Stirling /ESRC 3D face database. Capture and naming

3D imagery

3D images are captured with a Di3D imager (www.di3d.com), using four Canon EOS 550D cameras arranged as two stereo pairs. File format is wavefront obj. The raw, unconformed files are presented at half the available resolution, about 85000 vertices. The conformed files are reduced a standard face mesh with 3746 vertices.

The camera system has an issue with eyes. Because the eye surface is translucent, it cannot be located properly and the eye surface is interpreted as concave rather than convex. This cause a psychologically disturbing distortion of the projected image – people appear cross-eyed, and the pupil is often not circular. We are working on a fix for this, by enforcing the eye surface to be hemi-spherical. If anyone reading this has a solution, please tell us!

Lighting for the 3D images is provided by three flash softboxes: one above and one either side of the cameras.

Subjects were photographed wearing a white cap, to contain their hair, which does not render well in 3D – wisps of hair cause wild fluctuations in the computed surface. They were first photographed with a neutral expression, then with a smile, moth closed, then a full smile, then, prompted by Ekman-posed expressions from the Amsterdam set, asked to pose anger, disgust, fear, unhappy (sad) and surprise. The mouth-closed smile is because, like eyes, teeth render poorly in 3D, appearing bent.

A note about obj format. There are many flavours of this format: different programs differ in what they will open. In general, there are three files per image: an obj that specifies the 3D points, an image, typically jpg or bmp, and a .mtl material file which joins the two and gives some other information. This mtl file is not strictly needed and some programs don't bother, saving only an obj and a jpg file and assuming that the two will have the same name.

Photographic images.

Blue background images

While in the 3D imaging room, the same flash lighting and background was used to capture photographs using a Canon G3 camera at 90cm distance from the target. This has a native resolution of 2272x1704 pixels, and was set to f5 at 1/250 second, and, bar the occasional error, at 13 mm focal length. The images were cropped to 1000x1400 pixels, centred on the bridge of the nose. Usually, four images were captured; with and without a cap; neutral and smiling. Depending on hair style, there may be additional images with hair down, and those wearing spectacles were usually photographed with and without them.

Main sequence, varying pose and lighting, plus stereo images

The main sequence of images was taken using four cameras fired simultaneously, set at 0, 22.5, 45 and 90 degrees to the main source of lighting. Figure 1 and Figure 2 illustrate the layout.

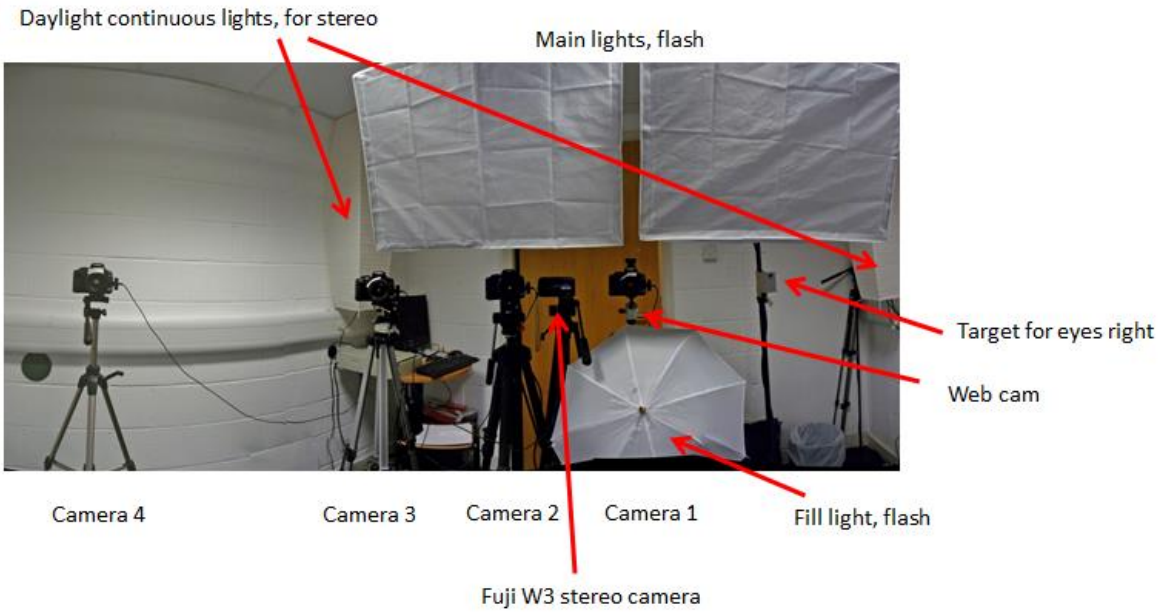

Figure 1 layout of cameras, pictured from the subject's viewpoint.

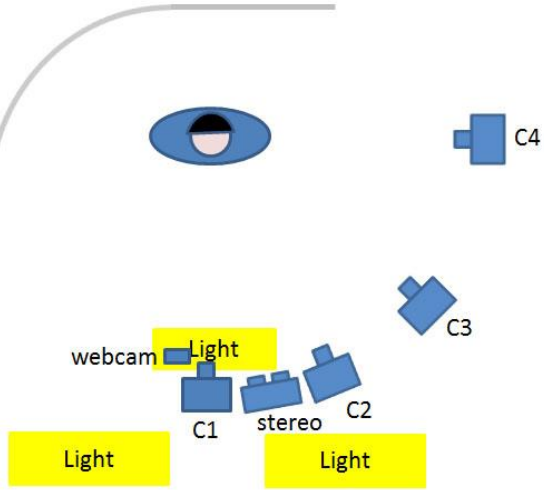

Figure 2 schematic arrangement of cameras, from above.

The sequence was as follows: participants were first photographed with a Fuji W3 stereo camera, using diffuse daylight fluorescent lighting, with and without a cap, neutral and smiling. They were then videoed using the Fuji, asked to move their head left and right, up and down, and then explain how to get to the campus library.

The daylight lights were switched off, leaving only dim overhead lighting. The camera array consisted of four Canon EOS 400D cameras, triggered simultaneously. 50mm lenses were pre-focussed on the sitter; exposure was 1/10 second at f22. Despite this slow shutter, a camera occasionally missed the flash, and sometimes the photographer failed to notice, so some shots are missing.

The sequence, in general (photographers sometimes got out of order), was as follows: neutral and happy, wearing a cap; then neutral and posed happy, angry, disgust, fear, unhappy (sad) and

surprise, facing forward, without cap. The expressions were prompted by showing an example from the Amsterdam Ekman-posed set. Then, facing forward, neutral, eyes forward and then 22 degrees right and left; then face camera 2, eyes forward, 22 degrees right and left, then face camera 3, then face camera 4. Then put on a black gown and hold a standard colour reference chart. These colour reference images are available at the full resolution, as taken. Put down the chart and face each camera in turn. The Camera 1 images are available as Canon Raw format, in case anyone needs uncompressed images. Finally, put the cap back on and face each camera in turn. There are a few extra images, some accidental exposures, some where the expression was not as intended but still interesting.

Note that for the eyes gaze images, the second set has them facing camera 2, and looking at both camera 1 and camera 3. The camera 1 and 3 images should be present also, meaning you have all combinations of gaze and viewpoint.

Images were captured at 3888x2592 pixels. A square crop was made to approximately centre the face in the frame, using the whole height of the image; this was then rescaled to 1200 pixels square. Relative head size of different people should therefore be maintained, especially from camera 4 – people did shift their position forward and backward somewhat, especially when trying to pose anger (forward) or fear (back). However, note that due to such vagaries of positioning and the precise focal length of lenses used, there is up to about a 10% variation in the size of a given person's head from different views.

Set taken under varied distance and lighting

Participants were next photographed in three locations, two indoors and one outdoor. Sometimes one of these is missing – for example it was sometimes too dark outside! In general the sequence was as follows: with slightly side lighting indoors, photograph at 70 mm and 130 mm focal length, under more overhead lighting, photograph at 70, 130 and 200 mmm focal length, then outdoors, in shade, with a foliage background, photograph at 70 and 130mm focal length. Shooting distance was approximately at 70mm, at 130mm and at 200mm, the idea being pretty much to fill the frame with the head, portrait orientation. Since the camera was hand held, some images were lost due to camera shake, some are included but are somewhat blurred. Images were minimally cropped and then resized to 800x1200 pixels.

Video files

We took five different bits of video of most participants. Both the 3D and the main sequence photography sessions were recorded by a standard resolution webcam. These are deliberately low quality, intended for testing the equivalent of recognition from CCTV. The sequence taken in the 3D camera room, labelled V1, is rather poorer quality than intended; for some reason the camera insisted on over-exposing, irrespective of any settings. We recorded a video of the person walking the length of a corridor, at their own pace, at somewhat higher but still low quality: using a Fuji digital camera. These walking videos have had their contrast reduced somewhat to soften the overhead lighting.

Using the Fuji W3 camera, we recorded a stereo video, of the person looking to left and right, up and down, then explaining how to get to the library from the current location: i.e both rigid and non-rigid motion.

Finally, we recorded a high-definition video, using a Sony camcorder. We used a half-silvered mirror video tunnel, which allowed us to record the participant from head-on, while they watched things on a computer monitor,

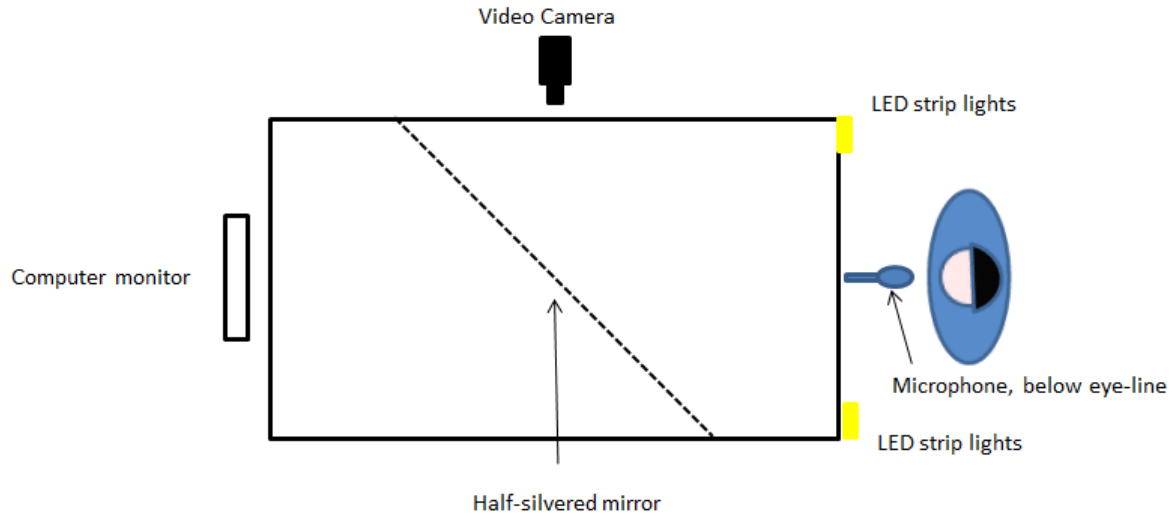

Figure 3.

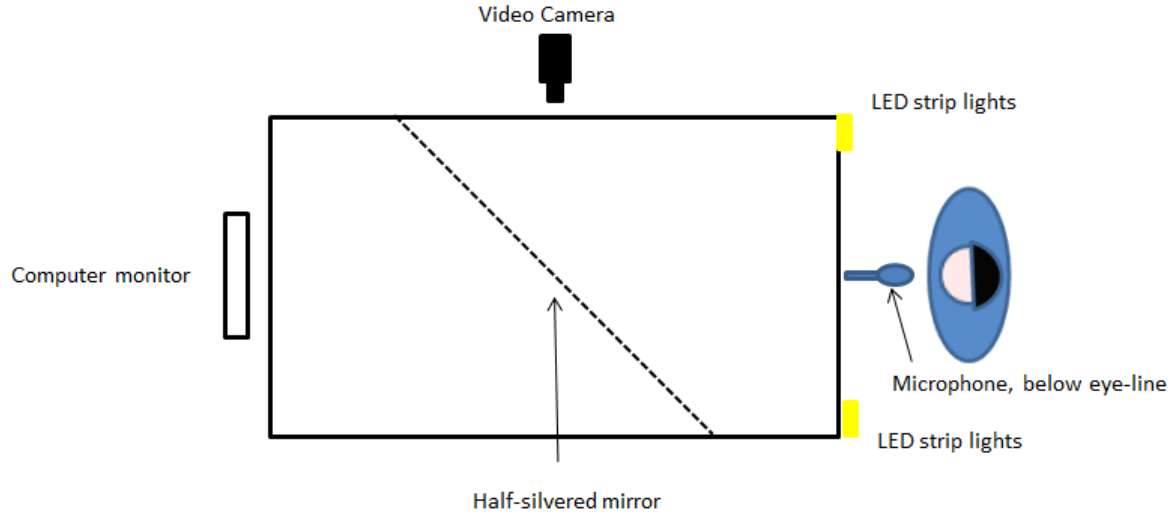

Figure 3 half-silvered mirror video capture setup

Participants were first asked to read out four sentences:

> The smell of freshly ground coffee never fails to entice me into the shop
>
> The length of her skirt caused the passers-by to stare
>
> They launched into battle with all the forces they could muster
>
> The most important thing to remember is to keep calm and stay safe.

They were then asked to describe their dream house, in about 30 seconds. A couple of participants took several minutes! They were then asked to rotate their head to the left and right, then up and down. They were next asked to imitate expressions, using models from the Amsterdam Ekman-posed set, anger, fear, surprise, sad, disgust and happy, plus a locally posed pain expression.

They were finally asked simply to watch a serious of video clips, chosen to elicit natural expressions. Some people responded more than others! The sequence lasted about 20 minutes; the scenes were as follows, with intended emotion in brackets:

A dog howling 'I love you' (happy)

Some puppies being thrown into a stream for real, presumably to drown them. (anger, disgust, horror)

A woman in an office talking too loud on the phone, which is then smashed by a colleague. (surprise, happy)

Someone logging into a computer, which makes a large explosion sound (surprise)

Bear Grylls eating some huge grubs from a log. (disgust)

A scene from Arachnophobia, where spiders come out of someone's mouth. (disgust, fear)

A video of the amputation of an arm (disgust)

A clip from 'Sea of Love', where a pigeon suddenly flies off (surprise)

A video of a man attempting a trick where he slaps a crocodile's mouth, but which bites his arm. (surprise, fear)

A video of a man luring a huge crocodile out of a river onto a muddy bank. (fear)

An edit of 'The Silence of the Lambs' sequence where Starling is trying to find the killer, in the dark. (fear)

The massacre sequence from 'Cry Freedom' (anger, disgust)

A video of Kate Moss and daughter being hounded by paparazzi at LA airport. (anger, disgust)

The death scene from 'The Champ' (sad)

'Who's going to take you home' – the song 'Drive' as used in Live Aid, 1985, to images of the refugee camp in Ethiopea. (sad)

Some clips from 'you've been framed', intended to lighten the mood at the end. (happy)

We have, frankly, barely scratched the surface of editing these videos. Each of them is about 2.5Gb. If anyone would like to help, please get in touch.

Naming

Blue background images

There were usually four images taken in the 3D room, with a blue background. These are indicated with a letter B, for blue: M1000_B_N, M1000_B_H_C. The expression is indicated by N for Neutral, H for happy (smile); the suffix C indicates a cap is worn. Some have an extra image or two, usually with hair down (suffix _H) or wearing spectacles (suffix _S). So someone happy with their hair down might be F1026_B_H_H, which is a bit clumsy, but makes sense, I hope!

Stereo images

Currently still in progress.

Main sequence images

Here is an example for an image from the main sequence: M1001_01_L0_V0S_N_CG

M1001: M indicates male, F, female. 1001 is the id number.

01 indicates a frame number, which is the same for all views of the same image. Thus, the four views of a neutral face looking at the front camera might be M1001_03_L0_V0S_N, M1001_03_L2_V2L_N, M1001_03_L4_V4L_N, M1001_03_L9_V9L_N. It also serves to differentiate two otherwise identical images, or two different attempts at an expression. The actual number does not have a meaning, other than reflecting the order in which they were taken; because of camera failures, etc, equivalent images will have different numbers across identities. There are cases where a camera failed without being spotted, in which case that particular image number will be missing for that view.

L0 is the lighting direction: L0 from camera 1, L2 from camera 2 (22.5 degrees), L4 from camera 3 (45 degrees) and L9 from camera 4 (90 degrees).

V0S is the view of the face. First character indicates the angle, 0, 2, 4, 6, 9, (for 0, 22.5, 45, 67.5 and 90 degrees), the second gives the direction relative to straight, S for straight, L for left and R for right. L means that the face is looking to the left of the camera as you look at it, and showing the left side of their face. See the table for the possible combinations. The 0 and S of V0S are redundant but keep the overall filename length the same.

Camera taking the picture:	Camera 1	Camera 2	Camera 3	Camera 4
Camera being looked at				
Camera 1	V0S	V2L	V4L	V9L
Camera 2	V2R	V0S	V2L	V6L
Camera 3	V4R	V2R	V0S	V4L
Camera 4	V9R	V6R	V4R	V0S

N is the expression. A: angry, D: disgust, F: fear, H: happy (smile), N: neutral, S: surprise, U: unhappy (sad)

The suffix may indicate wearing of a cap (C), a gown (G), spectacles (S) or with hair down (H). Where more than one occurs, they should be in alphabetical order, CGHS. Most images will not carry this suffix.

Alternatively, the suffix denotes eye position: EL, subject looking to their left; ER, looking to their right; E0, looking straight ahead; ES, eyes shut (not many have this). Images with eye direction do not have any of the other suffixes, e.g. a cap. Most images are 'E0' by default; only those that make up the eye movement group will be so labelled. So, for the set of images where the subject looks at camera 1, then to their right, then to their left, the camera 1 images might be called

M1001_10_L0_V0S_N_E0
M1001_11_L0_V0S_N_ER
M1001_12_L0_V0S_N_EL

The matching images from camera 2 would be:

M1001_10_L2_V2L_N_E0
M1001_11_L2_V2L_N_ER
M1001_12_L2_V2L_N_EL where this one would be looking at camera 2, while facing camera 1.

Varied distance and lighting set.

M1000_I1_130; M1000_O1_135. I indicates indoors, O, outdoors. 135 indicates the focal length, as recorded by the camera. The numbers are not necessarily sequential; duplicates and blurred images have been removed. Usually, the first two indoor images are under slightly side lighting and the next three more overhead. Lighting for the outdoor ones is uncontrolled, apart from being out of direct sunlight.

Movie files.

The webcam videos recorded during photography are labelled _V1, for the 3D camera set, and _V2, for the view/pose photography set. The walking videos are labelled _W. They are all .wmv format.

Peter Hancock

Last updated 13 August 2013