head pose and gaze directions.
- □ Our multi-modal *Static* model achieves a **significant improvement of 14.6% and 19.5% over the state of the art** on EYEDIAP *FT* and *CS* scenarios, respectively.
- □ Adding geometric features to appearance-based methods has a regularizing effect on accuracy results.
- □ Adding sequential information further **benefits the final performance by up to 4% compared to static-only input**, especially when head motion is present.

[1] K. A. Funes Mora, F. Monay, and J.M. Odobez. "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras". In *Proceedings ETRA*, pages 255-258. ACM, 2014.
[2] A. Bulat and G. Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)". In *ICCV*, 2017.
[3] W. H. Wollaston et al. "Xiii. on the apparent direction of eyes in a portrait". *Philosophical Transactions of the Royal Society of London*, 114:247-256, 1824
[4] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. "It's written all over your face: Full-face appearance-based gaze estimation". In *Proc. IEEE CVPRW*, 2017.