A 3D Morphable Model of the Eye Region

Erroll Wood[1], Tadas Baltrušaitis[2], Louis-Philippe Morency[2], Peter Robinson[1], and Andreas Bulling[3]

[1]Computer Lab, University of Cambridge, United Kingdom
[2]Language Technologies Institute, Carnegie Mellon University, United States
[3]Perceptual User Interfaces, Max Planck Institute for Informatics, Germany

Figure 1: *Fitting our morphable model to an image: Given an input image and facial landmarks (a), we first initialize our model (b). We then use analysis-by-synthesis to optimize shape, texture, pose and illumination parameters simultanously to match the observed image (c). Once both eyes are fit, they can be posed to re-target perceived eye gaze.*

Abstract

We present the first 3D morphable model that includes the eyes, enabling gaze estimation and gaze re-targetting from a single image. Morphable face models are a powerful tool and are used for a range of tasks including avatar animation and facial expression transfer. However, previous work has avoided the eyes, even though they play an important role in human communication. We built a new morphable model of the facial eye-region from high-quality head scan data, and combined this with a parametric eyeball model constructed from anatomical measurements and iris photos. We fit our models to an input RGB image, solving for shape, texture, pose, and scene illumination simultaneously. This provides us with an estimate of where a person is looking in a 3D scene without per-user calibration – a still unsolved problem in computer vision. It also allows us to re-render a person's eyes with different parameters, thus redirecting their perceived attention.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—Gaze Estimation

1. Introduction

Eyes and their movements convey our attention, and communicate social and emotional information [Kle86]. They are important in graphics, as virtual humans must appear realistic and engaging; and in computer vision, as we wish to estimate gaze or emotional state. Morphable face models are a powerful tool, being used in face recognition [PKA*09], avatar animation [CWLZ13], and expression re-targetting [TZN*]. However, previous work either portrays eyes as static geometry [PKA*09], or avoids them entirely by removing them from the mesh [CWLZ13, TZN*]. This is because the complex structure and movements of eyes are very challenging to model realistically.

We present the first 3D morphable model (3DMM) that includes the eyes, allowing us to model variation in facial appearance as well as eyeball pose. By fitting our 3DMM to an image, we can estimate gaze under challenging head-pose or illumination conditions. We can also re-target where someone is looking. This could be used for maintaining eye-contact during video-conferencing, or avoiding someone looking at a camera during filming (see Figure 1).

2. Synthesizing Images of the Eye Region

Our goal is to use our 3DMM to synthesize an image which matches an input RGB image. To render synthetic views of the eye region, we used parametric models of the facial eye region and eye-

ball, and a model of image formation. Our total set of model and scene parameters are $\Phi = \{\beta, \tau, \theta, \iota, \kappa\}$, where β are the eye region shape parameters, τ the texture parameters, θ the pose parameters, ι the illumination parameters, and κ the camera parameters. This leads us to 37 total parameters in our model.

2.1. Parameterized Eye Region and Eyeball

We built a generative model of the facial eye region by manually registering high resolution head scan meshes [WBZ*15] into a low resolution topology, containing the eye region only [WBM*16]. We represent color by using a texture map, allowing us to couple our efficient low-resolution mesh with a high-resolution texture. Once the scans have been brought into correspondence, we build linear models of shape $\mathcal{M}_s$ and texture $\mathcal{M}_t$ using *principal component analysis*. This allows us to generate 3D eye regions using our shape and texture parameters: $\mathcal{M}_s(\beta)$ and $\mathcal{M}_t(\tau)$. The eyeball is represented as a separate mesh constructed from standard anatomical measurements. We model iris color variation with a linear texture model $\mathcal{M}_{iris}(\tau)$ built from a set of aligned iris photos.

Both global and local pose information is stored in θ. The position and orientation of the eye region is given by its model-to-world transform, and the eyeball's rotation is defined by additional parameters θ_{pitch} and θ_{yaw}. When the eye looks up or down, the eyelid follows it – this is modelled using procedural geometric animation based on anatomic measurements [WBM*16]. As our eye region is a multi-part model, we also *shrinkrwap* the eyelid skin to the eyeball geometry, avoiding unwanted gaps or clipping issues. Finally, for gaze re-targetting, we render a transparent eyelash mesh controlled by a small number of guided hair particles.

2.2. Illumination and Image Formation

To complete the rendering process, we also model illumination and camera projection. We assume all materials are Lambertian, and model illumination (ι) as a simple combination of an ambient light and directional light. We fix the camera at the world origin, and assume knowledge of intrinsic camera calibration parameters (κ).

3. Fitting our Eye Region Model

Given an observed image I_{obs}, we wish to produce a synthesized image $I_{syn}(\Phi^*)$ that best matches it. We search for optimal model and scene parameters Φ^* using *analysis-by-synthesis*. To do this, we iteratively render a synthetic image $I_{syn}(\Phi)$, and compare it to I_{obs} using our energy function. We cast the problem as an unconstrained energy minimization problem for unknown Φ.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E(\Phi) \tag{1}$$

Our energy is formulated as a combination of a dense *image similarity metric*, and a sparse *landmark similarity metric*, with λ controlling their relative importance.

$$E(\Phi) = E_{image}(\Phi) + \lambda \cdot E_{ldmks}(\Phi, L) \tag{2}$$

Image Similarity Metric The primary goal for our optimization is to minimize the difference between I_{syn} and I_{obs}. I_{syn} contains a set of rendered foreground pixels P that we wish to compute image error over, and background pixels that we wish to ignore. We compute image similarity as the average absolute difference between foreground pixels $p \in P$.

$$E_{image}(\Phi) = \frac{1}{|P|} \sum_{p \in P} |I_{syn}(\Phi, p) - I_{obs}(p)| \tag{3}$$

Landmark Similarity Metric The face contains *landmark* points that can be localized reliably. We use a face tracker to localize 14 landmarks L around the eye region in image-space [BMR13]. For each landmark $l \in L$ we compute a corresponding synthesized landmark l' from our 3DMM. This energy is calculated as the distance between both sets of landmarks, and acts as a regularizer to prevent our pose θ from drifting too far from a reliable estimate.

$$E_{ldmks}(\Phi, L) = \sum_{i=0}^{|L|} \|l_i - l'_i\| \tag{4}$$

3.1. Optimization Procedure

Fitting our models is a challenging non-convex and high-dimensional optimization problem. To approach it we use gradient descent with an annealing step size. As calculating analytic derivatives for a scene as complex as ours is challenging, we use numerical central derivatives. Their efficient computation is made possible through the use of a tailored DirectX GPU rasterizer that can render I_{syn} at over 5000fps.

4. Conclusion

We have presented the first multi-part 3DMM that includes the eyes. Our model not only allows us to estimate the eye gaze, but also to retarget the perceived gaze in a photorealistic manner.

References

[BMR13] BALTRUŠAITIS T., MORENCY L.-P., ROBINSON P.: Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE ICCVW* (2013). 2

[CWLZ13] CAO C., WENG Y., LIN S., ZHOU K.: 3D shape regression for real-time facial animation. *ACM TOG* (2013). 1

[Kle86] KLEINKE C. L.: Gaze and eye contact: a research review. *Psychological bulletin 100*, 1 (1986), 78–100. 1

[PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3D Face Model for Pose and Illumination Invariant Face Recognition. *Proc. AVSS* (2009). 1

[TZN*] THIES J., ZOLLHÖFER M., NIESSNER M., VALGAERTS L., STAMMINGER M., THEOBALT C.: real-time expression transfer for facial reenactment. 1

[WBM*16] WOOD E., BALTRUŠAITIS T., MORENCY L.-P., ROBINSON P., BULLING A.: Learning an appearance-based gaze estimator from one million synthesised images. In *Proc. ETRA* (2016). 2

[WBZ*15] WOOD E., BALTRUŠAITIS T., ZHANG X., SUGANO Y., ROBINSON P., BULLING A.: Rendering of eyes for eye-shape registration and gaze estimation. In *ICCV* (2015). 2