Deep Multitask Gaze Estimation with a Constrained Landmark-Gaze Model

Yu Yu[1,2(✉)], Gang Liu[1], and Jean-Marc Odobez[1,2]

[1] Idiap Research Institute, Martigny, Switzerland
{yyu,gang.liu,odobez}@idiap.ch
[2] EPFL, Lausanne, Switzerland

Abstract. As an indicator of attention, gaze is an important cue for human behavior and social interaction analysis. Recent deep learning methods for gaze estimation rely on plain regression of the gaze from images without accounting for potential mismatches in eye image cropping and normalization. This may impact the estimation of the implicit relation between visual cues and the gaze direction when dealing with low resolution images or when training with a limited amount of data. In this paper, we propose a deep multitask framework for gaze estimation, with the following contributions. (i) we proposed a multitask framework which relies on both synthetic data and real data for end-to-end training. During training, each dataset provides the label of only one task but the two tasks are combined in a constrained way. (ii) we introduce a Constrained Landmark-Gaze Model (CLGM) modeling the joint variation of eye landmark locations (including the iris center) and gaze directions. By relating explicitly visual information (landmarks) to the more abstract gaze values, we demonstrate that the estimator is more accurate and easier to learn. (iii) by decomposing our deep network into a network inferring jointly the parameters of the CLGM model and the scale and translation parameters of eye regions on one hand, and a CLGM based decoder deterministically inferring landmark positions and gaze from these parameters and head pose on the other hand, our framework decouples gaze estimation from irrelevant geometric variations in the eye image (scale, translation), resulting in a more robust model. Thorough experiments on public datasets demonstrate that our method achieves competitive results, improving over state-of-the-art results in challenging free head pose gaze estimation tasks and on eye landmark localization (iris location) ones.

1 Introduction

Gaze is the essential indicator of human attention and can even provide access to thought processes [1–3]. In interactions, it is a non-verbal behavior that plays a major role in all communication aspects [4], and it has also been shown to be related to higher-level constructs, like personality, dominance, or rapport. Gaze is thus an important cue for human behavior analysis, and beyond traditional screen-gazing monitoring, 3D gaze estimation finds application in health

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 456–474, 2019.
https://doi.org/10.1007/978-3-030-11012-3_35

care [5], social interaction analysis [6], human computer interaction (HCI) or human robotic interaction (HRI) [7,8]. In another context, the new generation of smart phones like iPhone X and their extended applications raise further interest in gaze estimation under mobile scenarios [9–11].

Traditional gaze estimation methods include model-based geometrical methods and appearance based methods. The former are more accurate, but the techniques used so far to extract eye landmarks (eye corners, iris) require high resolution images (limiting the freedom of motion) and relatively open eyes since the gaze is estimated from sparse features which are most often detected in a separate task. The latter methods have been shown to be more robust to eye resolution or gazing direction (*e.g.* looking down with eyelid occlusion) variabilities. Thus, this is not surprising that recent works have explored inferring gaze from the eye image via deep regression [9,12–14]. Nevertheless, although progress has been reported, direct regression of gaze still suffers from limitations:

- Since the ground truth of gaze vector is hard to annotate, the amount of training data for gaze estimation is limited (number of people, illumination conditions, annotation accuracies) compared to other computer vision tasks. Although there has been some synthetic data [15] for gaze estimation, the appearance and gaze setting of synthetic data is somehow different to real data. Therefore, this currently hinders the benefits of deep learning for gaze.
- An accurate and unified eye cropping is difficult to achieve in real application. This means the size and location of the eye regions may significantly vary in the cropped eye images, due to bad eye/landmark localization, or when changing datasets. Since the gaze estimation is very sensitive to the subtle relative positions and shapes of eye landmarks, such variations can significantly alter the gaze estimation outcomes. Though data augmentation can partially handle this problem, an explicit model of this step may improve the generalization ability to new datasets, unperfect cropping, or new eyes.

To address these issues, we propose an end-to-end trainable deep multitask framework based on a Constrained Landmark-Gaze Model, with the following properties.

First, we address eye landmark (including iris center) detection and gaze estimation jointly. Indeed, since gaze values are strongly correlated with eye landmark locations, we hypothesize that modeling eye landmark detection (which is an explicit visual task) as an auxiliary task can ease the learning of a predictive model of the more abstract gaze information. To the best of our knowledge, this is the first time that multitask learning is applied to gaze estimation. Since there is no existing large scale dataset which annotates detailed eye landmarks, we rely on a synthetic dataset for the learning of the auxiliary task in this paper. Note that we only use the landmark annotations from the synthetic data because of the different gaze setting of synthetic data. The use of synthetic data also expands the amount of training data to some extent.

Second, instead of predicting eye landmarks and gaze in two network branches as in usual deep multitask learning, we build a Constrained Landmark-Gaze

Model (CLGM) modeling the joint variation of eye landmark location and gaze direction, which bridges the two tasks in a closer and more explicit way.

Third, we make our approach more robust to scale, translation and even head pose variations by relying on a deterministic decoder. More precisely, the network learns two sets of parameters, which are the coefficients of the CLGM model, and the scale and translation parameters defining the eye region. Using these parameters and the head pose, the decoder deterministically predicts the eye landmark locations and gaze via the CLGM. Note however that while all parameters account for defining the landmark positions, only the CLGM coefficients and the head pose are used for gaze prediction. Thus, gaze estimation is decoupled from irrelevant variations in scale and translation and geometrically modeled within the head pose frame.

Finally, note that while currently landmark detection is used as a secondary task, it could be used as a primary task as well to extract the features (eye corners, iris center) requested by a geometrical eye gaze model, which can potentially be more accurate. In particular, the CLGM could help predicting iris location even when the eyes are not fully open (see Fig. 7 for examples).

Thus, in summary, our contributions are as follows:

- A Constrained Landmark-Gaze Model modeling the joint variation of eye landmarks and gaze;
- Gaze estimation robust to translation, scale and head pose achieved by a CLGM based decoder;
- An end-to-end trainable deep multitask learning framework for gaze estimation with the help of CLGM and synthetic data.

Thorough experiments on public datasets for both gaze estimation and landmark (iris) localization demonstrate the validity of our approach.

The rest of the paper is organized as follows. We introduce related works in Sect. 2. The correlation between eye landmarks and gaze is studied in Sect. 3. The proposed method is presented in Sect. 4, while experimental results are reported in Sect. 5.

2 Related Work

We introduce the current related researches in gaze estimation and multitask learning as follows.

Gaze Estimation. In this paper, we mainly investigated the vision based non-invasive and non-active (i.e. without infra-red sources) remote gaze estimation methods. They can be grouped into two categories, the geometric based methods (GBM) and appearance based methods (ABM) [16].

GBM methods rely on a geometric model of the eye whose parameters (like eye ball center and radius, pupil center and radius) can be estimated from features extracted in training images [17–25] and can further be used to infer the gaze direction. They usually require high resolution eye images from near frontal

head poses to obtain stable and accurate feature detection, which limits the user mobility and their application to many settings of interest.

By learning a mapping from the eye appearance to the gaze, ABM methods [11,26] are more robust to lower resolution images. They usually extract visual features like Retinex feature [27] and mHOG [28], and train regression models such as Random Forest [29], Adaptive Linear Regression [30], Support Vector Regression [28] for gaze estimation. Very often ABM methods assumed static head pose, but recently head pose dependent image normalization have been tested with success, as done for instance in [31] where a 3D morphable model is used for pose normalization. In general, ABM methods require a large amount of data for training and they do not model the person gaze variation explicitly with a model.

Recent works started to use deep learning to regress gaze directly from the eye image [9,12–14,32]. Krafka et al. [9] proposed to learn the gaze fixation point on smart phone screens using the images taken from the front-facing camera. Their network takes 4 channels including full face image and eye images as input. To train their network, they collected a large dataset with 2.5M frames and an accurate estimator with 2 cm error is achieved. Nevertheless, this dataset does not provide the groundtruth for the 3D gaze direction, which is much harder to label than 2D gaze fixation point. Zhang et al. [13] proposed a dataset for 3D gaze estimation. However, the number of participants is much smaller, which may limit model generalization when being used for training. In Zhang's work, the head pose was linearly modeled when estimating gaze. This design is challenged in [12] where the correlation between the gaze and head pose is explicitly taken into account, which may reflect prior information between these two quantities, but does not account for eye shape or eye cropping variabilities.

The above works basically regress gaze directly from the eye appearances or full faces. In contrast, several methods [33–35] attempt to learn gaze via some intermediate representations or features, including eye landmarks [34,35]. A similar approach to this paper is proposed in [33] where the network learns the heatmaps of eye landmarks first. The gaze is then predicted based on the landmark heatmaps. Our paper differs from this work in that the eye landmarks and gaze are jointly modeled with a constrained framework and they are predicted in the same stage.

Multitask Learning. Multitask learning aims to improve the overall performance of one or each task by providing implicit data augmentation or regularizations [36]. Due to the flexibility of network architectures, a number of works have been proposed on deep multitask learning. The classical implementation is to share parameters in shallow layers and arrange task-specific branches in deeper layers. Many of the representative works are face-related research [37–41] since there are plenty of datasets with rich annotations in this area and the face attributes are also well correlated. Some other works, however, attempted to propose novel multitask learning architectures which could generalize well on other tasks. For example, the Cross-stitch Network [42] designed a cross-stitch unit to leverage the activations from multiple models thus the parameters are

shared softly. However, the network architecture and the placing of cross-stitch are still manually determined. Instead of hand designing the multitask learning architecture, Lu et al. [43] proposed to dynamically create network branches for tasks during training so fully adaptive feature sharing is achieved. Nevertheless, their approach did not model the interactions between tasks.

To the best of our knowledge, we are not aware of any multi-task learning framework for gaze estimation.

3 Correlation of Eye Landmarks and Gaze

Before introducing our method, we first study the correlation existing between the gaze and eye landmarks. We used the synthetic database UnityEyes [15] for correlation analysis since this database provides rich and accurate information regarding the landmark and gaze values. As this dataset relies on a synthetic yet realistic model of the eye ball and shape, and since this database has been used for the training of gaze estimators which have achieved very high performance on real datasets [15], we expect this correlation analysis to be rather accurate. In any case, in Sect. 4.3, we show how we can account for the discrepancy between the synthetic model and real data.

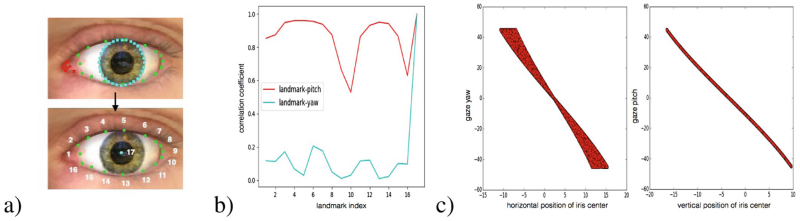

Fig. 1. Correlation between the eye landmark positions and gaze values, computed from the UnityEyes dataset. (a) selected landmarks (bottom) from the UnityEyes landmark set (top). (b) correlation coefficients between the landmark horizontal or vertical positions and the gaze yaw or pitch angles. (c) joint distribution map of the horizontal or vertical positions of the iris center and of the yaw or pitch gaze angles. (Color figure online)

Landmark Set. The UnityEyes annotates three types of eye landmarks, the caruncle landmarks, the eyelid landmarks and the iris landmarks, as shown in the first row of Fig. 1a. Considering that relying on many landmarks will not help improving the robustness and accuracy but simply increase the complexity of the method, we only selected a subset $\mathcal{I}$ of the available landmark instead. It contains 16 landmarks from the eyelid, and the iris center which is estimated from the iris contour landmarks. This is illustrated in the second row of Fig. 1a.

Landmark Alignment. We generated 50,000 UnityEyes samples with frontal head pose and a gaze value uniformly sampled within the $[-45°, 45°]$ range for both pitch and yaw. All the samples are aligned on a global center point c_{l}.

Correlation Analysis. We assign the landmark indices as shown in Fig. 1a. Then we compute the correlation coefficient between each landmark and gaze coordinates. More precisely, two correlation coefficients are computed: the gaze yaw - horizontal landmark position and the gaze pitch - vertical landmark position. They are displayed in Fig. 1b.

The following comments can be made. First, the position of the iris center (landmark 17) is strongly correlated with gaze, as expected. The correlation coefficient between the horizontal (respectively vertical) position of the iris center and the gaze yaw (respectively pitch) is close to 1. Furthermore, the joint data distribution between the iris center and gaze displayed in Fig. 1c indicates that they seem to follow a linear relationship, especially in the pitch direction. Second, the gaze pitch is also highly correlated with other landmarks (red curve in Fig. 1b). This reflects that looking up or looking down requires some eyelid movement which are thus quite indicative of the gaze pitch. Third, the gaze yaw is only weakly correlated with eyelid landmarks, which means that looking to the left or right is mainly conducted by iris movements.

In summary, we find that the eye landmarks are correlated with the gaze and therefore they can provide strong support cues for estimating gaze.

4 Method

The proposed framework is shown in Fig. 2. It consists of two parts. The first part is a neural network which takes an eye image as input and regresses two sets of parameters: the coefficients of our joint CLGM landmarks and gaze model, and the scale and translation defining the eye region. The second part is a deterministic decoder. Based on the Constrained Landmark-Gaze model, it reconstructs the eye landmark positions and gaze with the two sets of parameters and the head pose. Note that in the reconstruction, the eye landmarks are computed using all parameters while the gaze is only determined using the CLGM coefficients and the head pose. An end-to-end training of the network is performed by combining the losses on landmark localization and gaze estimation. In our approach, we assume that the head pose has been obtained in advance. Below, we provide more details about the different parts of the model.

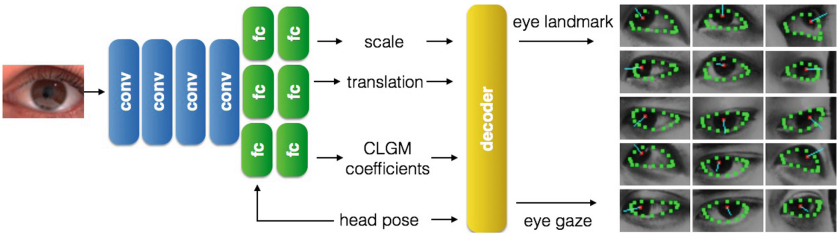

Fig. 2. Framework of the proposed method.

4.1 Constrained Landmark-Gaze Model

As with the 3D Morphable Model [44] or the Constrained Local Model [45] for faces, the eye shape can also be modeled statistically. Concretely, an eye shape can be decomposed as a weighted linear combination of a mean shape and a series of deformation bases according to:

$$\mathbf{v}_l(\alpha) = \mu^l + \sum_j \alpha_j \lambda_j^l \mathbf{b}_j^l, \tag{1}$$

where μ^l is the mean eye shape and λ_j^l represents the eigenvalue of the j^{th} linear deformation basis $\mathbf{b}_j^l$. The coefficients α denote the variation parameters determining eye shape while the superscript l means landmark.

As demonstrated in the previous section, the eye landmark positions are correlated with the gaze directions. In addition, we can safely assume that the landmark positions are also correlated. Therefore, we propose the Constrained Landmark-Gaze Model to explicitly model the joint variation of eye landmarks and gaze.

Concretely, we first extract the set of landmarks $\mathcal{I}$ from the N_s UnityEyes samples and align them with the global eye center. Denoting by $\mathbf{l}_{k,i} = (\mathbf{l}_{k,i}^x, \mathbf{l}_{k,i}^y)$, the horizontal and vertical positions of the i^{th} landmark of the k^{th} UnityEyes sample, and by $(\mathbf{g}_k^\phi, \mathbf{g}_k^\theta)$ the gaze pitch and yaw of the same sample, we can define the 1-D landmark-gaze array:

$$[\mathbf{l}_{k,1}^y, \cdots, \mathbf{l}_{k,N_l}^y, \mathbf{l}_{k,1}^x, \cdots, \mathbf{l}_{k,N_l}^x, \mathbf{g}_k^\phi, \mathbf{g}_k^\theta] \tag{2}$$

where N_l denotes the number of landmarks ($N_l = 17$), and the superscripts y, x, ϕ, θ represent the vertical position, horizontal position, pitch angle and yaw angle, respectively. This landmark-gaze vector has $2N_l + 2$ elements.

We then stack the vector of each sample into a matrix $\mathbf{M}$ of dimension $N_s \times (2N_l + 2)$, from which the linear bases $\mathbf{b}_j^{lg}$ representing the joint variation of eye landmark locations and gaze directions are derived through Principal Component Analysis (PCA). Thus, the eye shape and gaze of any eye sample can be modeled as:

$$\mathbf{v}_{lg}(\alpha) = \mu^{lg} + \sum_{j=1}^{2N_l+2} \lambda_j^{lg} \alpha_j \mathbf{b}_j^{lg} \tag{3}$$

where the superscript lg denotes the joint modeling of landmark and gaze. The definition of other symbols are similar to those in Eq. 1. Note that the resulting vector $\mathbf{v}_{lg}(\alpha)$ contains both the eye shape and gaze information.

In Eq. 3, the only variable is the vector of coefficients α. With a suitable learning algorithm, α can be determined to generate an accurate eye shape and gaze.

4.2 Joint Gaze and Landmark Inference Network

We use a deep convolutional neural network to jointly infer the gaze and landmark locations, as illustrated in Fig. 2. It comprises two parts: an encoder network inferring the coefficient α of the model in Eq. 3, as well as other geometric

parameters, and a decoder computing the actual landmark positions in the image and the gaze directions. The specific architecture for the encoder is described in Sect. 4.4. Below, we detail the decoder component and the loss used to train the network.

Decoder. We recall that the vector $\mathbf{v}_{lg}(\alpha)$ from the CLGM model only provides the aligned landmark positions. Thus, to model the real landmark positions in the cropped eye images, the head pose, the scale and the translation of the eye should be taken into account. In our framework, the scale s and translation $\mathbf{t}$ are inferred explicitly by the network, while the head pose is assumed to have already been estimated (see Fig. 2).

Given the head pose $\mathbf{h}$ and the inferred parameters α, s and $\mathbf{t}$ from the network, a decoder is designed to compute the eye landmark locations and gaze direction. Concretely, the decoder first uses α to computes the aligned eye shape and gaze according to Eq. 3. Then the aligned eye shape is further transformed with the head pose rotation matrix $\mathbf{R}(\mathbf{h})$, the scale s and the translation $\mathbf{t}$ to reconstruct the eye landmark positions in the input image:

$$\begin{bmatrix} \mathbf{l}_p^x \\ \mathbf{l}_p^y \end{bmatrix} = s \cdot \mathbf{Pr} \cdot \mathbf{R}(\mathbf{h}) \cdot \begin{bmatrix} \mathbf{v}_{lg}^x(\alpha) \\ \mathbf{v}_{lg}^y(\alpha) \\ 0 \end{bmatrix} + \mathbf{t} \tag{4}$$

$$\begin{bmatrix} cos(\mathbf{g}_p^\phi)sin(\mathbf{g}_p^\theta) \\ -sin(\mathbf{g}_p^\phi) \\ cos(\mathbf{g}_p^\phi)cos(\mathbf{g}_p^\theta) \end{bmatrix} = \mathbf{R}(\mathbf{h}) \cdot \begin{bmatrix} cos(\mathbf{v}_{lg}^\phi(\alpha))sin(\mathbf{v}_{lg}^\theta(\alpha)) \\ -sin(\mathbf{v}_{lg}^\phi(\alpha)) \\ cos(\mathbf{v}_{lg}^\phi(\alpha))cos(\mathbf{v}_{lg}^\theta(\alpha)) \end{bmatrix} \tag{5}$$

where $\mathbf{l}_p$ and $\mathbf{g}_p$ denote the predicted eye landmark positions and gaze respectively, and $\mathbf{Pr}$ is the projection matrix from 3D to 2D. From the equations above, note that the eye landmark positions are determined by all parameters while the gaze angles are only determined by the coefficient α and the head pose. Thus gaze estimation is geometrically coupled with the head pose as it should be, but is decoupled from the eye scale and translation.

Training Loss. To train the network, we define loss on both the predicted eye landmark positions and on the gaze according to

$$L(I) = w_l||\mathbf{l}_p - \mathbf{l}_g||_2 + w_g||\mathbf{g}_p - \mathbf{g}_g||_1 \tag{6}$$

where $\mathbf{l}_g$ and $\mathbf{g}_g$ represent the ground truth in the image I for the landmark positions and gaze respectively, and w_l and w_g denote the weights for the landmark loss and gaze loss respectively. Note from Eq. 6 that we do not provide any groundtruth for scale or translation during training (or to α), since the network automatically learns how to predict them from the landmark loss.

4.3 CLGM Revisited

As mentioned in Sect. 3, the UnityEyes models only reflects the main correlation between the gaze and eye landmarks. To account for real people and real images

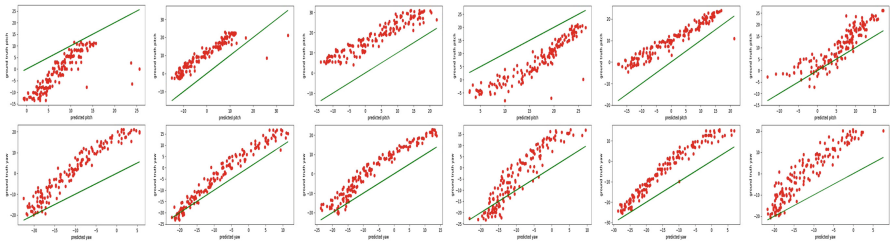

Fig. 3. Gaze bias between prediction and ground truth. 1^{st} row: pitch angle. 2^{nd} row: yaw angle. Green line: identity mapping (Color figure online)

and obtain a more accurate CLGM model, we perform an evaluation guided correction of the CLGM model. The main idea is to evaluate how our gaze prediction approach trained only with synthetic data (for both the CLGM model and the network model) performs on target real data. Then, by comparing the gaze predictions with the actual gaze data for a subject, we can estimate a gaze correction model mapping the prediction to the real ones. Such a parametric model can then be exploited on the UnityEyes data to correct the gaze values associated with a given eye landmarks configuration. A corrected CLGM model can then be obtained from the new data, and will implicitly model the joint variations of eye landmarks on the real data with the actual gaze on real data.

More concretely, we proceed as follows. We first train a gaze estimator (and landmark detector at the same time) with the proposed framework using only the UnityEyes synthetic data. Then the synthetic trained estimator is applied on a target database (UTMultiview, Eyediap) comprising N_{sub} subjects. For each subject, we can obtain gaze prediction/ground truth pairs, as illustrated in Fig. 3. According to these plots, we found a linear model (different for each subject) can be fitted between the prediction and the ground truth. In other words, the gaze predicted by the synthetic model is biased with respect to the real one but can be corrected by applying a linear model. Thus, to obtain a CLGM model linked to real people, for each subject j we fit two linear models f_j^{ϕ} and f_j^{θ} for the pitch and yaw prediction. Then, using the UnityEyes images, we construct a matrix $\mathbf{M}_j$ similar to the $\mathbf{M}$ matrix in Sect. 4.1, but stacking now the following landmark-gaze vectors instead of those in Eq. 2:

$$[\mathbf{l}_{k,1}^{y}, \cdots, \mathbf{l}_{k,N_l}^{y}, \mathbf{l}_{k,1}^{x}, \cdots, \mathbf{l}_{k,N_l}^{x}, f_j^{\phi}(\mathbf{g}_k^{\phi}), f_j^{\theta}(\mathbf{g}_k^{\theta})] \tag{7}$$

Then, a matrix $\mathbf{M}$ is build by stacking all $\mathbf{M}_j$ matrices, from which the corrected CLGM model taking into account real data is derived[1].

[1] Note that the corrected model relies on real data. In all experiments, the subject(s) used in the test set are never used for computing a corrected CLGM model.

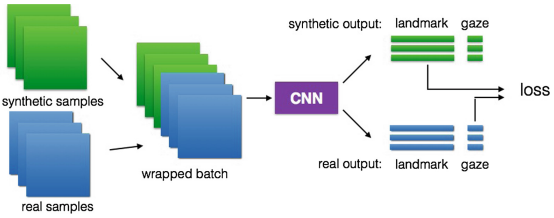

Fig. 4. Training combining synthetic and real data.

4.4 Implementation Detail

Auxiliary Database. To the best of our knowledge, the only public database annotating both eye landmark positions and gaze is MPIIGaze [13]. However, it only labels three eye landmarks per image on a subset of the dataset, which is not enough for our framework. Instead, we use the synthetic samples from UnityEyes as an auxiliary database. Concretely, we sample m real eye images from the main database and another m synthetic eye images from the auxiliary database in every training batch. After the feedforward pass, the landmark loss in Eq. 6 is only computed on synthetic samples (which have landmark annotations), whereas the gaze loss is only computed on real eye samples, as illustrated in Fig. 4. Note that we do not consider the gaze loss on synthetic samples (although they do have gaze groundtruth) to avoid a further potential bias towards the synthetic data.

Eye Image Cropping. The original UnityEyes samples cover a wide region around the eyes and we need a tighter cropping. To improve the generalization of the network, we give random cropping centers and sizes while cropping UnityEyes samples. Cropped images are then resized to fixed dimensions.

Network Configuration. We set the size of the input images as 36×60. The network is composed of 4 convolutional layers and 6 fully connected layers. The 4 convolutional layers are shared among the predictions of the CLGM coefficients, scale and translation. After the 4 convolutional layers, the network is split into 3 task-specific branches and each branch consists of 2 fully connected layers. Note that the head pose information is also concatenated with the feature maps before the first fully connected layer in the CLGM coefficient branch since the eye shape is also affected by the head pose. The network is learned from scratch in this paper.

5 Experiment Protocol

5.1 Dataset

Two public datasets of real images are used: UTMultiview [29] and Eyediap [46].

UTMultiview Dataset. It contains a large amount of eye appearances under different view points for 50 subjects thanks to a 3D reconstruction approach. This dataset provides the ground truth of gaze and head pose, both with large variability. In our experiment, we follow the same protocol as [29] which relies on a 3-fold cross validation.

Eyediap Dataset. It was collected in office conditions. It contains 94 videos from 16 participants. The recording sessions include continuous screen gaze target (CS, small gaze range) and 3D floating gaze target (FT, large gaze range), both based either on a close to static head pose (SP) and mobile head pose (MP) scenario. In experiment, we follow the same person-independent (PI) protocol as [31]. Concretely, for the CS case, we first train a deep network with all the SP-CS subjects but *leave one person out*. Then the network is tested on the left one in both SP-CS and MP-CS sessions (for cross session validation). We do the same for FT case (SP-FT and MP-FT sessions). Note that all eye images are rectified so that their associated head poses are frontal [31].

5.2 Synthetic Dataset and CLGM Training

As mentioned above, we use UnityEyes as the auxiliary dataset.

CLGM. For each experimental setting (datasets or sessions), we derive a CLGM model trained from frontal head pose samples using the gaze ranges of this setting. The resulting CLGM models is then further corrected as described in Sect. 4.3.

Auxiliary Training Samples. For multitask training, the auxiliary synthetic samples are generated with corresponding gaze and head pose ranges matching those of the dataset and session settings.

Synthetic Sample Refinement. One challenge when training using multiple datasets is the different data distribution. Although SimGAN [47] has been proposed to narrow down the distribution gap between synthetic images and real images, optimizing GAN models is difficult. Without suitable hyper parameters and tricks, the semantic of images after refining can be distorted. In our experiment, we simply adapt the UnityEyes synthetic images to UTMultiview samples by grayscale histogram equalization, and to Eyediap samples by Gaussian blurring.

5.3 Model Setup

In terms of gaze estimation models, we considered the models below. The architectures are given in Fig. 2 (proposed approach) and in Fig. 5 (contrastive approaches). Note that the architectures of the first three models below are the same whenever possible and all the models below are pretrained with synthetic data so that a fair comparison can be made.

CrtCLGM + MTL. This is our proposed multitask framework based on the corrected CLGM model, see Fig. 2.

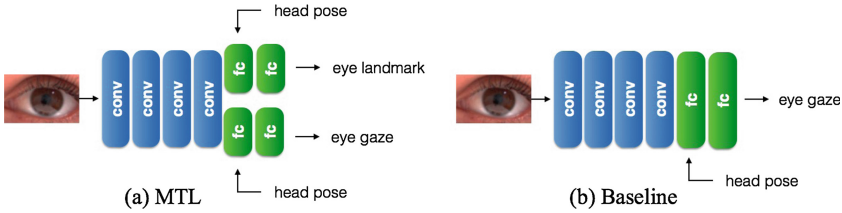

Fig. 5. Contrast models. (a) MTL architecture. (b) baseline architecture.

CLGM + MTL. This model is the same as above (CrtCLGM + MTL), except that the CLGM model is not corrected.

MTL. To contrast the proposed model, we implement a multitask learning network which also predicts landmarks and gaze jointly. Unlike the CrtCLGM + MTL, this model predicts the two labels directly by splitting the network into 2 separate branches after several shared layers. We also forward the head pose information to the features in both branches since both landmarks and gaze are affected by head pose.

Baseline. The baseline model performs a direct gaze regression from the eye appearances using the same base network architecture as in the two previous cases. The head pose is also used in this architecture.

MPIIGaze. For experiments on the Eyediap dataset, we also implemented the network of [13] to allow the comparison of different architectures. For the UTMultiview dataset, we directly report the result of this architecture from [13].

5.4 Performance Measurement

Gaze Estimation. We used the same accuracy measurement as [31] for gaze estimation. The gaze estimation error is defined as the angle between the predicted gaze vector and the groundtruth gaze vector.

Landmark Detection. We also measure the auxiliary task of our model. The GI4E database [48] is used to test the performance of iris center localization. To apply our method, we extracted eye images from the faces using dlib [49] and processed them with grayscale histogram equalization. The eye images are then forwarded to the UTMultiview trained network (frontal head pose assumed). In the evaluation, we adopt the maximum normalized error [23].

6 Results

We report the gaze estimation accuracy of UTMultiview and Eyediap in Fig. 6a and b respectively. Some qualitative results are demonstrated in Fig. 7. Please note that this paper targets at single eye gaze estimation. We think it is not suitable to compare with full face based methods since some datasets (UTMultiview)

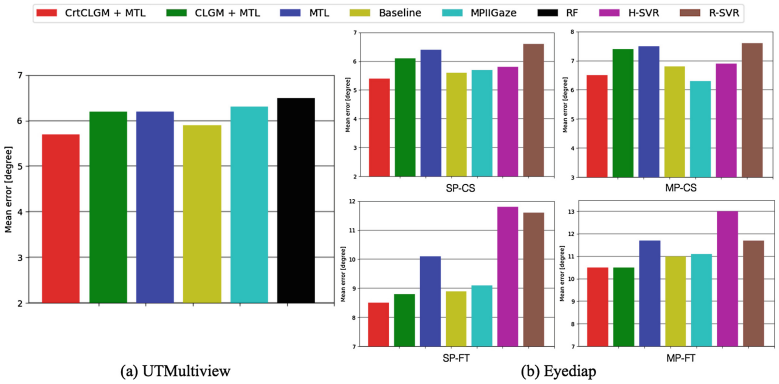

Fig. 6. Gaze estimation accuracy on UTMultiview and Eyediap.

do not provide the full face and the gaze definition can be different (e.g. gaze fixation points [9] and middle point of face as the origin of gaze direction [14]).

6.1 UTMultiview Dataset

From Fig. 6a, we note that the proposed CrtCLGM + MTL model shows the best performance (5.7°) among the contrast methods including two state-of-the-art works, MPIIGaze net [13] (6.3° with our implementation) and RF [29] (6.5°). We also note from Fig. 7 that accurate gaze estimation and landmark localization are achieved by our method regardless of eye scale, eye translation and large head pose.

In contrast, we find that the CLGM + MTL model performs worse than the CrtCLGM + MTL model. This is understandable since the optimization of the multitask loss can be difficult if the landmark-gaze correlations are different between the synthetic data and real data. Sometimes the optimization process competes between the two tasks and the final network can be bad for both tasks. This is also shown in Table 1 where the iris center localization of the CLGM + MTL model is not so accurate. This result demonstrates the importance of the CLGM correction.

When looking at the result of MTL model, it is a bit surprising that its error is on a par with the Baseline method and MPIIGaze net which only target gaze optimization. It thus seems that the MTL model failed to improve gaze estimation through a direct and pure feature sharing strategy. As shown in Fig. 5a, the landmark positions are regressed from the shared features directly in the landmark branch, which means some information such as eye scale and eye translation are contained in the shared features. Although this geometric information is important to landmark localization, they are irrelevant elements for gaze estimation and might even degrade it. Therefore, our mechanism which decouples the eye scale and translation from eye shape variation is thus necessary and important.

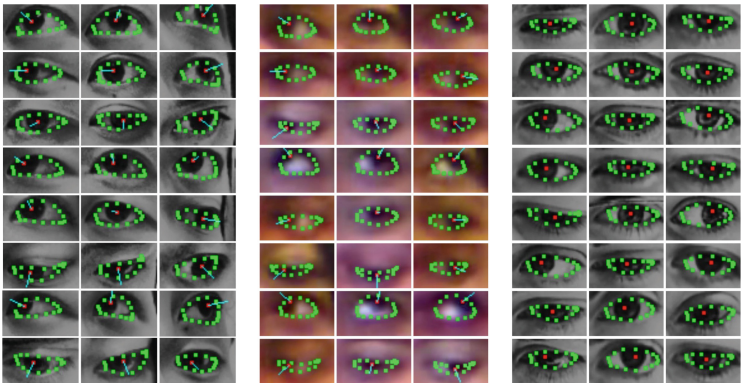

Fig. 7. Eye landmark detection and gaze estimation results on UTMultiview (***Left***), Eyediap (***Middle***) and GI4E (***Right***) dataset. The cyan line represents the estimated gaze direction.

Owing to the reasonable geometric modeling of the scale, translation and head pose, our method also demonstrates superior performance to the Baseline model and the MPIIGaze network. Note that our Baseline model is slightly better than the MPIIGaze net, possibly because in the Baseline, the head pose information is added earlier and thus processed by more layers. Thanks to the large data amount (including the synthetic data used for pretraining), all the network models perform better than the RF (random forest) method.

6.2 Eyediap Dataset

Two existing methods H-SVR and R-SVR [31] are used for comparison. From Fig. 6b, we note that the proposed CrtCLGM + MTL model achieves the best result in all sessions (5.4°, 6.5°, 8.5°, 10.5° respectively) except the SP-CS session (MPIIGaze: 6.3°). In particular, compared with other methods, the performance improvement is much larger in the floating target (FT) session than in the continuous screen (CS) session, indicating that our method can perform even better for applications requiring gaze in the 3D space, when large head pose and gaze angles are present.

When comparing the results of the CrtCLGM + MTL model and CLGM + MTL model, we note that the former is better for all the sessions which further corroborate the importance of CLGM correction. Compared with the UTMultiview dataset, the MTL model obtains much worse results than other network based methods (especially the Baseline and the MPIIGaze) in Eyediap dataset. Given that the Eyediap samples are much more blurry than the UTMultiview dataset, the direct regression of landmark positions without the constrained model is difficult and inaccurate, and the inaccurate landmark detection may confuse the shared architecture and ultimately instead of helping gaze inference, tends to degrade the results. In contrast, our CLGM model is better at handling

blurry data thanks to the introduction of an explicit geometrical model, and that learning the parameters of the CLGM model rather than the unconstrained landmark positions provides some form of regularization which prevents the network from overfitting. This also demonstrates the advantage of our method over traditional geometrical approaches where high resolution images are usually required.

When comparing the results across sessions (i.e. recording situations, see Sect. 5.1), we can observe that the accuracy of the floating target (FT) sessions are worse than the CS screen sessions, which is intrinsically due to the more difficult task (looking at a 3D space target) involving much larger head poses (potentially harming eye image frontalization) and gaze ranges. On the other hand, the results show that our method achieves the most robust performance in cross session validation (train on SP, test on MP).

6.3 Iris Center Localization

Lastly, we show the performance of the auxiliary task, landmark detection. Table 1 reports the accuracy of the iris center localization.

Table 1. Iris center localization on GI4E dataset.

Method	$d_{eye} \leq 0.05(\%)$	$d_{eye} \leq 0.1(\%)$	$d_{eye} \leq 0.25(\%)$
Timm et al. [24]	92.4	96.0	97.5
Villanueva et al. [25]	93.9	97.3	98.5
Gou et al. [23]	94.2	99.1	99.8
CLGM + MTL	92.5	99.7	100
CrtCLGM + MTL	**95.1**	**99.7**	**100**

From the table, our method achieves the best performance compared with the state-of-the-art works in all the three criteria which correspond to the range of pupil diameter, the range of iris diameter and the distance between eye center and eye corner [24] respectively. Concretely, most of the detections are within the pupil, few of them lie outside the iris and almost all falls inside the eye region. In contrast, the CLGM + MTL model is inferior to the CrtCLGM + MTL one in the $d_{eye} \leq 0.05(\%)$ measurement, which means more detections of the CLGM + MTL model deviate from the pupil. As discussed in Sect. 6.1, it can be explained by the differences in landmark-gaze correlations between the synthetic data and real data.

Some qualitative results are shown in Fig. 7. Note that we assumed that the head poses of all the samples were frontal since this label was not provided in this dataset. Even under this assumption we still achieved accurate iris center localization, which demonstrates that our method can be used in a wide scope of eye landmark detection applications where head pose information may not be available.

7 Conclusion

In this paper, we proposed a multitask learning approach for gaze estimation. This approach is based on a Constrained Landmark-Gaze Model which models the joint variation of the eye landmarks and gaze in an explicit way, which helps in (i) solving the absence of annotation on different datasets for some task (in our case, landmarks); (ii) better leveraging in this way the benefits of the multitask approach. This model differs from geometrical methods since landmarks and gaze are jointly extracted from eye appearance. Experiments demonstrate the capacity of our approach, which is shown to outperform the state-of-the-art in challenging situations where large head poses and low resolution eye appearances are presented. Our idea of CLGM model can also be extended to joint tasks like facial landmark detection and head pose estimation. For instance, using the FaceWarehouse [50] dataset as 3D face and landmark statistical model to generate faces with different identities and expressions which can be randomly rotated with different head poses. Since pose and landmarks are correlated, a constrained landmark-head pose model could be built and trained as we propose.

On the hand, although the head pose is not so important for landmark detection as shown in Table 1, we note from Eq. 5 that our model requires precise head pose label for gaze estimation, which may limit the application scope of our method. This problem can be possibly addressed by estimating the head pose from the eye appearance or full face as another task. We leave this as a future work.

Acknowledgement. This work was partly funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF), and by the European Unions Horizon 2020 research and innovation programme under grant agreement no. 688147 (MuMMER, mummer-project.eu).

References

1. Bixler, R., Blanchard, N., Garrison, L., D'Mello, S.: Automatic detection of mind wandering during reading using gaze and physiology. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI 2015, pp. 299–306. ACM, New York (2015)
2. Hiraoka, R., Tanaka, H., Sakti, S., Neubig, G., Nakamura, S.: Personalized unknown word detection in non-native language reading using eye gaze. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, pp. 66–70. ACM, New York (2016)
3. Velichkovsky, B.M., Dornhoefer, S.M., Pannasch, S., Unema, P.J.: Visual fixations and level of attentional processing. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, ETRA 2000, pp. 79–85. ACM, New York (2000)
4. Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychol. **26**(Suppl. C), 22–63 (1967)
5. Vidal, M., Turner, J., Bulling, A., Gellersen, H.: Wearable eye tracking for mental health monitoring. Comput. Commun. **35**(11), 1306–1311 (2012)

6. Ishii, R., Otsuka, K., Kumano, S., Yamato, J.: Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. ACM Trans. Interact. Intell. Syst. **6**(1), 4:1–4:31 (2016)
7. Andrist, S., Tan, X.Z., Gleicher, M., Mutlu, B.: Conversational gaze aversion for humanlike robots. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction, HRI 2014, pp. 25–32. ACM, New York (2014)
8. Moon, A., et al.: Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction, HRI 2014, pp. 334–341. ACM, New York (2014)
9. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H.: Eye tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2176–2184 (2016)
10. Tonsen, M., Steil, J., Sugano, Y., Bulling, A.: InvisibleEye: mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), vol. 1, no. 3 (2017)
11. Huang, Q., Veeraraghavan, A., Sabharwal, A.: Tabletgaze: unconstrained appearance-based gaze estimation in mobile tablets. arXiv preprint arXiv:1508.01244 (2015)
12. Zhu, W., Deng, H.: Monocular free-head 3D gaze tracking with deep learning and geometry constraints. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
13. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild, pp. 4511–4520 (2015)
14. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: full-face appearance-based gaze estimation (2016)
15. Wood, E., Baltrušaitis, T., Morency, L.P., Robinson, P., Bulling, A.: Learning an appearance-based gaze estimator from one million synthesised images. In: Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications, pp. 131–138 (2016)
16. Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
17. Venkateswarlu, R., et al.: Eye gaze estimation from a single image of one eye, pp. 136–143 (2003)
18. Funes Mora, K.A., Odobez, J.M.: Geometric generative gaze estimation (G3E) for remote RGB-D cameras, pp. 1773–1780, June 2014
19. Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P., Bulling, A.: A 3D morphable eye region model for gaze estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 297–313. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_18
20. Ishikawa, T.: Passive driver gaze tracking with active appearance models (2004)
21. Wood, E., Bulling, A.: Eyetab: model-based gaze estimation on unmodified tablet computers, pp. 207–210 (2014)
22. Gou, C., Wu, Y., Wang, K., Wang, F.Y., Ji, Q.: Learning-by-synthesis for accurate eye detection. In: ICPR (2016)
23. Gou, C., Wu, Y., Wang, K., Wang, K., Wang, F., Ji, Q.: A joint cascaded framework for simultaneous eye detection and eye state estimation. Pattern Recogn. **67**, 23–31 (2017)
24. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. In: VISAPP (2011)

25. Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., Cabeza, R.: Hybrid method based on topography for robust detection of iris center and eye corners. ACM Trans. Multimedia Comput. Commun. Appl. **9**(4) (2013)
26. Tan, K.H., Kriegman, D.J., Ahuja, N.: Appearance-based eye gaze estimation, pp. 191–195 (2002)
27. Noris, B., Keller, J.B., Billard, A.: A wearable gaze tracking system for children in unconstrained environments. Comput. Vis. Image Underst. **115**(4), 476–486 (2011)
28. Martinez, F., Carbone, A., Pissaloux, E.: Gaze estimation using local features and non-linear regression, pp. 1961–1964 (2012)
29. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3D gaze estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1821–1828 (2014)
30. Lu, F., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression, pp. 153–160 (2011)
31. Funes-Mora, K.A., Odobez, J.M.: Gaze estimation in the 3D space using RGB-D sensors. Int. J. Comput. Vis. **118**(2), 194–216 (2016)
32. Palmero, C., Selva, J., Bagheri, M.A., Escalera, S.: Recurrent CNN for 3d gaze estimation using appearance and shape cues, p. 251 (2018)
33. Park, S., Zhang, X., Bulling, A., Hilliges, O.: Learning to find eye region landmarks for remote gaze estimation in unconstrained settings, pp. 21:1–21:10 (2018)
34. Wang, K., Zhao, R., Ji, Q.: A hierarchical generative model for eye image synthesis and eye gaze estimation, June 2018
35. Park, S., Spurr, A., Hilliges, O.: Deep pictorial gaze estimation, September 2018
36. Ruder, S.: An overview of multi-task learning in deep neural networks, June 2017
37. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. CoRR abs/1603.01249 (2016)
38. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, 30 May–3 June 2017, pp. 17–24 (2017)
39. Wang, F., Han, H., Shan, S., Chen, X.: Deep multi-task learning for joint prediction of heterogeneous face attributes. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, 30 May–3 June 2017, pp. 173–179 (2017)
40. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
41. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: CVPR, pp. 676–684. IEEE Computer Society (2015)
42. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. CoRR abs/1604.03539 (2016)
43. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.S.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. CoRR abs/1611.05377 (2016)
44. IEEE: A 3D Face Model for Pose and Illumination Invariant Face Recognition. IEEE, Genova, Italy (2009)
45. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models, January 2006

46. Funes Mora, K.A., Monay, F., Odobez, J.M.: 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, March 2014
47. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. CoRR abs/1612.07828 (2016)
48. Villanueva, A., Ponz, V., Sesma-Sanchez, L., Ariz, M., Porta, S., Cabeza, R.: Hybrid method based on topography for robust detection of iris center and eye corners. ACM Trans. Multimedia Comput. Commun. Appl. **9**(4), 25:1–25:20 (2013)
49. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: CVPR, pp. 1867–1874. IEEE Computer Society (2014)
50. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: a 3D facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2014)