Person-independent 3D Gaze Estimation using Face Frontalization

László A. Jeni
Carnegie Mellon University
Pittsburgh, PA, USA
laszlojeni@cmu.edu

Jeffrey F. Cohn
University of Pittsburgh
Pittsburgh, PA, USA
jeffcohn@pitt.edu

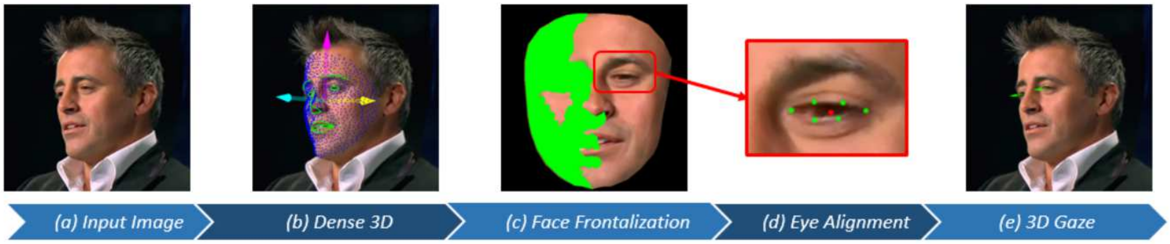

Figure 1: From a 2D image of a person's face (a) a dense, part-based 3D deformable model is aligned (b) to reconstruct a partial frontal view of the face (c). Binary features are extracted around eye and pupil markers (d) for the 3D gaze calculation (e).

Abstract

Person-independent and pose-invariant estimation of eye-gaze is important for situation analysis and for automated video annotation. We propose a fast cascade regression based method that first estimates the location of a dense set of markers and their visibility, then reconstructs face shape by fitting a part-based 3D model. Next, the reconstructed 3D shape is used to estimate a canonical view of the eyes for 3D gaze estimation. The model operates in a feature space that naturally encodes local ordinal properties of pixel intensities leading to photometric invariant estimation of gaze. To evaluate the algorithm in comparison with alternative approaches, three publicly-available databases were used, Boston University Head Tracking, Multi-View Gaze and CAVE Gaze datasets. Precision for head pose and gaze averaged 4 degrees or less for pitch, yaw, and roll. The algorithm outperformed alternative methods in both datasets.

1. Introduction

Gaze and eye contact communicate interpersonal engagement and emotion, express intimacy, reveal attention and cognitive processes, signal objects or events of interest, and regulate social interaction [14, 22, 30, 31]. Automated tracking of gaze in highly constrainted contexts, such as while seated in front of a computer monitor or when wearing specialized eyewear, is well developed using commercial software [43]. In less constrained contexts, such as during social interaction or in a car while driving, automated gaze tracking presents a challenging and vital problem. Efforts to detect gaze in social and automotive contexts are just beginning [25].

Gaze estimation methods can be categorized into model-based and appearance-based approaches [18]. Model-based 3D gaze estimation methods use 3D eyeball models and estimate gaze direction using geometric eye features [13, 17]. They typically use infrared light sources together with a high-resolution camera to locate the 3D eyeball position and its line of sight via personal calibration. Although this approach can accurately estimate gaze direction, it necessitates specialized hardware that limits its range of application. If the iris contour alone is used to detect line of sight, the need for specialized eye wear can be relaxed [12, 49, 20, 45]. The latter is effective for short distance scenarios in which high-resolution observations are available. Their effectiveness in mid-distance scenarios is unclear.

Appearance-based methods compute non-geometric image features from input eye images to estimate gaze direction. This approach frames the gaze estimation problem into one of learning a mapping function from the input eye images to the target gaze directions. The corresponding mapping can be learned using different regression techniques, including artificial neural networks [4], adaptive linear regression [26], interpolation [40], and Gaussian process regression [36, 46].

For appearance-based 3D gaze estimation, the 3D position of the eye must be found in order to estimate the gaze target in the camera-coordinate system. With the recent advancement of monocular 3D head pose tracking [29] and the increasing availability of depth cameras with head pose tracking capabilities [3], the means of capturing 3D head poses are becoming readily available. Indeed, recent appearance-based 3D gaze estimation methods use 3D head poses obtained as an additional input for gaze estimation [27, 16].

Appearance-based methods have the advantage of requiring only a single camera and natural illumination. They can use images with common or even low resolution. They typically regard a whole eye image as a high-dimensional input vector and learn the mapping between these vectors and the gaze positions.

Appearance-based methods, however, often lack robustness to head motion and illumination variation. Because appearance-based methods require precise alignment of the eye region, head pose variation is particularly challenging. Without precise alighnment, they are prone to large error. Varying illumination conditions, such as in a driving scenario [20], also affect their performance. Using active near-infrared imaging can alleviate this issue [45].

3D estimation from 2D video is a promising alternative. This is made possible in part by recent advances in 2D shape alignment that use discriminative shape regression methods [10, 38, 48, 32, 9]. These techniques predict a face shape in a cascade manner: They begin with an initial guess about shape and then progressively refine that guess by regressing a shape increment step-by-step from a feature space. The feature space can be hand designed, and may utilize SIFT features [48] or learned from the data [10, 6, 32].

Our approach exploits cascade shape regression for 3D gaze and head pose estimation. The method was made possible, in part, by training on the Multi-view Gaze (MVG) Dataset [37] that contains 8,000 3D face meshes. The method was validated in a series of tests. We found that eye alignment and 3D absolute gaze estimation from 2D images effectively handles previously unseen faces that may span a variety of poses and illuminations.

This paper advances two main novelties. First, the method estimates self-occluded markers and computes a canonical view using the estimated 3D head pose. This eliminates the need of learning a manifold of appearance features for gaze estimation. Second, the proposed method operates in a binary feature space that is robust to photometric variations and different scales. We demonstrate that precise 3D gaze estimation is possible from this space.

The paper is organized as follows: Section 2 details the cascade framework and different regression techniques. Section 3 describes the 3D gaze dataset and manual annotations we used for training our system. The efficiency of our novel solution method is illustrated by numerical experiments in Section 4. Conclusions are drawn in Section 5.

Notations. Vectors ($\mathbf{a}$) and matrices ($\mathbf{A}$) are denoted by bold letters. An $\mathbf{u} \in \mathbb{R}^d$ vector's Euclidean norm is $\|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^d u_i^2}$. $\mathbf{B} = [\mathbf{A}_1; \ldots; \mathbf{A}_K] \in \mathbb{R}^{(d_1+\ldots+d_K)\times N}$ denotes the concatenation of matrices $\mathbf{A}_k \in \mathbb{R}^{d_k \times N}$.

2. Methods

2.1. Part-based Linear Face Models

We are interested in building a dense linear shape model. A shape model is defined by a 3D mesh and, in particular, by the 3D vertex locations of the mesh, called landmark points. Consider the 3D shape as the coordinates of 3D vertices that make up the mesh:

$$\mathbf{x} = [x_1; y_1; z_1; \ldots; x_M; y_M; z_M], \tag{1}$$

or, $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_M]$, where $\mathbf{x}_i = [x_i; y_i; z_i]$. We have T samples: $\{\mathbf{x}(t)\}_{t=1}^T$.

We assume that – apart from scale, rotation, and translation – all samples $\{\mathbf{x}(t)\}_{t=1}^T$ can be approximated by means of a linear subspace.

The 3D point distribution model (PDM) describes non-rigid shape variations linearly and composes it with a global rigid transformation, placing the shape in the image frame:

$$\mathbf{x}_i = \mathbf{x}_i(\mathbf{p}, \mathbf{q}) = s\mathbf{R}(\bar{\mathbf{x}}_i + \mathbf{\Phi}_i \mathbf{q}) + \mathbf{t} \quad (i = 1, \ldots, M), \tag{2}$$

where $\mathbf{x}_i(\mathbf{p}, \mathbf{q})$ denotes the 3D location of the i^{th} landmark and $\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{t}\}$ denotes the rigid parameters of the model, which consist of a global scaling s, angles of rotation in three dimensions ($\mathbf{R} = \mathbf{R}_1(\alpha)\mathbf{R}_2(\beta)\mathbf{R}_3(\gamma)$), a translation $\mathbf{t}$. The non-rigid transformation is denoted with $\mathbf{q}$. Here $\bar{\mathbf{x}}_i$ denotes the mean location of the i^{th} landmark (i.e. $\bar{\mathbf{x}}_i = [\bar{x}_i; \bar{y}_i; \bar{z}_i]$ and $\bar{\mathbf{x}} = [\bar{\mathbf{x}}_1; \ldots; \bar{\mathbf{x}}_M]$). The d pieces of $3M$ dimensional basis vectors are denoted with $\mathbf{\Phi} = [\mathbf{\Phi}_1; \ldots; \mathbf{\Phi}_M] \in \mathbb{R}^{3M \times d}$. Vector $\mathbf{q}$ represents the 3D distortion of the face in the $3M \times d$ dimensional linear subspace.

In Eq. (2) one can assume that the prior of the parameters follow a normal distribution with mean $\mathbf{0}$ and variance $\mathbf{\Lambda}$ at a parameter vector $\mathbf{q}$: $p(\mathbf{q}) \propto N(\mathbf{0}, \mathbf{\Lambda})$ and can use Principal Component Analysis (PCA) to determine the d pieces of $3M$ dimensional basis vectors ($\mathbf{\Phi} = [\mathbf{\Phi}_1; \ldots; \mathbf{\Phi}_M] \in \mathbb{R}^{3M \times d}$). This approach has been used successfully in a broad range of face alignment techniques, such as Active Appearance Models [28] or 3D Morphable Models [5]. This procedure would result in a holistic shape model with a high compression rate, but on the on the other hand, its components have a global reach and they lack of semantic meaning.

The deformations on the face can be categorized into two separate subsets: rigid (the shape of the face) and non-rigid (facial expressions) parts. We reformulate Eq. (2) to model these deformations separately:

$$\mathbf{x}_i = \mathbf{x}_i(\mathbf{p}, \mathbf{r}, \mathbf{s}) = s\mathbf{R}(\bar{\mathbf{x}}_i + \mathbf{\Theta}_i \mathbf{r} + \mathbf{\Psi}_i \mathbf{s}) + \mathbf{t} \quad (i = 1, \ldots, M), \tag{3}$$

where the d pieces of $3M$ dimensional basis vectors ($\mathbf{\Theta} = [\mathbf{\Theta}_1;\ldots;\mathbf{\Theta}_M] \in \mathbb{R}^{3M\times d}$) describes the rigid, and the the e pieces of $3M$ dimensional basis vectors ($\mathbf{\Psi} = [\mathbf{\Psi}_1;\ldots;\mathbf{\Psi}_M] \in \mathbb{R}^{3M\times e}$) describes the non-rigid deformations.

For the deformable model building we build on the work of Jeni et al. [21]. We used the BP4D-Spontaneous [50] dataset that consists of high-resolution 3D face scans and comes with per frame Facial Action Unit Coding (FACS [15]).

To build the rigid part, we selected neutral frames from each subject based on the FACS annotation and applied PCA to determine the basis vectors ($\mathbf{\Theta}$) and their mean ($\bar{\mathbf{x}}$). This provides us a holistic linear subspace that describes the variation of the face shape only. Note that the neutral face is only required during the model building, it is not required for testing.

To build a linear subspace that describes the non-rigid deformations ($\mathbf{\Psi}$) we followed the method of Tena et al [41]. The goal is to build a model that composed of a collection of PCA part-models that are independently trained but share soft boundaries. This model generalizes to unseen data better than the traditional holistic approach. Before the model building, we subtracted the person's own neutral face to remove all the personal variation from the data. Note, that if we would have used the global mean face for the subtraction ($\bar{\mathbf{x}}$) that would leave some of the rigid variation in the dataset. In our experiment we obtained 13 compact clusters, similar to the ones reported in [41].

2.2. Dense Cascade Regression

In this section we describe the general framework of dense cascade regression for face alignment. We build on the work of Xiong and De la Torre [48]. Given an image $\mathbf{d} \in \mathbb{R}^{a\times 1}$ of a pixels, $\mathbf{d}(\mathbf{y}) \in \mathbb{R}^{b\times 1}$ indexes b markers in the image. Let $\mathbf{h}$ be a feature extraction function (e.g. HOG, SIFT or binary features) and $\mathbf{h}(\mathbf{d}(\mathbf{y})) \in \mathbb{R}^{Fb\times 1}$ in the case of extracting features of length F. During training we will assume that the ground truth locations of the b markers are known. We refer to them as $\mathbf{y}_\star$.

We used a face detector on the training images to provide an initial configuration of the markers ($\mathbf{y}_0$), which correspond to the frontal projection of the 3D reference face ($\bar{\mathbf{x}}$ in eq. (3).

In this framework, face alignment can be framed as minimizing the following function over ($\Delta\mathbf{y}$):

$$f(\mathbf{y}_0 + \Delta\mathbf{y}) = \|\mathbf{h}(\mathbf{d}(\mathbf{y}_0 + \Delta\mathbf{y})) - \beta_\star\|_2^2 \qquad (4)$$

where $\beta_\star = \mathbf{h}(\mathbf{d}(\mathbf{y}_\star))$ represents the feature values in the ground truth markers.

The feature extraction function ($\mathbf{h}$) can be highly nonlinear and minimizing eq. (4) would require numerical approximations, which are computational expensive. Instead we learn a series of linear regressor matrices ($\mathbf{R}_i$), that produce a sequence of updates starting from $\mathbf{y}_0$ and converging to $\mathbf{y}_\star$ in the training data:

$$\Delta\mathbf{y}_i = \mathbf{R}_{i-1}\beta_{i-1} + \mathbf{b}_{i-1} \qquad (5)$$

$$\mathbf{y}_i = \mathbf{y}_{i-1} + \Delta\mathbf{y}_i \rightarrow \mathbf{y}_\star \qquad (6)$$

In our case, the annotation $\mathbf{y}$ consists of the projected 2D locations of the 3D markers and their corresponding visibility information:

$$\mathbf{y} = [x_1; y_1; v_1; \ldots; x_M; y_M; v_M], \qquad (7)$$

where $v_i \in [0,1]$ indicates if the marker is visible ($v_i = 1$) or not ($v_i = 0$).

To ensure the consistency and semantic correspondence between the 3D and 2D markers we followed the protocol of Jeni et al. [21].

In all our experiments we used localized binary features [7] for training the regression cascades. In comparison with dense features (such as SIFT or HOG), the central computational advantage of binary descriptors stems from their ability to encode the comparison of two intensity values which could be bytes (8 bits), floats (32 bits) or doubles (64 bits) in an extremely compact form as a single bit. This quantization has an additional benefit as it naturally encodes the local ordinal [51] properties of pixel intensities leading to obvious photometric invariant properties.

2.3. 3D Model Fitting

The dense cascade regressor defined in the previous section provides projected 2D locations of the 3D markers. To reconstruct the 3D shape from the 2D shape ($\mathbf{z}$) we need to minimize the reconstruction error using eq. (3):

$$\arg\min_{\mathbf{p},\mathbf{r},\mathbf{s}} \sum_{i=1}^{M} \|\mathbf{P}\mathbf{x}_i(\mathbf{p},\mathbf{r},\mathbf{s}) - \mathbf{z}_i\|_2^2 \qquad (8)$$

Here $\mathbf{P}$ denotes the projection matrix to 2D, and $\mathbf{z}$ is the target 2D shape. An iterative method can be used to register 3D model on the 2D landmarks. The algorithm iteratively refines the 3D shape and 3D pose until convergence, and estimates the rigid ($\mathbf{p} = \{s, \alpha, \beta, \gamma, \mathbf{t}\}$) and non-rigid transformations ($\mathbf{r}$ and $\mathbf{s}$).

This equation assumes that there is a semantic correspondence between the 2D and 3D markers. The lack of correspondence requires a correction step [8], usually in a form of a selection matrix [45], that selects the right 3D markers corresponding to the 2D ones.

In our case the semantic correspondence has been established during model building time: the markers provided by the cascade regressor are 2D projections of the 3D markers. Furthermore, the cascade regressor estimates the visibility of markers. We can incorporate this information in eq. (8), by constraining the process to the visible markers:

$$\arg\min_{\mathbf{p},\mathbf{r},\mathbf{s}} \sum_{i\in\boldsymbol{\xi}} \|\mathbf{P}\mathbf{x}_i(\mathbf{p},\mathbf{r},\mathbf{s}) - \mathbf{z}_i\|_2^2 \qquad (9)$$

where $\boldsymbol{\xi} = \{j|v_j = 1\}$ denotes the subset of marker-indices that are visible (see eq. (7)).

2.4. Face Frontalization

For appearance-based 3D gaze estimation, the 3D position of the eye must be found in order to estimate the gaze target in the camera-coordinate system. Allowing head motion for these methods is more difficult, since they require precise alignment of the eye region.

To overcome these issues, we use the reconstructed, dense 3D shape and the estimated head pose to synthesize a canonical, frontal looking view of the face (see Fig 1.c). This step allows us to estimate the gaze direction relative to the head pose and at the same time, it removes most of the rigid deformations in the facial appearance caused by the non-frontal head orientation.

The method can be described as follows: First, from the reconstructed 3D shape and the head pose, we render the corresponding depth maps, which represent the surfaces from the camera view point. Using the depth maps, we calculate a vertex level occlusion map of the mesh. In the next step, we calculate the canonical view of the face by removing all the rigid movements: in the parameter vector **p**, we set the scale to 1, remove the rotations ($\alpha = \beta = \gamma = 0$) and the translations. Finally, we map the original image to the canonical view using perspective texture mapping, and remove the occluded parts using the occlusion map. The result is a frontal looking face, with minimal to moderate level missing regions, based on the degree of head pose (see Fig 1.c).

One could utilize face completion techniques, such as rank minimization [33] or enforce soft-symmetries [19], to hallucinate the missing regions. This can affect the gaze estimation precision. We estimate gaze from the visible eye or eyes only. We detail this procedure in the next subsection.

2.5. Gaze Estimation from Feature Space

Appearance based gaze estimation methods typically regard a whole eye image as a high-dimensional input vector and learn the mapping between these vectors and the gaze positions. The raw pixel intensities are sensitive to illumination changes and usually require manifold learning techniques [26].

We propose a different approach, that operates in a binary feature space that is robust to photometric variations. From the visible part of frontalized image and the dense 3D shape, we calculate the positions of 6 eye contour markers and 1 pupil marker for each visible eye. We extract binary features around these points and train a linear Support Vector Regressor (SVR) [11] from these features to the 3D gaze direction. The gaze direction is given in the head coordinate system in the form of spherical coordinates (θ is the azimuthal angle, ϕ is the polar angle and the radius were normalized to unit length).

3. Dataset

We used the Boston University Head Tracking dataset [24] for the evaluation of the head-pose estimation, and the Multi-view Gaze (MVG) [37] and the CAVE Gaze datasets [35] for 3D gaze direction estimation.

3.1. Boston University Head Tracking Dataset

We used the Boston University Head Tracking dataset [24] to evaluate the performance of the proposed method for 3D head-pose estimation. The database contains short video sequences of different subjects with uniform (45 videos from five subjects) and varying (27 videos from three subjects) illuminations at a resolution of 320×240 pixels. Subjects were asked to perform various head movements, including translation and rotation, without distinctive facial expressions.

The dataset contains ground truth information of the head position and orientation, collected by the Flock of Birds magnetic tracker attached on the subject's head.

3.2. Multi-view Gaze Dataset

The Multi-view Gaze (MVG) Dataset [37] consists of a total of 50 (15 female and 35 male) people ranging in age approximately from 20 to 40 years old. A chin rest was used to stabilize the head position located at 60 cm apart from the monitor. During recording sessions, participants were instructed to look at a visual target displayed on the monitor. The screen was divided into a 16×10 regular grid, and the visual target moved to the center of each grid in a random order. The white circle shrank after the target stops at each position, and cameras were triggered at the time the circle disappeared. As a result, G = 160 (gaze directions) $\times 8$ (cameras) images were acquired from each participant at SXGA resolution, together with the 3D positions of the visual targets. The gaze directions spanned approximately ± 25 degrees horizontally and $\pm$ 15 degrees vertically, and this covered the range of natural gaze directions. Images are recorded by a fully calibrated 8 multi-camera system, and the 3D reconstruction of eye regions was done by using a patch-based multi-view stereo algorithm.

3.3. CAVE Gaze dataset

The CAVE Gaze dataset [35] consists of 5,880 high resolution images from 56 subjects (32 male, 24 female). Subjects were ethnically and racially diverse (European-American, African-American, South-Asian, Asian, and

Hispanic Latino) and 21 of them wore glasses. Head pose was stabilized using a chin rest. Subjects were imaged individually using a Canon EOS Rebel T3i camera and a Canon EF-S 18135 mm IS f/3.55.6 zoom lens, from five different horizontal positions while looking into seven horizontal and three vertical gaze directions.

4. Experiments

We executed a number of evaluations to judge the strength of the proposed method for 3D head pose estimation and 3D gaze estimation. Studies concern (i) 3D head pose estimation, (ii) the performance of 3D gaze estimation under various conditions, including head orientation, different image resolutions and illumination changes.

4.1. Head-pose estimation using Dense Models

In this experiment we evaluate the performance of the proposed method for head tracking using real faces from the Boston University (BU) head tracking database [24]. First, we used the uniform illumination subset (45 sequences) of the BU database and compared the estimated head pose to the ground truth provided in the dataset.

The mean absolute angular error of the head pose estimation is shown in Table 1 in comparison with results from different sources. The accuracies of Cascia et al. [24] and Xiao et al. [47] are taken from [29].

In the second part of the experiment we used the varying illumination subset (27 sequences) to evaluate the effect of the changing lighting conditions on the pose estimation. The mean absolute angular error of the Pitch, Yaw and Roll angle estimation is 2.72°, 4.87° and 2.24° respectively.

Method	Pitch	Yaw	Roll	Mean
La Cascia et al. (CT) [24]	6.1	**3.3**	9.8	6.4
Xiao et al. (CT) [47]	3.2	3.8	**1.4**	**2.8**
Asteriadis et al. (DVF) [2]	3.82	4.56	-	4.19
Kumano et al. (PF) [23]	4.2	7.1	2.9	4.73
Sung et al. (AAM+CT) [39]	5.6	5.4	3.1	4.7
Valenti et al. (CT) [44]	5.26	6.10	3.00	4.79
An & Chung (ET) [1]	7.22	5.33	3.22	5.26
Saragih et al. (CLM) [34]	4.5	5.2	2.6	4.1
Vincente et al. (SDM) [45]	6.2	4.3	3.2	4.6
This work (Dense 3D)	**2.66**	3.93	2.41	3.00

Table 1: Comparison of different head tracking results on the Boston University dataset. The numbers represent the mean absolute angular error of the head pose estimation in degrees. The accuracies of Cascia et al. [24] and Xiao et al. [47] are taken from [29]. Acronyms: CT - Cylindrical Tracker, DVF - Distance Vector Field, PF - Particle Filter, AAM - Active Appearance Model, ET - Ellipsoidal Tracker, CLM - Constrained Local Model, SDM - Supervised Descent Method.

The results demonstrated that the proposed method is able to estimate head pose with high accuracy, even under varying lighting conditions. The method achieved the best result for Pitch angle estimation, and the second best result overall. We note that the Yaw and Roll angle estimation slightly lower than the tracker proposed in [47], however our method is able to simultaneously estimate the head pose and reconstruct a dense 3D mesh of the face.

4.2. 3D Gaze Estimation on the MVG Dataset

We are interested in eye alignment and gaze estimation in a less constrained, social context, where moderate head-pose variations are present. In this experiment we evaluated the eye alignment precision on rotated images. We rendered views from the annotated 8000 meshes, covering ±30 yaw and ±20 pitch rotations in 10 degrees of increment, resulting in 280,000 training samples. Figure 2 shows some of the synthesized images with the ground truth eye and pupil markers.

In these experiments we evaluated the precision of relative eye gaze estimation, where the gaze direction is given in the head coordinate system in the form of spherical coordinates (θ is the azimuthal angle, ϕ is the polar angle and the radius were normalized to unit length).

This dataset consist of partial faces, that include the eye regions only. To be able to track these images we trained our 3D model and the cascade regressor on the BP4D-Spontaneous dataset using the same reduced ares. This area corresponds to the two compact clusters we obtained during the part-based model building.

We tracked the images and calculated the frontal views. Figure 3 shows the cumulative error distribution (CED) curves for the different markers. During the RMSE calculations the markers were normalized to have 100 pixel of inter ocular distance (IOD) in the 3D shape space. The tracker acquired high precision for each parts.

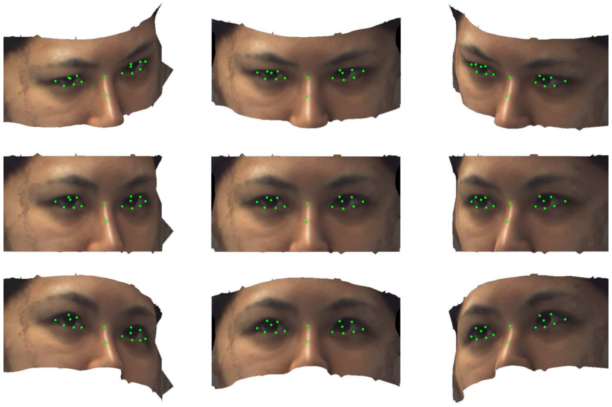

Figure 2: Rotated views and the corresponding markers.

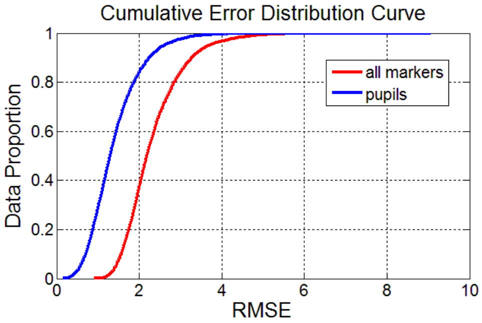

Figure 3: Cumulative error distribution curves on the rotated images. IOD were normalized to 100 pixels before the RMSE calculation. Note that the pupil markers can be estimated close to 1 pixel precision in average.

First, we used only the frontal renders and evaluated two different regression techniques. These were a Tikhonov regularized least square regression (LSR) [42] and linear Support Vector Regression (SVR) [11]. We extracted local binary features around the estimated landmarks using a 32x32 window. In each test we used a subject independent, five-fold cross validation. In the case of SVR regressor, we searched for the best parameter C between 2^{-10} and 2^{10} on a logarithmic scale with equidistant steps and selected the parameter having the lowest regression error on the remaining 4-folds.

The first part of Table 2 shows the gaze estimation results in mean absolute angular errors. Both methods achieved low recognition errors.

Encouraged by the results, in the next experiment we used all the rotated renders and evaluated the two different regression techniques in the same manner.

The second part of Table 2 shows the gaze estimation results in mean absolute angular errors.

Performance scores show that 3D gaze can be estimated with high precision and there were no significant differences between the frontal and rotated views.

	Frontal		Rotated	
	LSR	SVR	LSR	SVR
θ	4.1392	4.028	4.2232	4.0828
ϕ	4.4593	4.5453	4.6031	4.6919
Avg.	4.2992	4.2867	4.4132	4.3874

Table 2: 3D gaze estimation results on the MVG dataset using different type of regressors. The values are given in degrees.

	Frontal		Rotated	
	LSR	SVR	LSR	SVR
θ	3.8564	3.5133	4.4353	4.3404
ϕ	3.9462	4.2716	4.0983	3.8768
Avg.	3.9013	3.8925	4.2668	4.3874

Table 3: 3D gaze estimation results using different type of regressors on the CAVE Gaze dataset. The values are given in degrees.

4.3. 3D Gaze Estimation on the CAVE Dataset

The CAVE Gaze dataset comes with discrete gaze direction values given in the camera coordinate system. We used an 8-fold subject-independent cross-validation scheme to evaluate the different regression methods. In the first experiment we used only the frontal images for training and testing, in the second one we used all available images from the dataset.

The first part of Table 3 shows the absolute gaze estimation results in mean absolute angular errors using the frontal images only. The gaze can be estimated approximately 3.9 degrees of precision.

In the second experiment we repeated the procedure using all the available images in the dataset. The second part of Table 3 shows estimations for the different regressors. Average estimation error was 4.1 degrees.

Our results on the CAVE dataset is 1.5 degrees lower than the results we acquired on the Multi-view Gaze (MVG) Dataset [37]. This might be due to the image quality differences: the CAVE Gaze dataset was recorded with a 18-megapixel DSLR camera, comparing to PointGrey Flea3 cameras used in the MVG setup.

4.4. Low Resolution Experiment

In real-life scenarios we have to deal with low resolution images. In turn, we evaluated the previously trained 3D gaze regressors on different scales. In the previous experiments, all the training images were normalized to have 200 pixel of inter ocular distance (IOD) in the frontalized space. The IOD correspond to the distance between the eye centers. Note, that even if one the eye is occluded, we still can calculate the IOD based on the 3D shape.

In this experiment we varied the IOD of the testing images between 200 and 50 pixels. We also evaluated the performance, when both the training and testing images were normalized to have IOD 50 pixels.

Figure 4 depicts the resolution of the eye with different IODs and Figure 5 shows the CED curves for the different scenarios.

Table 4 shows the results for the different training and testing resolutions. Even with the lowest resolution, high precision estimation is possible.

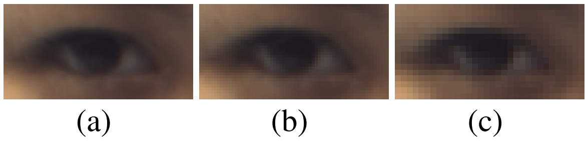

Figure 4: The resolution of the eye with (a) IOD 200 pixels, (b) IOD 100 pixels and (c) IOD 50 pixels.

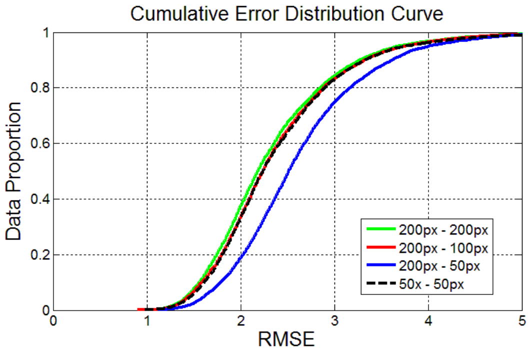

Figure 5: CED curves for the different scales. Green, red and blue curves show the performance, when the SDM were trained on IOD 200 pixel images, and tested on IOD 200 pixels, 100 pixels and 50 pixels images, respectively. The black dashed curve depicts the performance, where both the train and test images were normalized to IOD 50 pixels.

Image Sizes	Methods	
(training - testing)	LSR	SVR
IOD 200px - 200px	5.6277	5.3735
IOD 200px - 100px	5.7039	5.4487
IOD 200px - 50px	5.8223	5.5067
IOD 50px - 50px	5.8726	5.6265

Table 4: Recognition results using different training and testing image resolutions.

Sugano et al. [37] achieved a mean angular error of 6.5 degrees on the MVG dataset by using a learning-by-synthesis method for person-independent, 3D gaze estimation from low resolution images.

In our results, the last two rows in Table 4 correspond to the same conditions, where the images were normalized to have 50 pixels IOD. Our method achieved 5.6 degrees of error in average.

4.5. Varying illumination

Up until now, we used only uniform illumination in the training and testing set. In this experiment we evaluated the robustness of the gaze estimation under varying illumination conditions. We re-synthesized the rotated renders and varied the level of ambient and directional diffuse light and repeated the experiment described in Section 4.2. In this experiment all training images were normalized to have 200 pixels IOD.

The Tikhonov regularized LSR and the SVR achieved 6.1324, 5.7764 mean absolute angular errors, respectively. Varying illuminations present a more challenging conditions for the estimation.

Note that in many real-life situations, the light source in a scene gives rise to a specular highlight on the eyes. Since the MVG dataset does not come with detailed specular maps, we could not include this condition in our current study.

5. Conclusions

To afford real-time, person-independent 3D gaze estimation from 2D images, we developed a dense cascade regression approach in which facial landmarks remain invariant across poses.

We frontalize the reconstructed 3D shape to a canonical view using the estimated 3D head pose. This eliminates the need of learning a manifold of appearance features for gaze estimation. High precision can be achieved for 3D gaze estimation from a feature space that is robust to photometric variations and different scales. We used binary features that encode local ordinal properties of pixel intensities leading to obvious photometric invariant properties.

We validated the method in a series of experiments that evaluate its precision of 3D head pose estimation and 3D gaze estimation under varying head poses, resolutions and illumination changes. Experimental findings strongly support the validity of real-time, 3D gaze estimation from 2D images.

6. Acknowledgments

Preparation of this publication was supported in part by the National Institute of Mental Health of the National Institutes of Health under Award Number MH096951 and the Carnegie Mellon University People Image Analysis Consortium. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

References

[1] K. H. An and M. J. Chung. 3d head tracking and pose-robust 2d texture map-based face recognition using a simple ellipsoid model. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 307–312. IEEE, 2008.

[2] S. Asteriadis, D. Soufleros, K. Karpouzis, and S. Kollias. A natural head pose and eye gaze dataset. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, page 1. ACM, 2009.

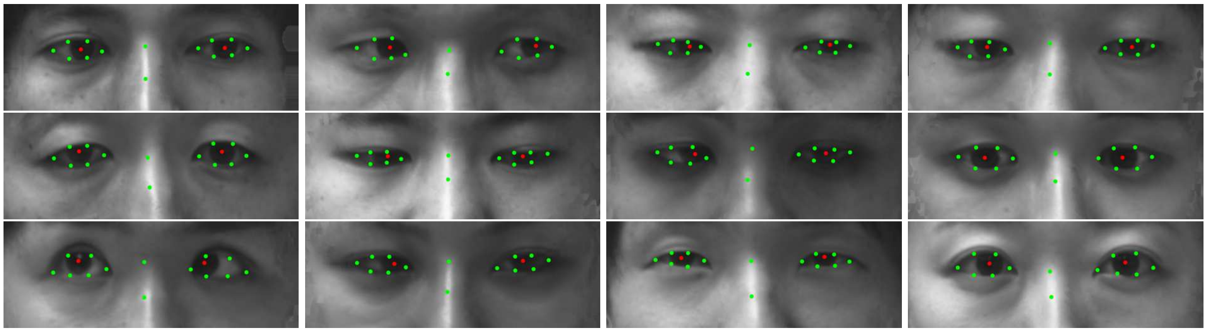

Figure 6: Automatic eye alignment results after the frontalization from the Multi-view Gaze dataset. The pupil markers are highlighted in red for display purpose.

[3] T. Baltrusaitis, P. Robinson, and L. Morency. 3d constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2610–2617. IEEE, 2012.

[4] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. Technical report, DTIC Document, 1994.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[6] X. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1513–1520, Dec 2013.

[7] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua. Brief: Computing a local binary descriptor very fast. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1281–1298, July 2012.

[8] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43:1–43:10, July 2014.

[9] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Trans. Graph.*, 32(4):41:1–41:10, July 2013.

[10] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2887–2894, June 2012.

[11] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[12] J. Chen and Q. Ji. 3d gaze estimation with a single camera without ir illumination. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[13] J. Chen and Q. Ji. Probabilistic gaze estimation without active personal calibration. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 609–616. IEEE, 2011.

[14] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[15] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System (FACS): Manual*. Salt Lake City (USA): A Human Face, 2002.

[16] K. A. Funes Mora and J. Odobez. Gaze estimation from multimodal kinect data. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 25–30. IEEE, 2012.

[17] E. D. Guestrin and E. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *Biomedical Engineering, IEEE Transactions on*, 53(6):1124–1133, 2006.

[18] D. W. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):478–500, 2010.

[19] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. *arXiv preprint arXiv:1411.7964*, 2014.

[20] T. Ishikawa, S. Baker, I. Matthews, and T. Kanade. Passive driver gaze tracking with active appearance models. In *In Proceedings of the 11th World Congress on Intelligent Transportation Systems*, 2004.

[21] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *11th IEEE International Conference on Automatic Face and Gesture Recognition*.

[22] C. L. Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[23] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato. Pose-invariant facial expression recognition using variable-intensity templates. *International journal of computer vision*, 83(2):178–194, 2009.

[24] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(4):322–336, 2000.

[25] N. Li and C. Busso. Analysis of facial features of drivers under cognitive and visual distractions. In *Multimedia*

and Expo (ICME), 2013 IEEE International Conference on, pages 1–6. IEEE, 2013.

[26] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Inferring human gaze from appearance via adaptive linear regression. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 153–160. IEEE, 2011.

[27] F. Lu, Y. Sugano, T. Okabe, and Y. Sato. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1008–1011. IEEE, 2012.

[28] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[29] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):607–626, 2009.

[30] U. J. Pfeiffer, B. Timmermans, G. Bente, K. Vogeley, and L. Schilbach. A non-verbal turing test: differentiating mind from machine in gaze-based social interaction. *PloS one*, 6(11):e27591, 2011.

[31] U. J. Pfeiffer, K. Vogeley, and L. Schilbach. From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, 37(10):2516–2528, 2013.

[32] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1685–1692, June 2014.

[33] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Face frontalization for alignment and recognition. *arXiv preprint arXiv:1502.00852*, 2015.

[34] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2).

[35] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280. ACM, 2013.

[36] Y. Sugano, Y. Matsushita, and Y. Sato. Appearance-based gaze estimation using visual saliency. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(2):329–341, 2013.

[37] Y. Sugano, Y. Matsushita, and Y. Sato. Learning-by-synthesis for appearance-based 3d gaze estimation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1821–1828, June 2014.

[38] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483, June 2013.

[39] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[40] K.-H. Tan, D. Kriegman, and N. Ahuja. Appearance-based eye gaze estimation. In *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*, pages 191–195. IEEE, 2002.

[41] J. R. Tena, F. De la Torre, and I. Matthews. Interactive region-based linear 3d face models. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, pages 76:1–76:10, New York, NY, USA, 2011. ACM.

[42] A. N. Tikhonov. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer, 1995.

[43] Tobii Eye Tracking Research. (2014). Eye tracking products. November 13, 2014. `http://www.tobii.com/en/eye-tracking-research/global/products/`.

[44] R. Valenti, N. Sebe, and T. Gevers. Combining head pose and eye location information for gaze estimation. *Image Processing, IEEE Transactions on*, 21(2):802–815, 2012.

[45] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, and D. Levi. Driver gaze tracking and eyes off the road detection system.

[46] O. Williams, A. Blake, and R. Cipolla. Sparse and semi-supervised visual mapping with the sˆ 3gp. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 230–237. IEEE, 2006.

[47] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology*, 13(1):85–94, 2003.

[48] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539, June 2013.

[49] H. Yamazoe, A. Utsumi, T. Yonezawa, and S. Abe. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 245–250. ACM, 2008.

[50] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692 – 706, 2014. Best of Automatic Face and Gesture Recognition 2013.

[51] A. Ziegler, E. Christiansen, D. Kriegman, and S. J. Belongie. Locally uniform comparison image descriptor. In *Advances in Neural Information Processing Systems*, pages 1–9, 2012.