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

David Greenwood[1]
Sarah Taylor[1]
Iain Matthews[2]

Introduction

Morphable models to represent faces have a rich presence in the computer vision literature, yet fitting such models to unconstrained images remains a challenging problem.

We present a method to fit a parametric shape model to a collection of images of an individual.

Our method does not require training weights, or manual annotation of landmarks, and is particularly useful if no camera calibration is available.

Method

Our goal is to find the parameters for a shape model, for a collection of images of a single identity. Our shape model is Flame [1].

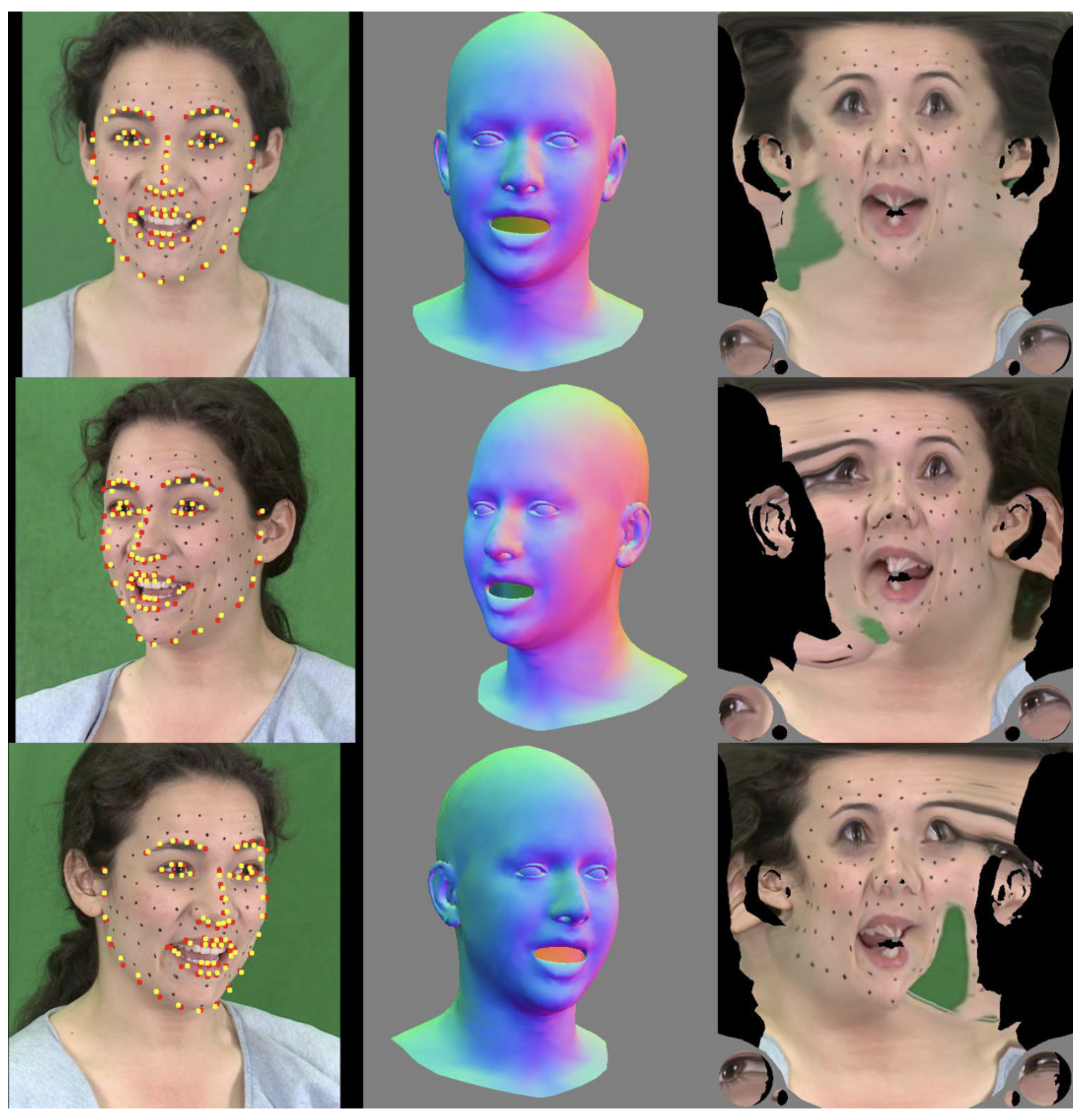
We jointly solve for the shape parameters and a camera model for each image. The camera parameters are used to project the shape to UV space with a differential rasteriser.

Exploiting Image Knowledge

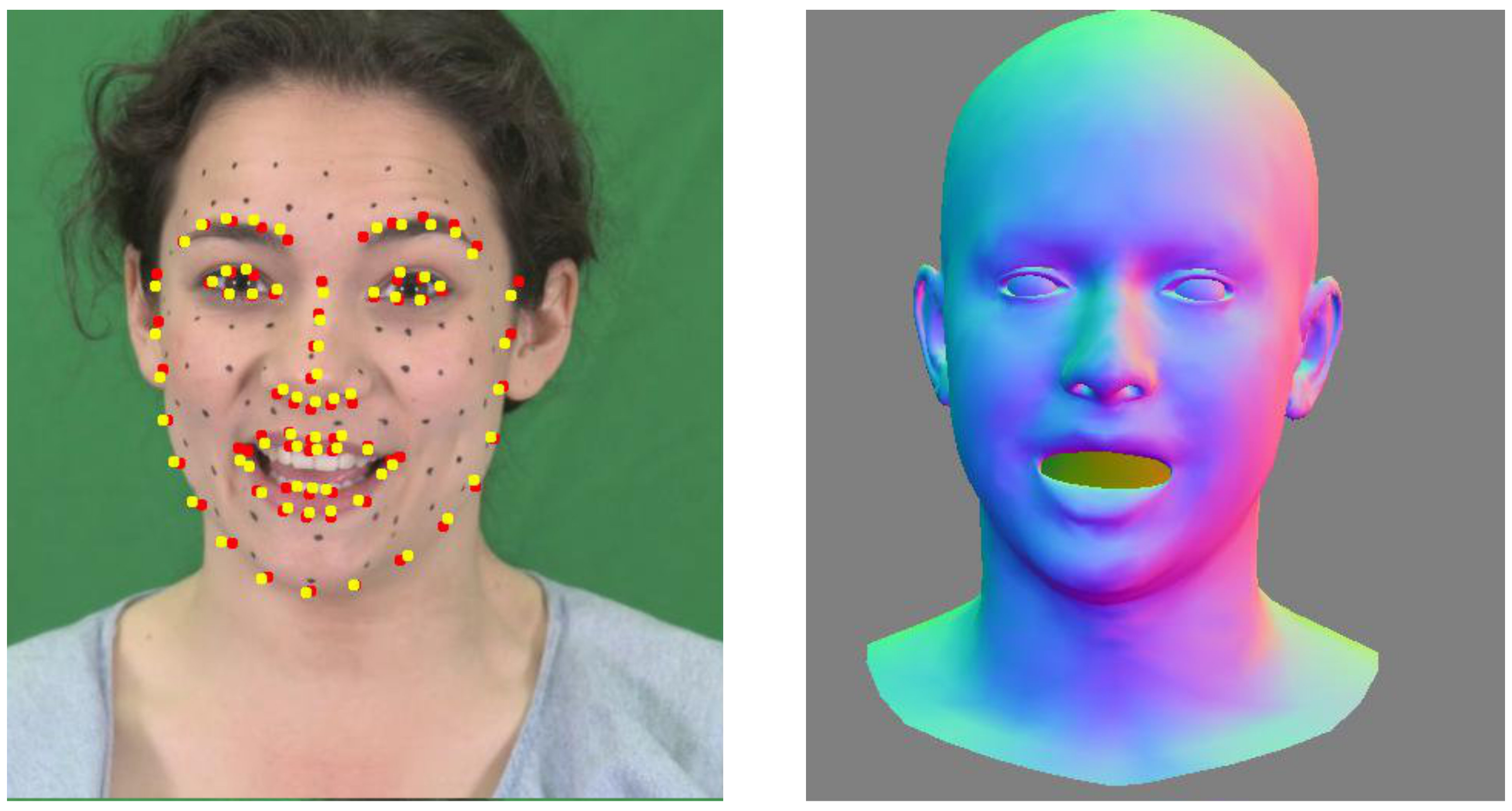
It is possible to take advantage of knowledge of the capture of the group of images.

In this example, we know that the images are from synchronised cameras, so each image shows both a single identity and a single expression.

We can share both the identity and the expression parameters when solving for all images.



Landmark Loss

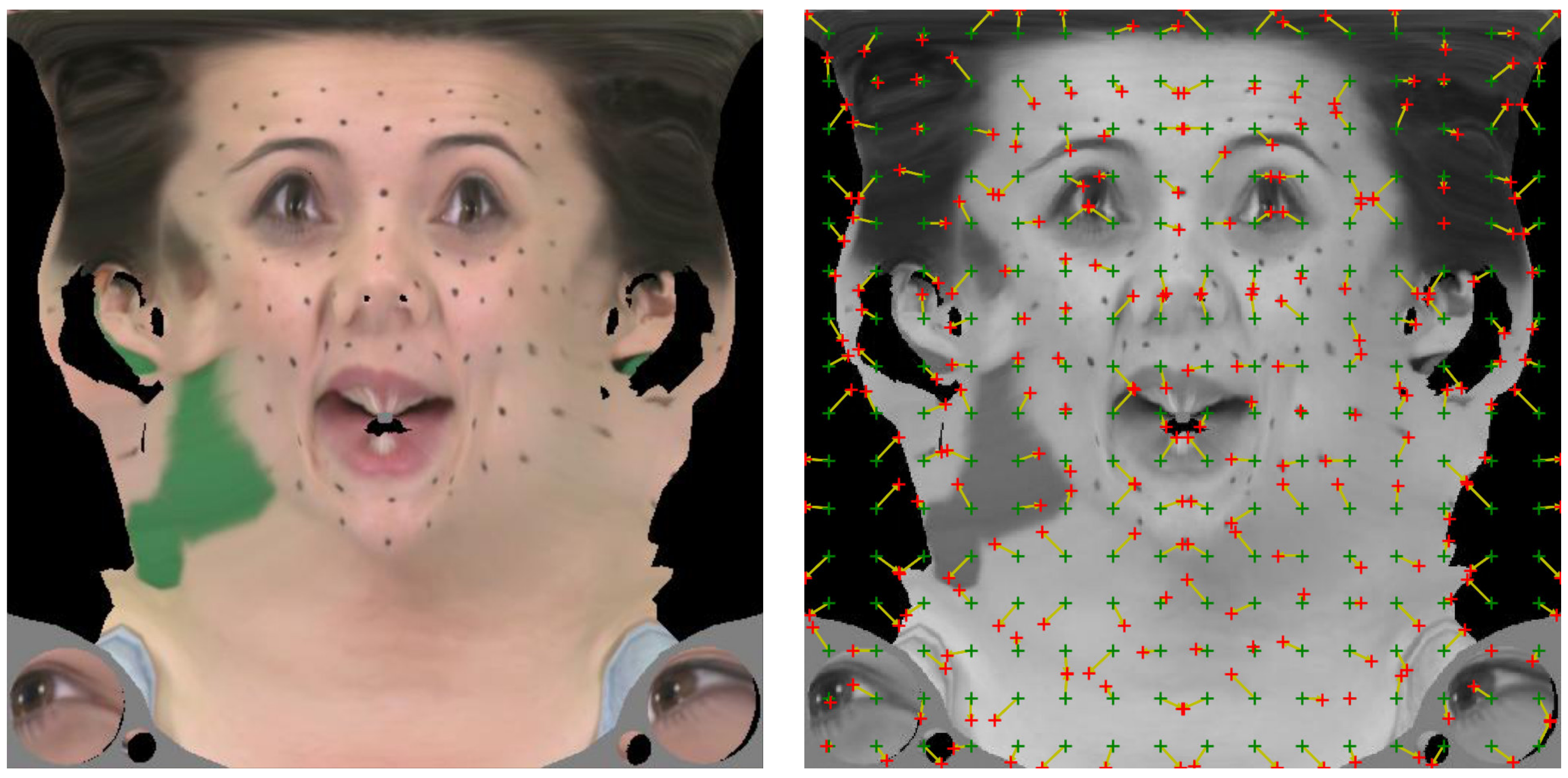


Off-the-peg tools are used to find sparse facial landmarks. These landmarks share semantics that can be located directly on the shape surface. To find the first term *A* in our loss function:

- find *n=68* landmarks in the image
- project the shape landmarks to image space
- normalise the image space to [0, 1]
- mean squared distance between landmarks for all images

$$A = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

CNN Feature Loss



In UV space we extract dense CNN features using key point re-localisation [2].

We follow the maximal pixel through layers of VGG16, by successively re-localising at each max-pool layer.

The result is the most active pixel within each 32 pixel square of a uniform grid over the UV image.

The second term in our loss function is then the mean squared distance between each corresponding feature, for all UV images.

$$B = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

Combined Loss

The loss terms are combined with a scaling coefficient, setting the landmark loss to be initially more significant.

$$\lambda A + \gamma B$$

We back propagate the loss through the UV projection and landmark projection to optimise the shape and camera parameters.

We use Adam with a learning rate of 0.01. Convergence takes approximately 200 function evaluations, depending on the number of images involved.

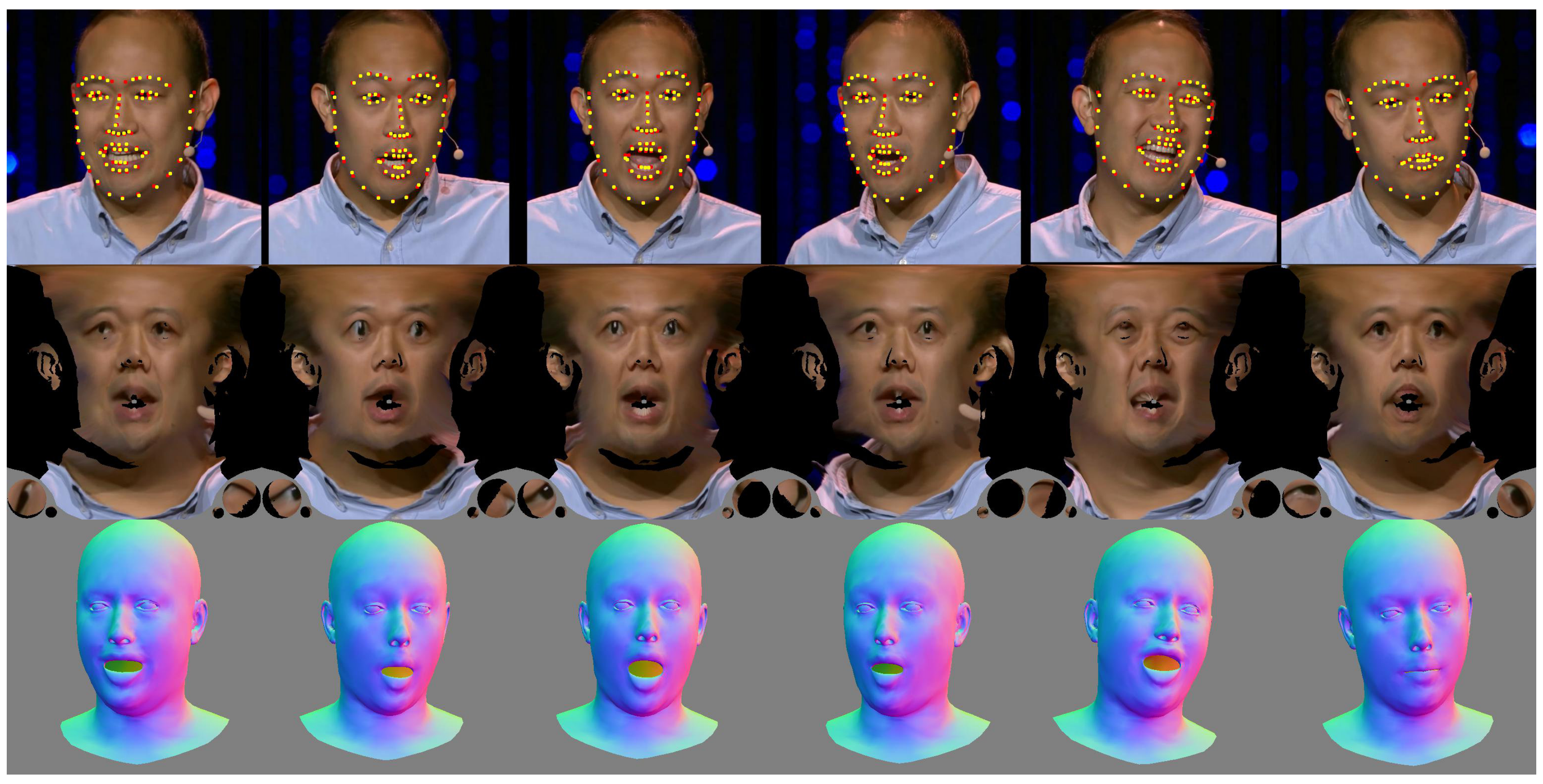
Unconstrained Images

In this example, we show a collection of images randomly extracted from an online video sequence.

Aside from ensuring the images belong to a single identity, no assumptions are made regarding the camera parameters.

In fact we expect that the images may be derived from cuts from several cameras, have camera motion, and cropping or zooming.

Hence, we solve for each image the intrinsic and extrinsic parameters, and an expression parameterisation.



References

[1] 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

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

[1] University of East Anglia
[2] Epic Games

University of East Anglia