Joint Estimation of Face and Camera Pose from a Collection of Images

David Greenwood[1]
David.Greenwood@uea.ac.uk
Sarah Taylor[1]
S.L.Taylor@uea.ac.uk
Iain Matthews[2]
Iain.Matthews@uea.ac.uk

[1] School of Computing Science, University of East Anglia
[2] Epic Games, Pittsburgh

Morphable models to represent faces have a rich presence in the computer vision literature, stemming from the seminal work of [1]. A 3DMM is some combination of a shape and appearance model, much like Active Appearance Models (AAMs), where AAMs may also include a method of fitting the model [4]. One may gain an impression that deriving 3D shape and appearance has become trivial with the proliferation of consumer focused applications. However, high accuracy fitting and tracking of faces retains considerable motivation in, for example, the games and motion picture industry. Emerging use cases include projecting ones presence in virtual environments and interactive interfaces using realistic avatars. Many of the more impressive examples of accurate model fitting involve elaborate multi-view stereo camera rigs, that are expensive, complex to calibrate and produce data at a rate that is difficult to manage. In contrast, much of the data available to the public domain is originated by simpler means. Examples of this data can be found on video sharing sites, or can be collected using more modest equipment. The compromise that must be made is to accept a number of unknowns, that may include lighting variation or camera parameters. We present a work in progress method to fit a parametric shape model to a collection of images of a subject. We jointly solve for the shape *and* camera parameters, and develop the appearance by completing the UV texture map. Our method does not require training weights, or manual annotation of landmarks, and is particularly useful if no camera calibration is available.

Aligning a 3D surface with an image is an ill-posed problem. For rigid objects, the solution can be recovered using Structure from Motion (SfM), usually by finding corresponding points in a number of images and solving the shape and the camera views as a bundle. Deforming objects, such as faces, present greater difficulties, so a statistical shape model can provide a significant prior. By projecting the view of the shape to an unwrapped UV space, the alignment of corresponding features across camera images can be represented as a minimisation of Euclidean distance. Each point in a UV triangle represents a unique 3D point on the surface of the shape, provided the geometry and UVs are well formed, without overlap. A useful by-product of working in UV space is the completion of the UV texture map, and appearance model.

We choose arguably the most complete parametric face shape model, FLAME [3], extract the data, and re-implement the pose function using PyTorch [5] to take advantage of fast automatic differentiation. Using the same framework, we also implement a perspective pinhole camera model, with trainable intrinsic and extrinsic parameters (we do not, at this time, model distortion). The third component in our method is a differentiable rasterizer, to project the camera views to UV space. Together, our system allows us to jointly solve the camera parameters and the face parameters by back propagation from our loss function. We retain the separation of identity and expression, solve for a single identity, but possibly many expressions. Our method makes use of off-the-peg face landmarks [2] to solve the gross alignment of camera parameters to camera images. No assumptions about camera parameters need be made, although we can take advantage of prior knowledge, for example, if a camera is fixed. These facial landmarks share semantics that can be located directly on the shape surface, for example lip outlines, and these form the first term in our loss function. In UV space we extract dense CNN features using keypoint re-localisation [6]. Using the last max pool layer of VGG16 gives the maximal feature in a 32 pixel region. Minimising the distance between the corresponding maximal feature in each of the projected views is the second term of our loss function. Each fitting starts with zero knowledge of the camera images. We use Adam with a learning rate of 0.01. Convergence takes approximately 200 function evaluations, and is somewhat dependent on the number n of images chosen. We have fitted the shape to collections of $n = 2$ to 12, both in a single shape expression and n camera views, or n expressions and n camera views.

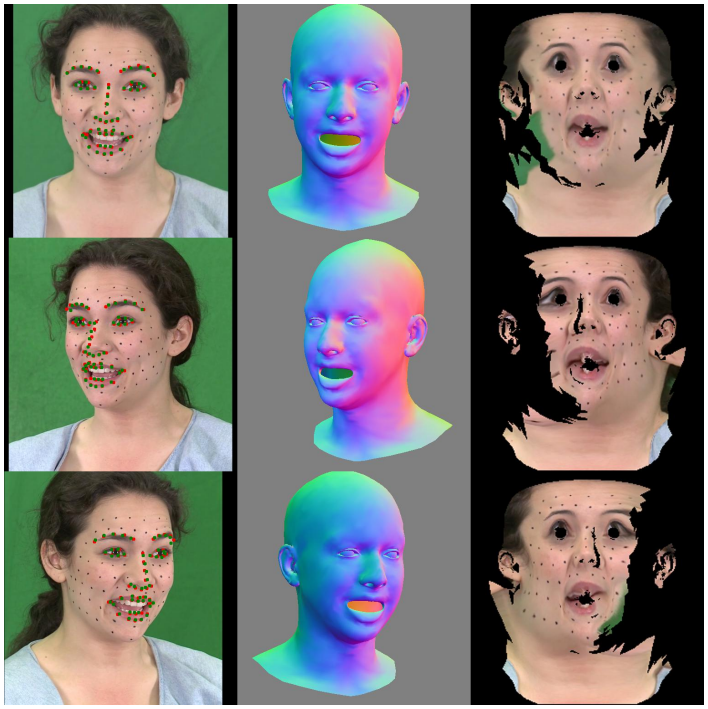

Figure 1: From a collection of images of an identity, we fit a shape model by aligning facial landmarks and features projected to UV space. Projecting to UV space allows UV texture completion and appearance modelling for further fitting or tracking.

Once a good fit has been achieved, the UV texture map can be combined for use in a conventional rendering pipeline, or each projection can contribute to an appearance model, giving a subject dependent AAM for fitting and tracking further images. This is especially useful for data collected 'in the wild' where generalised appearance modelling is particularly challenging.

[1] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[2] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[3] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36 (6), 2017.

[4] Iain A. Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[6] Aji Resindra Widya, Akihiko Torii, and Masatoshi Okutomi. Structure from motion using dense cnn features with keypoint relocalization. *IPSJ Transactions on Computer Vision and Applications*, 10:6, May 2018.