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

Cristina Palmero[1,2], Javier Selva[1], Mohammad Ali Bagheri[3,4], and Sergio Escalera[1,2]

[1] *Dept. Mathematics and Informatics, Universitat de Barcelona, Spain*
[2] *Computer Vision Center, Edifici O, Campus UAB, Bellaterra, Barcelona, Spain*
[3] *Dept. Electrical and Computer Engineering, University of Calgary, Canada*
[4] *Dept. Engineering, University of Larestan, Iran*

Abstract

Gaze behavior is an important non-verbal cue in social signal processing and human-computer interaction. In this paper, we tackle the problem of person- and head pose-independent 3D gaze estimation from remote cameras, using a multi-modal recurrent convolutional neural network (CNN). We propose to combine face, eyes region, and face landmarks as individual streams in a CNN to estimate gaze in still images. Then, we exploit the dynamic nature of gaze by feeding the learned features of all the frames in a sequence to a many-to-one recurrent module that predicts the 3D gaze vector of the last frame. Our multi-modal static solution is evaluated on a wide range of head poses and gaze directions, achieving a significant improvement of 14.6% over the state of the art on EYEDIAP dataset, further improved by 4% when the temporal modality is included.

1 Introduction

Eyes and their movements are considered an important cue in non-verbal behavior analysis, being involved in many cognitive processes and reflecting our internal state (Liversedge & Findlay (2000)). More specifically, eye gaze behavior, as an indicator of human visual attention, has been widely studied to assess communication skills (Rutter & Durkin (1987)) and to identify possible behavioral disorders (Guillon *et al.* (2014)). Therefore, gaze estimation has become an established line of research in computer vision, being a key feature in human-computer interaction (HCI) and usability research (Jacob & Karn (2003); Majaranta & Bulling (2014)).

Recent gaze estimation research has focused on facilitating its use in general everyday applications under real-world conditions, using off-the-shelf remote RGB cameras and removing the need of personal calibration. In this setting, appearance-based methods, which learn a mapping from images to gaze directions, are the preferred choice (Ono *et al.* (2006)). However, they need large amounts of training data to be able to generalize well to in-the-wild situations, which are characterized by significant variability in head poses, face appearances and lighting conditions. In recent years, CNNs have been reported to outperform classical methods. However, most existing approaches have only been tested in restricted HCI tasks, where users look at the screen or mobile phone, showing a low head pose variability. It is yet unclear how these methods would perform in a wider range of head poses.

On a different note, until very recently, the majority of methods only used static eye region appearance as input. State-of-the-art approaches have demonstrated that using the face along with a higher resolution image of the eyes (Krafka *et al.* (2016)), or even just the face itself (Zhang *et al.* (2017)), increases performance. Indeed, the whole-face image encodes more information than eyes alone, such as illumination and head pose. Nevertheless, gaze behavior is not static. Eye and head movements allow us to direct our gaze to target locations of interest. It has been demonstrated that humans can better predict gaze when being shown image sequences of other people moving their eyes (Anderson *et al.* (2016)). However, it is still an open

question whether this sequential information can increase the performance of automatic methods.

In this work, we show that the combination of multiple cues benefits the gaze estimation task. In particular, we use face, eye region and facial landmarks from still images. Facial landmarks model the global shape of the face and come at no cost, since face alignment is a common pre-processing step in many facial image analysis approaches. Furthermore, we present a subject-independent, free-head recurrent 3D gaze regression network to leverage the temporal information of image sequences. The static streams of each frame are combined in a late-fusion fashion using a multi-stream CNN. Then, all feature vectors are input to a many-to-one recurrent module that predicts the gaze vector of the last sequence frame.

In summary, our contributions are two-fold. First, we present a Recurrent-CNN network architecture that combines appearance, shape and temporal information for 3D gaze estimation. Second, we test static and temporal versions of our solution on the EYEDIAP dataset (Funes Mora *et al.* (2014a)) in a wide range of head poses and gaze directions, showing consistent performance improvements compared to related appearance-based methods. To the best of our knowledge, this is the first third-person, remote camera-based approach that uses temporal information for this task. Table 1 outlines our main method characteristics compared to related work.

2 Related work

Gaze estimation methods are typically categorized as model-based or appearance-based (Hansen & Ji (2010); Ferhat & Vilariño (2016); Kar & Corcoran (2017)). **Model-based approaches** use a geometric model of the eye, usually requiring either high resolution images or a person-specific calibration stage to estimate personal eye parameters (Yoo & Chung (2005); Morimoto *et al.* (2002); Venkateswarlu *et al.* (2003); Wood & Bulling (2014); Wang & Ji (2017)). In contrast, **appearance-based methods** learn a direct mapping from intensity images or extracted eye features to gaze directions, thus being potentially applicable to relatively low resolution images and mid-distance scenarios. Different mapping functions have been explored, such as neural networks (Baluja & Pomerleau (1994)), adaptive linear regression (ALR) (Lu *et al.* (2011b)), local interpolation (Tan *et al.* (2002)), gaussian processes (Williams *et al.* (2006); Sugano *et al.* (2013)), random forests (Huang *et al.* (2017); Sugano *et al.* (2014)), or k-nearest neighbors (Wood *et al.* (2016b)). Main challenges of appearance-based methods for 3D gaze estimation are head pose, illumination and subject invariance without user-specific calibration. To handle these issues, some works proposed compensation methods (Lu *et al.* (2011a)) and warping strategies that synthesize a canonical, frontal looking view of the face (Mora & Odobez (2012); Funes-Mora & Odobez (2016); Jeni & Cohn (2016)). Hybrid approaches based on analysis-by-synthesis have also been evaluated (Wood *et al.* (2016a)).

Currently, data-driven methods are considered the state-of-the-art for person- and head pose-independent appearance-based gaze estimation. Consequently, a number of gaze estimation datasets have been introduced in recent years, either in controlled (Smith *et al.* (2013)) or semi-controlled settings (Funes Mora *et al.* (2014b)), in the wild (Zhang *et al.* (2015); Krafka *et al.* (2016)), or consisting of synthetic data (Sugano *et al.* (2014); Wood *et al.* (2015, 2016b)). Zhang *et al.* (2015) showed that CNNs can outperform other mapping methods, using a multi-modal CNN to learn the mapping from 3D head poses and eye images to 3D gaze directions. Krafka *et al.* (2016) proposed a multi-stream CNN for 2D gaze estimation, using individual eye, whole-face image and the face grid as input. As this method was limited to 2D screen mapping, Zhang *et al.* (2017) later explored the potential of just using whole-face images as input to estimate 3D gaze directions. Using a spatial weights CNN, they demonstrated their method to be more robust to facial appearance variation caused by head pose and illumination than eye-only methods. While the method was evaluated in the wild, the subjects were only interacting with a mobile device, thus restricting the head pose range. Deng & Zhu (2017) presented a two-stream CNN to disjointly model head pose from face images and eyeball movement from eye region images. Both were then aggregated into 3D gaze direction using a gaze transform layer. The decomposition was aimed to avoid head-correlation overfitting of previous

Method	3D gaze direction	Unrestricted gaze target	Full face	Eye region	Facial landmarks	Sequential information
Zhang *et al.* (2015)	✓	✗	✗	✓	✗	✗
Krafka *et al.* (2016)	✗	✗	✓	✓	✗	✗
Zhang *et al.* (2017)	✓	✗	✓	✗	✗	✗
Deng & Zhu (2017)	✓	✓	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓	✓

Table 1: Characteristics of recent related work on person- and head pose-independent appearance-based gaze estimation methods using CNNs.

data-driven approaches. They evaluated their approach in the wild with a wider range of head poses, obtaining better performance than previous eye-based methods. However, they did not test it on public annotated benchmark datasets.

In this paper, we propose a multi-stream recurrent CNN network for person- and head pose-independent 3D gaze estimation for a mid-distance scenario. We evaluate it on a wider range of head poses and gaze directions than screen-targeted approaches. As opposed to previous methods, we also rely on temporal information inherent in sequential data.

3 Methodology

In this section, we present our approach for 3D gaze regression based on appearance and shape cues for still images and image sequences. First, we introduce the data modalities and formulate the problem. Then, we detail the normalization procedure prior to the regression stage. Finally, we explain the global network topology as well as the implementation details. An overview of the system architecture is depicted in Figure 1.

3.1 Multi-modal gaze regression

Let us represent gaze direction as a 3D unit vector $\mathbf{g} = [g_x, g_y, g_z]^T \in \mathbb{R}^3$ in the Camera Coordinate System (CCS), whose origin is the central point between eyeball centers. Assuming a calibrated camera, and a known head position and orientation, our goal is to estimate $\mathbf{g}$ from a sequence of images $\{\mathbf{I}^{(i)} \mid \mathbf{I} \in \mathbb{R}^{W \times H \times 3}\}$ as a regression problem.

Gazing to a specific target is achieved by a combination of eye and head movements, which are highly coordinated. Consequently, the apparent direction of gaze is influenced not only by the location of the irises within the eyelid aperture, but also by the position and orientation of the face with respect to the camera. Known as the Wollaston effect (Wollaston *et al.* (1824)), the exact same set of eyes may appear to be looking in different directions due to the surrounding facial cues. It is therefore reasonable to state that eye images are not sufficient to estimate gaze direction (Zhang *et al.* (2017); Krafka *et al.* (2016)). Instead, whole-face images can encode head pose or illumination-specific information across larger image areas than those available just in the eye region.

The drawback of appearance-only methods is that global structure information is not explicitly considered. In that sense, facial landmarks can be used as global shape cues to encode spatial relationships and geometric constraints. Current state-of-the-art face alignment approaches are robust enough to handle large appearance variability, extreme head poses and occlusions, being especially useful when the dataset used for gaze estimation does not contain such variability. Facial landmarks are mainly correlated with head orientation, eye position, eyelid openness, and eyebrow movement, which are valuable features for our task.

Therefore, in our approach we jointly model appearance and shape cues (see Figure 1). The former is represented by a whole-face image $\mathbf{I}_F$, along with a higher resolution image of the eyes $\mathbf{I}_E$ to identify subtle changes. Due to dealing with wide head pose ranges, some eye images may not depict the whole eye, containing mostly background or other surrounding facial parts instead. For that reason, and contrary to previous approaches that only use one eye image (Zhang *et al.* (2015); Sugano *et al.* (2014)), we use a single image composed of two patches of centered left and right eyes. Finally, the shape

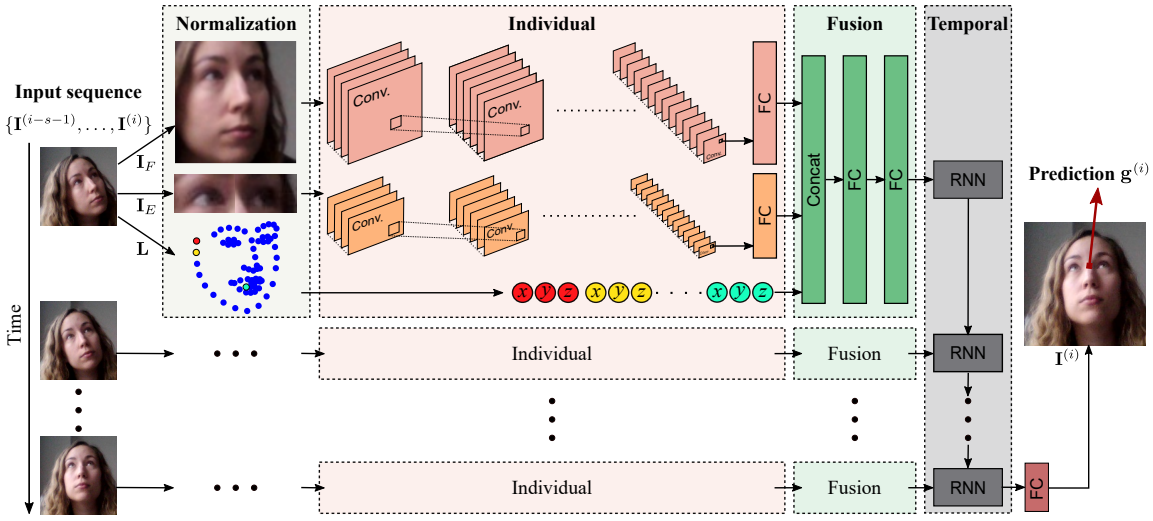


Figure 1: Overview of the proposed network. A multi-stream CNN jointly models full-face, eye region appearance and face landmarks from still images. The combined extracted features from each frame are fed into a recurrent module to predict last frame's gaze direction.

cue is represented by 3D face landmarks obtained from a 68-landmark model, denoted by $\mathbf{L} = \{(l_x, l_y, l_z)_c \mid \forall c \in [1, ..., 68]\}$.

In this work we also consider the dynamic component of gaze. We leverage the sequential information of eye and head movements such that, given appearance and shape features of consecutive frames, it is possible to better predict the gaze direction of the current frame. Therefore, the 3D gaze estimation task for a 1-frame sequence is formulated as $\mathbf{g}^{(i)} = f\left(\{\mathbf{I}_F{}^{(i)}\}, \{\mathbf{I}_E{}^{(i)}\}, \{\mathbf{L}^{(i)}\}\right)$, where i denotes the i-th frame, and f is the regression function.

3.2 Data normalization

Prior to gaze regression, a normalization step in the 3D space and the 2D image, similar to (Sugano *et al.* (2014)), is carried out. This is performed to reduce the appearance variability and to allow the gaze estimation model to be applied regardless of the original camera configuration.

Let $\mathbf{H} \in \mathbb{R}^{3x3}$ be the head rotation matrix, and $\mathbf{p} = [p_x, p_y, p_z]^T \in \mathbb{R}^3$ the reference location with respect to the original CCS. The goal is to find the conversion matrix $\mathbf{M} = \mathbf{SR}$ such that (a) the X-axes of the virtual camera and the head become parallel using the rotation matrix $\mathbf{R}$, and (b) the virtual camera looks at the reference location from a fixed distance d_n using the Z-direction scaling matrix $\mathbf{S} = diag(1, 1, d_n/\|\mathbf{p}\|)$. $\mathbf{R}$ is computed as $\mathbf{a} = \hat{\mathbf{p}} \times \mathbf{H}^T\mathbf{e_1}$, $\mathbf{b} = \hat{\mathbf{a}} \times \hat{\mathbf{p}}$, $\mathbf{R} = [\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{\mathbf{p}}]^T$, where $\mathbf{e_1}$ denotes the first orthonormal basis and $\langle\,\hat{\cdot}\,\rangle$ is the unit vector.

This normalization translates into the image space as a centered cropped image patch of size $W_n \times H_n$ where head roll rotation has been removed. This is done by applying a perspective warping to the input image $\mathbf{I}$ using the transformation matrix $\mathbf{W} = \mathbf{C}_o\mathbf{M}{\mathbf{C}_n}^{-1}$, where $\mathbf{C}_o$ and $\mathbf{C}_n$ are the original and virtual camera matrices, respectively.

The 3D gaze vector is also normalized as $\mathbf{g}_n = \mathbf{Rg}$. After image normalization, the line of sight can be represented in a 2D space. Therefore, $\mathbf{g}_n$ is further transformed to spherical coordinates (θ, ϕ) assuming unit length, where θ and ϕ denote the horizontal and vertical direction angles, respectively. This 2D angle representation, delimited in the range $[-\pi/2, \pi/2]$, is computed as $\theta = \arctan(g_x/g_z)$ and $\phi = \arcsin(-g_y)$, such that $(0, 0)$ represents looking straight ahead to the CCS origin.

3.3 Recurrent Convolutional Neural Network

We propose a Recurrent CNN Regression Network for 3D gaze estimation. The network is divided in 3 modules: (1) *Individual*, (2) *Fusion*,

and (3) *Temporal*.

First, the *Individual* module learns features from each appearance cue separately. It consists of a two-stream CNN, one devoted to the normalized face image stream and the other to the joint normalized eyes image. Next, the *Fusion* module combines the extracted features of each appearance stream in a single vector along with the normalized landmark coordinates. Then, it learns a joint representation between modalities in a late-fusion fashion. Both *Individual* and *Fusion* modules, further referred to as *Static* model, are applied to each frame of the sequence. Finally, the resulting feature vectors of each frame are input to the *Temporal* module based on a many-to-one recurrent network. This module leverages sequential information to predict the normalized 2D gaze angles of the last frame of the sequence using a linear regression layer added on top of it.

3.4 Implementation details

3.4.1 Network details

Each stream of the *Individual* module is based on the VGG-16 deep network (Parkhi *et al.* (2015)), consisting of 13 convolutional layers, 4 max pooling layers, and 1 fully connected (FC) layer with Rectified Linear Unit (ReLU) activations. The full-face stream follows the same configuration as the base network, having an input of 224×224 pixels and a 4096D FC layer. In contrast, the input joint eye image is smaller, with a final size of 120×48 pixels, so the number of parameters is decreased proportionally. In this case, its last FC layer produces a 1536D vector. A 204D landmark coordinates vector is concatenated to the output of the FC layer of each stream, resulting in a 5836D feature vector. Consequently, the *Fusion* module consists of 2 5836D FC layers with ReLU activations and 2 dropout layers between FCs as regularization. Finally, to model the temporal dependencies, we use a single GRU layer with 128 units.

The network is trained in a stage-wise fashion. First, we train the *Static* model and the final regression layer end-to-end on each individual frame of the training data. The convolutional blocks are pre-trained with the VGG-Face dataset (Parkhi *et al.* (2015)), whereas the FCs are trained from scratch. Second, the training data is re-arranged by means of a sliding window with stride 1 to build input sequences. Each sequence is composed of $s = 4$ consecutive frames, whose gaze direction target is the gaze direction of the last frame of the sequence $\left(\{\mathbf{I}^{(i-s-1)}, \ldots, \mathbf{I}^{(i)}\},\ \mathbf{g}^{(i)}\right)$. Using this re-arranged training data, we extract features of each frame of the sequence from a frozen *Individual* module, fine-tune the *Fusion* layers, and train both, the *Temporal* module and a new final regression layer from scratch. This way, the network can exploit the temporal information to further refine the fusion weights.

We trained the model[1] using ADAM optimizer with an initial learning rate of 0.0001, dropout of 0.3, and batch size of 64 frames. The number of epochs was experimentally set to 21 for the first training stage and 10 for the second. We use the average Euclidean distance between the predicted and ground-truth 3D gaze vectors as loss function.

3.4.2 Input pre-processing

For this work we use head pose and eye locations in the 3D scene provided by the dataset. The 3D landmarks are extracted using the state-of-the-art method of Bulat & Tzimiropoulos (2017), which is based on stacked hourglass networks (Newell *et al.* (2016)).

During training, the original image is pre-processed to get the two normalized input images. The normalized whole-face patch is centered 0.1 meters ahead of the head center in the head coordinate system, and $\mathbf{C}_n$ is defined such that the image has size of 250×250 pixels. The difference between this size and the final input size allows us to perform random cropping and zooming to augment the data (explained in Section 4.1). Similarly, each normalized eye patch is centered in their respective eye center locations. In this case, the virtual camera matrix is defined so that the image is cropped to 70×58, while in practice the final patches have size of 60×48. Landmarks are normalized using the same procedure and further pre-processed with mean subtraction and min-max normalization per axis. Finally, we divide them by a scaling factor w such that all coordinates are in the range $[0, w]$. This way, all concatenated feature values are in a similar range. After inference, the predicted normalized 2D angles are de-normalized back to

[1]Code and models will be public available after publication of the paper.

the original 3D space.

4 Experiments

In this section, we evaluate the cross-subject 3D gaze estimation task on a wide range of head poses and gaze directions. Furthermore, we validate the effectiveness of the proposed architecture comparing both static and temporal approaches. We report the error in terms of mean angular error between predicted and ground-truth 3D gaze vectors. Note that due to the requirements of the temporal model not all the frames obtain a prediction. Therefore, for a fair comparison, the reported results for static models disregard such frames when temporal models are included in the comparison.

4.1 Training data

There are few publicly available datasets devoted to 3D gaze estimation and most of them focus on HCI with a limited range of head pose and gaze directions. Therefore, we use VGA videos from the publicly-available EYEDIAP dataset (Funes Mora *et al.* (2014a)) to perform the experimental evaluation, as it is the only one containing video sequences with a wide range of head poses and showing the full face. This dataset consists of 3-minute videos of 16 subjects looking at two types of targets: continuous *screen* targets on a fixed monitor (CS), and *floating* physical targets (FT). The videos are further divided into *static* (S) and *moving* (M) head pose for each of the subjects. Subjects 12-16 were recorded with 2 different lighting conditions.

For evaluation, we filtered out those frames that fulfilled at least one of the following conditions: (1) face or landmarks not detected; (2) subject not looking at the target; (3) 3D head pose, eyes or target location not properly recovered; and (4) eyeball rotations violating physical constraints ($|\theta| \leq 40°$, $|\phi| \leq 30°$, MSC (2000)). Note that we purposely do not filter eye blinking moments to learn their dynamics with the temporal model, which may produce some outliers with a higher prediction error due to a less accurate ground truth. Figure 2 shows the distribution of gaze directions and head poses for both filtered CS and FT cases.

We applied data augmentation to the training set with the following random transformations: horizontal flip, shifts of up to 5 pixels, zoom of up to 2%, brightness changes by a factor in the range $[0.4, 1.75]$, and additive Gaussian noise with $\sigma^2 = 0.03$.

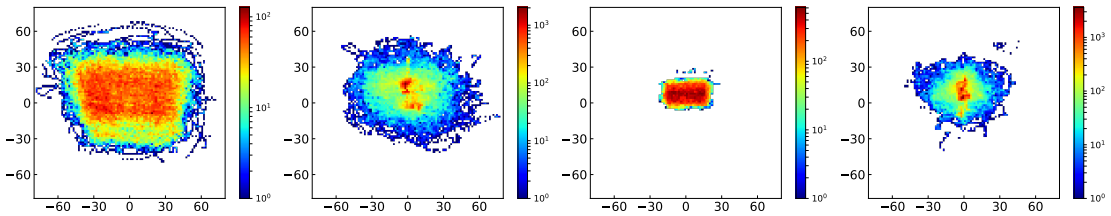
4.2 Evaluation of static modalities

First, we evaluate the contribution of each static modality on the FT scenario. We divided the 16 participants into 4 groups, such that appearance variability was maximized while maintaining a similar number of training samples per group. Each static model was trained end-to-end performing 4-fold cross-validation using different combinations of input modalities. Since the number of fusion units depends on the number of input modalities, we also compare different fusion layer sizes. The effect of data normalization is also evaluated by training a not-normalized face model where the input image is the face bounding box with square size the maximum distance between 2D landmarks.

As shown in Figure 3, all models that take normalized full-face information as input achieve better performance than the eyes-only model. More specifically, the combination of face, eyes and landmarks outperforms all the other combinations by a small but significant margin (paired Wilcoxon test, $p < 0.0001$). The standard deviation of the best-performing model is reduced compared to the face and eyes model, suggesting a regularizing effect due to the addition of landmarks. The not-normalized face-only model shows the largest error, proving the impact of normalization to reduce the appearance variability. Furthermore, our results indicate that the increase of fusion units is not correlated with a better performance.

4.3 Static gaze regression: comparison with existing methods

We compare our best-performing static model with three baselines. **Head.** Treating the head pose directly as gaze direction. **PR-ALR.** Method that relies on RGB-D data to rectify the eye images viewpoint into a canonical head pose using a 3DMM. It then learns an RGB gaze appearance model using ALR (Mora & Odobez (2012)). Predicted 3D vectors for *FT-S* scenario are provided by EYEDIAP dataset. **MPIIGaze.** State-of-the-art full-face 3D gaze estimation method (Zhang *et al.* (2015)). They



(a) **g** (*FT*) (b) **h** (*FT*) (c) **g** (*CS*) (d) **h** (*CS*)

Figure 2: Ground-truth eye gaze **g** and head orientation **h** distribution on the filtered EYEDIAP dataset for *CS* and *FT* settings, in terms of x- and y- angles.

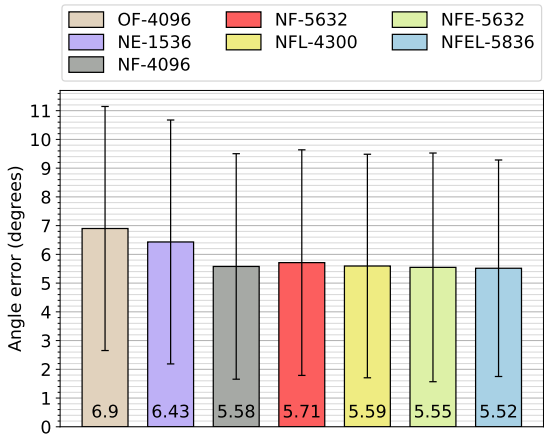


Figure 3: Performance evaluation of the *Static* network using different input modalities (*O - Not normalized, N - Normalized, F - Face, E - Eyes, L - 3D Landmarks*) and size of fusion layers on the *FT* scenario.

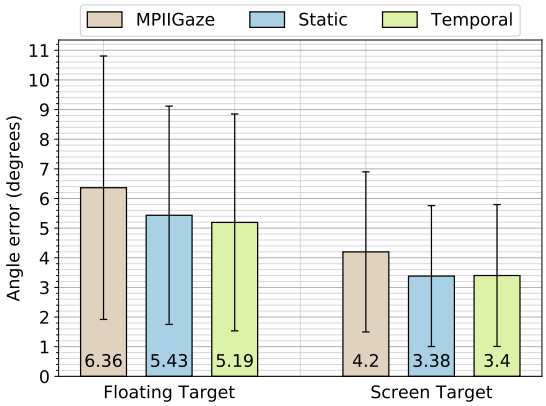


Figure 4: Performance comparison among MPIIGaze method (Zhang *et al.* (2015)) and our *Static* and *Temporal* versions of the proposed network for *FT* and *CS* scenarios.

use an Alexnet-based CNN model with spatial weights to enhance information in different facial regions. We fine-tuned it with the filtered EYEDIAP subsets using our training parameters and normalization procedure.

In addition to the aforementioned *FT*-based evaluation setup, we also evaluate our method on the *CS* scenario. In this case there are only 14 participants available, so we divided them in 5 groups and performed 5-fold cross-validation. In Figure 4 we compare our method to MPIIGaze, achieving a statistically significant improvement of 14.6% and 19.5% on *FT* and *CS* scenarios, respectively (paired Wilcoxon test, $p < 0.0001$). We can observe that a restricted gaze target benefits the performance of all methods, compared to a more challenging unrestricted setting with a wider range of head poses and gaze directions.

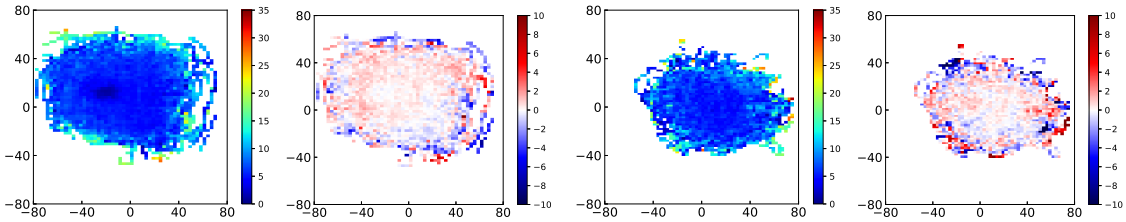
Table 2 provides a detailed comparison on every participant, performing leave-one-out cross-validation on the *FT* scenario for *static* and *moving* head separately. Results show that, as expected, facial appearance and head pose have a noticeable impact on gaze accuracy, with average error differences of up to 7.7° among participants.

4.4 Evaluation of the temporal network

In this section, we evaluate the contribution of adding the temporal module to the static model. To do so, we trained a lower-dimensional version of the static network with comparable performance to the original, reducing the number of units of the second fusion layer to 2918. Re-

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	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 2: Gaze angular error comparison for *static* (top half) and *moving* (bottom half) head pose for each subject in the *FT* scenario. Best results in bold.



(a) Gaze space (b) Head orientation space

Figure 5: Angular error distribution across gaze (a) and head orientation (b) spaces in the *FT* setting, in terms of x- and y- angles. For each space, we depict the *Static* model performance (left) and the contribution of the *Temporal* model versus *Static* (right). In the latter, positive difference means higher improvement of the *Temporal* model.

sults are reported in Figure 4 and Table 2. One can observe that using sequential information is helpful on the *FT* scenario, outperforming the static model by a statistically significant 4.4% (paired Wilcoxon test, $p < 0.0001$). This contribution is more noticeable in the *moving* head setting, proving that the temporal model can benefit from head motion information. In contrast, such information seems to be less meaningful in the *CS* scenario, where the obtained error is already very low for a cross-subject setting and the amount of head movement declines.

Figure 5 further explores the error distribution of the static network and the impact of sequential information. We can observe that the accuracy of the static model drops with extreme head poses and gaze directions, which can also be correlated to having less data in those areas. Compared to the static model, the temporal model particularly benefits gaze targets from mid-range upwards. Its contribution is less clear for extreme targets, probably again due to data imbalance.

Finally, we evaluated the effect of different recurrent architectures for the temporal model. In particular, we tested 1 (128 units) and 2 (256-128 units) LSTM and GRU layers, with 1 GRU layer obtaining slightly superior results (up to 0.12°). We also assessed the effect of sequence length fixing s in the range $\{4, 7, 10\}$, with $s = 7$ performing worse than the other two (up to 0.14°).

5 Conclusions

In this work, we studied the combination of full-face and eye images along with facial landmarks for person- and head pose-independent 3D gaze estimation. Consequently, we proposed a multi-stream recurrent CNN network that leverages the sequential information of eye and head movements. Both static and temporal versions of our approach significantly outperform current state-of-the-art 3D gaze estimation methods on a wide range of head poses and gaze directions. We showed that adding geometry features to appearance-based methods has a regularizing effect on the accuracy. Adding sequential information further benefits the final performance compared to static-only input, especially from mid-range upwards and in those cases where head motion is present. The effect in very extreme head

poses is not clear due to data imbalance, suggesting the importance of learning from a continuous, balanced dataset including all head poses and gaze directions of interest. To the best of our knowledge, this is the first attempt to exploit the temporal modality in the context of gaze estimation from remote cameras. As future work, we will further explore extracting meaningful temporal representations of gaze dynamics, considering 3DCNNs as well as the encoding of deep features around particular tracked face landmarks (Jung *et al.* (2015)).

Acknowledgements

This work has been partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER, UE), CERCA Programme / Generalitat de Catalunya, and the FP7 people program (Marie Curie Actions), REA grant agreement no FP7-607139 (iCARE — improving Children Auditory REhabilitation). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

References

Anderson, Nicola C, Risko, Evan F, & Kingstone, Alan. 2016. Motion influences gaze direction discrimination and disambiguates contradictory luminance cues. *Psychonomic bulletin & review*, **23**(3), 817–823.

Baluja, Shumeet, & Pomerleau, Dean. 1994. Non-intrusive gaze tracking using artificial neural networks. *Pages 753–760 of: Advances in Neural Information Processing Systems*.

Bulat, Adrian, & Tzimiropoulos, Georgios. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). *In: International Conference on Computer Vision*.

Deng, Haoping, & Zhu, Wangjiang. 2017. Monocular Free-head 3D Gaze Tracking with Deep Learning and Geometry Constraints. *Pages 3162–3171 of: Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE.

Ferhat, Onur, & Vilariño, Fernando. 2016. Low cost eye tracking. *Computational intelligence and neuroscience*, **2016**, 17.

Funes-Mora, Kenneth A, & Odobez, Jean-Marc. 2016. Gaze estimation in the 3D space using RGB-D sensors. *International Journal of Computer Vision*, **118**(2), 194–216.

Funes Mora, Kenneth Alberto, Monay, Florent, & Odobez, Jean-Marc. 2014a. EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. *In: Proceedings of the ACM Symposium on Eye Tracking Research and Applications*. ACM.

Funes Mora, Kenneth Alberto, Monay, Florent, & Odobez, Jean-Marc. 2014b. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. *Pages 255–258 of: Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM.

Guillon, Quentin, Hadjikhani, Nouchine, Baduel, Sophie, & Rogé, Bernadette. 2014. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, **42**, 279–297.

Hansen, Dan Witzner, & Ji, Qiang. 2010. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, **32**(3), 478–500.

Huang, Qiong, Veeraraghavan, Ashok, & Sabharwal, Ashutosh. 2017. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, **28**(5-6), 445–461.

Jacob, Robert JK, & Karn, Keith S. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Pages 573–605 of: The mind's eye*. Elsevier.

Jeni, László A, & Cohn, Jeffrey F. 2016. Person-independent 3d gaze estimation using face frontalization. *Pages 87–95 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Jung, Heechul, Lee, Sihaeng, Yim, Junho, Park, Sunjeong, & Kim, Junmo. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. *Pages 2983–2991 of: Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE.

Kar, Anuradha, & Corcoran, Peter. 2017. A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms. *IEEE Access*, **5**, 16495–16519.

Krafka, Kyle, Khosla, Aditya, Kellnhofer, Petr, Kannan, Harini, Bhandarkar, Suchendra, Matusik, Wojciech, & Torralba, Antonio. 2016. Eye Tracking for Everyone. *Pages 2176–2184 of: Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE.

Liversedge, Simon P, & Findlay, John M. 2000. Saccadic eye movements and cognition. *Trends in cognitive sciences*, **4**(1), 6–14.

Lu, Feng, Okabe, Takahiro, Sugano, Yusuke, & Sato, Yoichi. 2011a. A head pose-free approach for appearance-based gaze estimation. *Pages 1–11 of: BMVC*.

Lu, Feng, Sugano, Yusuke, Okabe, Takahiro, & Sato, Yoichi. 2011b. Inferring human gaze from appearance via adaptive linear regression. *Pages 153–160 of: Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE.

Majaranta, Päivi, & Bulling, Andreas. 2014. Eye tracking and eye-based human–computer interaction. *Pages 39–65 of: Advances in physiological computing*. Springer.

Mora, Kenneth Alberto Funes, & Odobez, Jean-Marc. 2012. Gaze estimation from multimodal kinect data. *Pages 25–30 of: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE.

Morimoto, Carlos Hitoshi, Amir, Arnon, & Flickner, Myron. 2002. Detecting eye position and gaze from a single camera and 2 light sources. *Pages 314–317 of: Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4. IEEE.

MSC, IMO. 2000. *Circ. 982 Guidelines on ergonomic criteria for bridge equipment and layout*.

Newell, Alejandro, Yang, Kaiyu, & Deng, Jia. 2016. Stacked hourglass networks for human pose estimation. *Pages 483–499 of: European Conference on Computer Vision*. Springer.

Ono, Yasuhiro, Okabe, Takahiro, & Sato, Yoichi. 2006. Gaze estimation from low resolution images. *Pages 178–188 of: Pacific-Rim Symposium on Image and Video Technology*. Springer.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. 2015. Deep Face Recognition. *In: British Machine Vision Conference*.

Rutter, Derek R, & Durkin, Kevin. 1987. Turn-taking in mother–infant interaction: An examination of vocalizations and gaze. *Developmental psychology*, **23**(1), 54.

Smith, Brian A, Yin, Qi, Feiner, Steven K, & Nayar, Shree K. 2013. Gaze locking: passive eye contact detection for human-object interaction. *Pages 271–280 of: Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM.

Sugano, Yusuke, Matsushita, Yasuyuki, & Sato, Yoichi. 2013. Appearance-based gaze estimation using visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, **35**(2), 329–341.

Sugano, Yusuke, Matsushita, Yasuyuki, & Sato, Yoichi. 2014. Learning-by-synthesis for appearance-based 3d gaze estimation. *Pages 1821–1828 of: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE.

Tan, Kar-Han, Kriegman, David J, & Ahuja, Narendra. 2002. Appearance-based eye gaze estimation. *Pages 191–195 of: Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*. IEEE.

Venkateswarlu, Ronda, *et al.* 2003. Eye gaze estimation from a single image of one eye. *Pages 136–143 of: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE.

Wang, Kang, & Ji, Qiang. 2017. Real Time Eye Gaze Tracking with 3D Deformable Eye-Face Model. *Pages 1003–1011 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Williams, Oliver, Blake, Andrew, & Cipolla, Roberto. 2006. Sparse and Semi-supervised Visual Mapping with the Sˆ 3GP. *Pages 230–237 of: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE.

Wollaston, William Hyde, *et al.* 1824. XIII. On the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*, **114**, 247–256.

Wood, Erroll, & Bulling, Andreas. 2014. Eyetab: Model-based gaze estimation on unmodified tablet computers. *Pages 207–210 of: Proceedings of the Symposium on Eye Tracking Research and Applications.* ACM.

Wood, Erroll, Baltrusaitis, Tadas, Zhang, Xucong, Sugano, Yusuke, Robinson, Peter, & Bulling, Andreas. 2015. Rendering of eyes for eye-shape registration and gaze estimation. *Pages 3756–3764 of: Proceedings of the IEEE International Conference on Computer Vision.*

Wood, Erroll, Baltrušaitis, Tadas, Morency, Louis-Philippe, Robinson, Peter, & Bulling, Andreas. 2016a. A 3d morphable eye region model for gaze estimation. *Pages 297–313 of: European Conference on Computer Vision.* Springer.

Wood, Erroll, Baltrušaitis, Tadas, Morency, Louis-Philippe, Robinson, Peter, & Bulling, Andreas. 2016b. Learning an appearance-based gaze estimator from one million synthesised images. *Pages 131–138 of: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications.* ACM.

Yoo, Dong Hyun, & Chung, Myung Jin. 2005. A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Computer Vision and Image Understanding*, **98**(1), 25–51.

Zhang, Xucong, Sugano, Yusuke, Fritz, Mario, & Bulling, Andreas. 2015. Appearance-based gaze estimation in the wild. *Pages 4511–4520 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Zhang, Xucong, Sugano, Yusuke, Fritz, Mario, & Bulling, Andreas. 2017. It's written all over your face: Full-face appearance-based gaze estimation. *In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*