A FACS Valid 3D Dynamic Action Unit Database with Applications to 3D Dynamic Morphable Facial Modeling

Darren Cosker
Department of Computer Science
University of Bath
dpc@cs.bath.ac.uk

Eva Krumhuber
School of Humanities and Social Sciences
Jacobs University
e.krumhuber@jacobs-university.de

Adrian Hilton
Centre for Vision, Speech and Signal Processing
University of Surrey
a.hilton@surrey.ac.uk

Abstract

This paper presents the first dynamic 3D FACS data set for facial expression research, containing 10 subjects performing between 19 and 97 different AUs both individually and in combination. In total the corpus contains 519 AU sequences. The peak expression frame of each sequence has been manually FACS coded by certified FACS experts. This provides a ground truth for 3D FACS based AU recognition systems. In order to use this data, we describe the first framework for building dynamic 3D morphable models. This includes a novel Active Appearance Model (AAM) based 3D facial registration and mesh correspondence scheme. The approach overcomes limitations in existing methods that require facial markers or are prone to optical flow drift. We provide the first quantitative assessment of such 3D facial mesh registration techniques and show how our proposed method provides more reliable correspondence.

1. Introduction

Facial analysis using 3D models has become a popular research topic in recent years. Some of the primary benefits of such models include potentially improved robustness to pose and illumination changes during recognition [3], estimation of 3D facial shape from 2D images [2, 20], and motion capture [13]. Given this emerging popularity, a great need exists for rigorous and standardized 3D dynamic facial data sets that the computer vision community can use for experimentation.

There are a range of available data sets for 2D facial analysis – both static and dynamic – containing variation in pose, illumination, expression and disguise (e.g. see [19, 14, 17]. Expression recognition in particular is a highly active research area, with many works based on movement descriptions from the Facial Action Coding System (FACS) [12]. FACS was primarily introduced by psychologists to describe different configurations of facial actions or Action Units (AUs). FACS lists 44 AUs that form the basis of 6 prototypical facial expressions: happiness, sadness, fear, surprise, anger and disgust. Numerous attempts exist to classify these movements in both static and dynamic 2D sequences [17, 1, 19]. Perhaps the most thorough set collected to date is the Extended Cohn-Kanade Dataset (CK+) [17], which contains 593 sets of expressions with the peaks manually FACS coded to establish AU presence.

The ability to FACS code data automatically has a wide potential in social psychological research on the understanding of facial expressions. One major reason for this is that manual coding is highly time consuming and often not practical for long dynamic sequences. FACS is also now often used as the movement basis for 3D facial models in movies, making automatic analysis relevant to motion-capture and performance mapping [11]. However, while available data for 2D analysis is widespread, there are only a handful of 3D facial data sets available [8, 23]. Data sets portraying 3D dynamic movement are fewer still [22], do not contain AU level motions, and are not FACS coded.

There is therefore clearly a need for dynamic 3D FACS data sets comparable to the state of the art in 2D. However, given such a corpus, approaches are also required for the modeling and utilization of this data. A popular model for 3D facial analysis is the morphable model [3]. This uses a basis of static 3D laser range scans of different subjects to learn a statistical space of shape and texture deformation. However, in order to build such a model the scans must

first be non-rigidly registered to a common space. This process is required to achieve 3D mesh correspondence. While Blanz and Vetter [3] rely solely on optical flow to densely register images, Patel and Smith [20] improve accuracy by employing a set of manually labeled facial feature points.

Even though 3D morphable models are potentially powerful tools for facial analysis, previous work to date has only used static 3D scans of faces to build models. There is therefore great potential for extending the framework to incorporate dynamic data. However, the problem with building such models lies again in non-rigid registration. In the context of dynamic 3D data, this requires the creation of spatio-temporal dense feature correspondences throughout the sequences. The problem is more complex than using static scans alone since registration must reliably track highly variable nonlinear skin deformations [13].

One approach for achieving correspondence given dynamic 3D sequences is to register facial images using optical flow vectors tracked dynamically in multiple 2D stereo views [7, 24]. However, drift in the flow (caused by e.g. violation of the brightness consistency assumption) typically accumulates over time introducing errors. Borshukov et al [6] overcome this problem by manually correcting the mesh positions when drift occurs. More recently, Bradley et al [7] mosaiced the views of 14 HD cameras to create high resolution images for skin pore tracking. By back calculating optical flow to the initial image drift is also reduced. In addition, mesh regularization ensures that faces do not flip due to vertices overlapping. Other solutions to the registration problem include the use of facial markers and special make-up to track consistent points [18].

Existing non-rigid registration methods for dynamic facial data therefore have drawbacks: they rely on optical flow which is prone to drift over time, or use painted facial markers to acquire stable points. There is therefore clear scope for improvement. Previous work on these methods has also only been applied to animation, where errors can be hand corrected. A more quantitative assessment of their merits would therefore also be of benefit to the computer vision community. Finally, given a reliable means to non-rigidly register 3D facial data efficiently, the opportunity for building dynamic 3D morphable models becomes possible.

1.1. Contributions

This paper makes several contributions: It presents the first dynamic 3D FACS data set for facial expression research, portraying 10 subjects performing between 19 and 97 different AUs both individually and in combination. In total the data set contains 519 AU sequences. Compared with other state of the art 2D [17] and 3D [22] facial data sets which contain more subjects, we provide substantially more expressions per subject. As well as allowing comprehensive experimentation on per person facial movement, the data allows for thorough research in a range of tasks: large scale 3D model building, registration of 3D faces, and tracking of 3D models to 2D video.

The peak frame of each sequence has been manually FACS coded by certified FACS experts. These are individuals whom have passed the FACS final test [12]. This provides the first ground truth for 3D FACS based AU recognition systems, as well as a valuable resource for building 3D dynamic morphable models for motion capture and synthesis using AU based parameters.

Secondly, our paper provides a description of the first framework for building dynamic 3D morphable facial models. This extends the state of the art in static 3D morphable model construction to incorporating dynamic data. In describing this framework, we also propose an Active Appearance Model (AAM) [9] based approach for densely and reliably registering captured 3D surface data. This method has several advantages over existing dynamic 3D facial registration methods: (1) it requires no paint or special markers on the face, and (2) it shows improved performance over optical flow based strategies which accumulate drift over time [7, 6, 24]. We compare the AAM based method to the state of the art in optical flow and mesh regularization schemes. This provides the first quantitative assessment of popular methods adopted in this area. We also include a comparison to techniques used in static 3D morphable model construction [2], and highlight limitations in directly applying these approaches given dynamic data.

2. Dynamic 3D FACS Dataset (D3DFACS)

2.1. Capture Protocol and Contents Overview

Our aim was to capture a variety of facial movements as performed by a range of posers. For this, we recruited 4 expert FACS coders and 6 FACS-untrained participants for our data set. The performer age range was 23 to 41 years (average age 29.3 years), and consisted of 6 females and 4 males, all of Caucasian European origin. The expert coders, having extensive knowledge of FACS, allowed us to elicit more complex AU combinations than would be possible for FACS unfamiliar people. Each FACS expert spent time before the session practicing the combinations as well as possible. The FACS unfamiliar participants were provided coaching before the session on a reduced set of AUs and expressions. For a discussion on how easily people find performing different AUs, the reader is referred to [15].

In total we recorded between 80 and 97 AU sequences (including Action Descriptors (ADs) [12]) for each FACS expert performer, and between 19 and 38 sequences for each FACS non-expert. This number depended on the ability of a performer to produce the desired sequence, which either targeted a specific single AU, or a combination of AUs. We selected the combinations based on criteria for (1) the six

basic emotions outlined by Ekman et al [12], and (2) non-additive appearance changes. These latter combinations are particularly interesting since they reveal new appearance characteristics for their joint activation that cannot be traced back to the sum of single AUs (e.g. 1+4). In total 519 sequences were captured, comprising of 1184 AUs in total. Table 1 shows the frequency of each AU in the data set.

2.2. Dynamic 3D Capture and Data Format

Each FACS performer was recorded using a 3DMD dynamic 3D stereo camera [21] (see Figure 1). The system consists of six cameras split between two pods, with 3 cameras stacked vertically on each pod. Each pod produces a 3D reconstruction of one half of the face. The top and bottom cameras of each pod are responsible for stereo reconstruction and middle cameras are responsible for capturing UV color texture. The system samples at 60 FPS and provides (1) OBJ format 3D mesh data consisting of the two pod half-face meshes joined together, and (2) corresponding BMP format UV color texture map data for each frame. The texture mapping provided by the system is originally a stereo one, meaning that it consists of the color camera views from the two pods joined together into one image. We modify this by converting the mapping into a cylindrical one. This means that each mesh has a UV texture map equivalent to placing a cylinder around the head and projecting the color information on the cylinder. The mesh data consists of approximately 30K vertices per mesh, and each UV map is 1024x1280 pixels. Figure 2 shows example images of the FACS performers, including corresponding mesh and UV map data.

Figure 1. Dynamic 3D Stereo Camera used for data collection. Six cameras combine to provide 3D reconstructions of the face, with a recording rate of 60 FPS.

For each sequence the camera was set to record for between 5 and 10 seconds depending on the complexity of the AU. Performers were asked to repeat AU targets as many times as possible during this period. A mirror was set up in front of the actor so that they could monitor their own expressions before and during each capture. Recording took between 2 and 7 hours per participant. After all data recording, the sequences which most visually matched the targets from onset to peak were extracted for scoring by a FACS expert. This led to the following data set:

- 519 AU sequences (single and in combination) from 10 people, including 4 expert coders and 6 non-experts.
- Each sequence is approximately 90 frames long at 60 FPS and consists of OBJ mesh and BMP cylindrical UV texture map data.
- AU codes for each peak frame of each sequence are scored by a FACS expert.

Instructions for acquiring the database may be found at http://www.cs.bath.ac.uk/~dpc/D3DFACS/. In the remainder of the paper we describe our framework for building dynamic 3D morphable facial models. We also introduce our AAM based approach for mesh registration in dynamic sequences and compare it to: (1) existing work on facial mesh correspondence and (2) registration techniques employed in static 3D data for morphable modeling [20, 2].

3. 3D Dynamic Morphable Model Framework

In the following Section, we first provide an overview of static 3D morphable model construction before describing extensions to dynamic sequences. In static 3D morphable model construction, as proposed by Blanz and Vetter [2], a set of 200 facial scans (each of a different person) is taken from a Cyberware 2020PS laser range scanner. These are represented in a 2D space and aligned to a common coordinate frame using a dense optical flow alignment. Patel and Smith [20] improve the accuracy of the alignment by manually placing 2D landmarks on the faces, and then using a Thin Plate Spline (TPS) based warping scheme [4]. Procrustes analysis is also performed to remove head pose variation. After correspondence, the 2D UV space which also contains a mapping to 3D shape is sampled to generate the 3D mesh information. Both the UV texture data and the 3D mesh data are then represented using linear Principle Component Analysis (PCA) models.

We propose several extensions to this process for building dynamic 3D morphable models. Given several single dynamic sequences (e.g. an AU combination) consisting of multiple 3D meshes and corresponding UV texture maps:

Step 1: For each mesh, generate a mapping from each pixel in 2D UV space to a vertex position in 3D space. This mapping is $I(\mathbf{u}) = \mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^3$ is a 3D vector coordinate, I is a UV map, and $\mathbf{u}$ is a coordinate (u, v). The function can be generated using a Barycentric coordinate mapping between mesh faces in 2D UV space and faces in 3D vertex space (see Section 4.1).

Step 2: Perform stand-alone non-rigid registration of each separate UV texture map sequence. This process identifies and tracks image features through neighboring image

AU	Description	Total	AU	Description	Total	AU	Description	Total
1	Inner Brow Raiser	45	17	Chin Raiser	118	31	Jaw Clencher	4
2	Outer Brow Raiser	36	18	Lip Pucker	26	32	Lip Bite	5
4	Brow Lowerer	56	19	Tongue Out	3	33	Cheek Blow	4
5	Upper Lid Raiser	42	20	Lip Stretcher	30	34	Cheek Puff	3
6	Cheek Raiser	16	21	Neck Tightener	6	35	Cheek Suck	3
7	Lid Tightener	38	22	Lip Funneler	15	36	Tongue Bulge	4
9	Nose Wrinkler	36	23	Lip Tightener	47	37	Lip Wipe	3
10	Upper Lip Raiser	97	24	Lip Pressor	22	38	Nostril Dilator	29
11	Nasolabial Deepener	16	25	Lips Part	164	39	Nostril Compressor	9
12	Lip Corner Puller	77	26	Jaw Drop	63	43	Eyes Closed	13
13	Cheek Puffer	5	27	Mouth Stretch	14	61	Eyes Turn Left	4
14	Dimpler	32	28	Lip Suck	8	62	Eyes Turn Right	4
15	Lip Corner Depressor	28	29	Jaw Thrust	4	63	Eyes Turn Up	4
16	Lower Lip Depressor	42	30	Jaw Sideways	5	64	Eyes Turn Down	4

Table 1. AU frequencies identified by manual FACS coders in the D3DFACS data set (based on FACS descriptions in Ekman et al [12]).

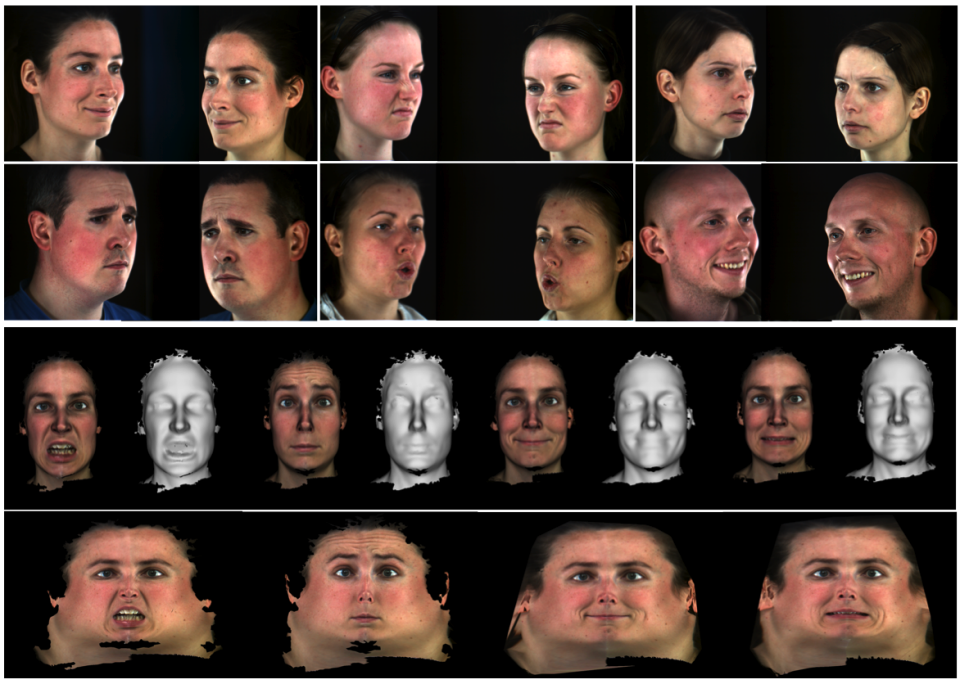

Figure 2. Examples from the D3DFACS data set. The top two rows show camera views from 6 participants. The bottom two rows show 3D mesh data (textured and un-textured), and corresponding UV texture maps.

sequence frames. In this paper we propose a dense AAM based approach to achieve registration and compare to state of the art approaches. Directly applying optical flow for registration without a mesh regularization term (as in [2]) produces drift artifacts in the meshes and images. Even with a regularization term (as in [7, 24]) tracking accuracy still depends on optical flow quality which can be error prone (see Section 4.2). Registration is with respect to a neutral expression image selected from the sequence.

Step 3: Perform global non-rigid registration of the dynamic sequences. One of the neutral sequence poses is chosen as a global template to which each of the UV sequences is then registered using a single dense warping per sequence. This registered UV space provides data for the linear texture PCA model (see Section 4.2)

Step 4: Regularly sample the UV space to calculate 3D vertices for each corresponding mesh. The more accurate the pixel based registration is, the more accurate the mesh correspondence (see Section 4.2).

Step 5: Perform rigid registration of the 3D mesh data. Since sequences at this point have 3D mesh correspondence, Procrustes analysis [5] may be applied to align the meshes in an efficient manner. This removes head pose variation in the dynamic sequences.

Step 6: Build linear PCA models for shape and texture using the registered 3D mesh and UV texture data.

We now expand on the above process concentrating primarily on the procedures for non-rigid registration.

4. 3D Registration and Correspondence

4.1. Creating a 2D to 3D Mapping (Step 1)

A sequence of data consists of a set of meshes $\mathcal{X} = [\mathbf{X}_1, \dots \mathbf{X}_n]$, where $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T \dots \mathbf{x}_m^T]^T, \mathbf{x}_i = [x_x^i, x_y^i, x_z^i]^T \in \mathbb{R}^3$. There also exists a set of UV texture maps $\mathbb{I} = [I_1 \dots I_n]$, and a set of UV coordinates $\mathbb{U} = [\mathbf{U}_1, \mathbf{U}_2, \dots \mathbf{U}_n]$, where $\mathbf{U} = [\mathbf{u}_1^T, \mathbf{u}_2^T, \mathbf{u}_m^T]^T$, and $\mathbf{u}_i = [u_i, v_i] \in \mathbb{R}^2$. The UV texture maps supply color data to the mesh in the form of images, with the UV coordinates linking individual vertices $\mathbf{x}_i$ to unique points on these images $\mathbf{u}_i$. In the above definitions, n is the number of meshes and corresponding UV maps in a sequence. Similarly, m is the number of vertices in a mesh and the number of corresponding UV coordinates.

There also exists a set of common triangular faces per mesh $\mathbf{F}_i, i = 1 \dots n$, where faces in the 3D vertex space correspond to the same faces in the 2D texture space. The entire set of faces for a sequence may also be defined as $\mathbb{F} = [\mathbf{F}_1, \dots \mathbf{F}_n]$.

We approach 3D correspondence as a 2D image registration problem. From a theoretical point of view, perfect one-to-one pixel registration between successive face images relates to perfect 3D mesh correspondence. The goal is to achieve as near an optimal correspondence as possible.

It is therefore useful from an implementation point of view to work primarily in 2D space. We first generate 3D images $I_{3D}(\mathbf{u}) = \mathbf{x}$. This is achieved by taking each face in turn, and for each pixel within its triangle in 2D space calculating the corresponding 3D position using a Barycentric coordinate mapping. Repeating for each triangle results in a dense 2D pixel to 3D vertex mapping for the entire UV map. Operations performed on I are from now on also applied to I_{3D}, including optical flow and TPS warping.

4.2. Non-Rigid Alignment (Steps 2 and 3)

We now describe several strategies for non-rigid alignment, including our proposed method. In Section 5 we then provide experimental results comparing these.

Optical Flow: Blanz and Vetter [2] calculate smoothed optical flow to find corresponding features between images of 200 different people. However, the formulation and choice of features is tuned to the particular data. In this work we consider a more standardized approach and extend to dynamic sequences. We calculate concatenated Lukas-Kanade (LK) [16] flow fields that warp images between $I+i$ and I_0, where I_0 is the neutral expression image (UV map). Flow is summed for the images between $I + i$ and I_0, providing the concatenated flow. Smoothing of the flow field is applied in the form of local averaging in both the spatial and temporal domains. Flow fields calculated from the UV maps are also then applied to the I_{3D} images.

Optical Flow and Regularization: Bradley et al [7], Zhang et al [24] and Borshukov et al [6] use optical flow from stereo image pairs to update a mesh through a sequence. They use this technique for animation applications. The mesh is initialized in frame 1, and its vertices moved to optimal positions in successive frames using flow vectors merged from each stereo view. The update is also combined with a mesh regularization constraint to avoid flipped faces. We extend this approach by using a single UV space for optical flow calculation and mesh updating as opposed to merging stereo flow fields. For regularization, we sparsely sample the flow field and interpolate the positions of in-between points using TPS warping (see *AAM and TPS* next). This ensures that flow vectors follow the behavior of the sparse control points, but as with previous approaches does not guarantee against tracking errors accumulating due to optical flow drift.

AAM and TPS: Patel and Smith [20] achieve correspondence in 3D morphable model construction by manually landmarking 3D images and aligning them using TPS based warping. TPS is a type of Radial Basis Function (RBF). RBFs can be used to define a mapping between any point defined with respect to a set of basis control points (e.g. landmarks in one image), and its new position given a change in the control points (e.g. landmarks in the target image). Thus, a dense mapping between pixels in two images may be defined. TPS itself provides a kernel that models this mapping based on a physical bending energy term (for more detail see [4]).

This approach has several advantages over optical flow based correspondence: (1) the TPS warp provides a smooth and dense warping field with no drift artifacts in the mesh or images, and (2) manual point placement guarantees correspondence of key facial features whereas optical flow is prone to drift. We extend this to dynamic sequences by building an AAM and using it to automatically track feature points through a dynamic sequence of multiple UV maps. Pixels (or (u, v) coordinates) within the control points are corresponded with those in neighboring frames using the TPS mapping, which warps each I and I_{3D} to a common coordinate frame (the neutral expression).

AAMs are well known in the computer vision literature, and a thorough overview may be found in [9]. We use the same principles here and define our AAM as:

$$\mathbf{l} = \bar{\mathbf{l}} + \mathbf{P}_l \mathbf{W} \mathbf{Q}_l \mathbf{c} \qquad \mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c} \tag{1}$$

where $\mathbf{l}$ is a vector of image landmarks, $\bar{\mathbf{l}}$ are the mean landmarks learned from training, $\mathbf{g}$ is a vector of image pixels inside the region defined by $\mathbf{l}$, and $\bar{\mathbf{g}}$ are the mean image pixels learned from training. The eigenvectors of the training sets of vectors $\mathbf{l}$ and $\mathbf{g}$ are the matrices $\mathbf{P}_l$ and $\mathbf{P}_g$ respectively. The matrix $\mathbf{W}$ is a set of scaling weights, the matrix $\mathbf{Q}$ represents the eigenvectors of the joint distribution of landmark and image data, and $\mathbf{c}$ is the appearance parameter.

Fitting AAMs to new images is a well covered topic in the computer vision literature (see [9, 17]). In this work we define a simple minimization approach which compares the current guess to the models best reconstruction of this:

$$E = \min_{c}(\mathbf{g}_l - (\mathbf{P}^T(\mathbf{g}_l - \bar{\mathbf{g}})) \quad (2)$$

where $\mathbf{g}_l$ is portion of the image I within the area defined by $\mathbf{l}$ (the current guess). Calculating $\mathbf{g}_l$ requires first calculating $\mathbf{l}$ using $\mathbf{c}$ (in (1)), and then warping this region into the space defined by the mean landmarks $\bar{\mathbf{l}}$. In order to optimize E we use the Levenberg-Marquardt algorithm. The process of tracking results in a set of labeled feature based landmarks per frame (neutral expression). These can be then used to warp each image to the common coordinate frame, thus achieving dense non-rigid correspondence.

4.3. Sampling, Rigid Alignment and Statistical Modeling (Steps 4, 5 and 6)

Given a set of non-rigidly aligned sequences, these are aligned again to a single common coordinate frame. This is selected to be a neutral expression from the full training sequence. The space of aligned images I_{3D} is then uniformly sampled. This sampling defines the topology and density of the facial mesh, recalling that $I_{3D}(\mathbf{u}) = \mathbf{x}$. Since each I_{3D} refers to a different set of 3D points, aligning these and then sampling in a uniform manner results in a unique set of registered 3D meshes. Similarly, there now also exists a common set of faces $\mathbf{F}$ for each mesh.

The entire set of 3D mesh data can now be rigidly aligned using Procrustes analysis (see [5] for a detailed description). Following [2] the registered 3D mesh $\mathbb{X}$ and UV texture data $\mathbb{I}$ may now be expressed using two PCA models:

$$\mathbf{X}' = \bar{\mathbf{X}} + \mathbf{P}_X\mathbf{b}_X \qquad \mathbf{I}' = \bar{\mathbf{I}} + \mathbf{P}_I\mathbf{b}_I \quad (3)$$

where $\bar{\mathbf{X}}$ is the mean mesh, $\bar{\mathbf{I}}$ is the mean UV image texture, $\mathbf{P}_X$ and $\mathbf{P}_I$ are the eigenvectors of $\mathbb{X}$ and $\mathbb{I}$, and $\mathbf{b}_x$ and $\mathbf{b}_I$ are vectors of weights. The eigenvectors of $\mathbb{X}$ and $\mathbb{I}$ are ordered by the proportion of total variance in the data they represent. Removing some of their columns therefore means that the projected weights $\mathbf{b}_x$ and $\mathbf{b}_I$ can be made much smaller than $\mathbf{x}$ and $\mathbf{I}$. Rewriting (3) allows us to perform this parameterization to a lower dimensional space:

$$\mathbf{b}_X = \mathbf{P}_X^T(\mathbf{X}' - \bar{\mathbf{X}}) \qquad \mathbf{b}_I = \mathbf{P}_I^T(\mathbf{I}' - \bar{\mathbf{I}}) \quad (4)$$

This provides a convenient lower dimensional representation for storing dynamic facial movements and performing optimization when fitting the model to new sequences.

5. Experiments

In this Section we perform baseline experiments comparing the three registration approaches described in section 4.2. These are (1) standard optical flow concatenation which extends [2], (2) a combined optical flow and regularization approach similar to [7, 24], and (3) the new AAM-TPS combination approach proposed in this paper.

For test purposes we selected 8 dynamic AU sequences from our data set consisting of approximately 65 frames each. For optical flow we use the pyramidal Lucas-Kanade (LK) algorithm as in [7]. We first wished to compare how well the AAM and LK algorithms tracked facial feature points versus a ground truth. To create the ground truth we manually annotated each frame from each sequence with landmark points at 47 key facial features around the eyes, nose and mouth. This test would give an indication of how stable points are over time, and whether drift occurs as reported in previous work. For the AAM test, an individual model with 47 landmarks was trained for each sequence using 3 manually selected frames – typically at the beginning, middle and end. Points were manually initialized in frame 1 for both the AAM and LK tests. Table 2 shows the mean Euclidian error between ground truth points and tracked points (in pixels) for each frame. It can be seen that the AAM error is consistently lower than the LK error. Figure 3 shows examples of how the LK error accumulates over the course of tracking, supporting the optical flow drift observations in [6, 7, 24]. This is evidence that the AAM method provides a more stable tracking approach over time, and is a valuable tool for reducing drift.

We next wished to evaluate how well each method performed registration of the image sequences from a qualitative point of view. Figure 4 shows example registrations of peak frames to neutral frames for four sequences using (1) dense concatenated LK flow fields between the peak and neutral frame (see Section 4.2 - Optical Flow), (2) concatenated LK optical flow combined with TPS regularization (see Section 4.2 - Optical Flow and Regularization), and (3) feature points tracked with an AAM and registered using TPS (see Section 4.2 - AAM and TPS).

It can be seen from Figure 4 that the LK method used alone produces noticeable drift artifacts. We observed that this is due to pixels overlapping each other, and is a result of the flow field being concatenated over consecutive neighboring frames. One approach to avoid this in the future may be to add a temporal constraint to the flow calculation which observes learned facial deformations. The LK+TPS method overcomes the drawback of pixel overlap due to (1) tracked points being initially far apart, and (2) the TPS method regularizing the positions of pixels in between the tracked points. Alignment is much improved over LK alone. However, as highlighted by the red dotted circles, accumulated optical flow drift causes some facial features (such as the lower lip and cheeks) to distort. The AAM-TPS method provides the most stable registration, as demonstrated qualitatively by an absence of drift artifacts and pixel overlaps. We have also used this technique in a

perceptual face experiment [10] and participants reported no visible issues with the model.

Finally, we used the AAM-TPS approach to create a morphable model (see Section 4.3). We parameterized the original sequences using (4) and then re-synthesized them using (3). Figure 5 shows example outputs from the model. In order to show how the mesh deforms smoothly with the tracked facial features we also show corresponding examples using a UV map of a checkered pattern. The deformations in the pattern clearly demonstrate that the mesh is following the correct facial movement.

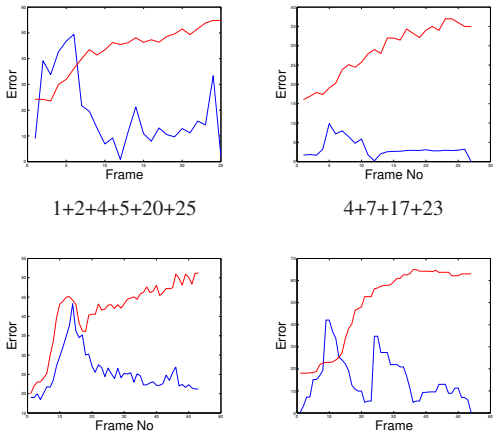

Figure 3. AAM-TPS (blue line) and Lucas Kanade (red line) tracking errors for 4 sequences. It can be seen that the optical flow method accumulates error as the sequence moves on, whereas the AAM value remains consistently lower.

AU Sequence	AAM-TPS	LK
1+2+4+5+20+25	18.6	43.1
20+23+25	53.8	63.8
9+10+25	25.4	41.7
18+25	15.3	43.6
16+10+25	23.2	38.8
12+10	15.2	48.7
4+7+17+23	3.4	28.5
1+4+15	2.3	26.2

Table 2. AAM-TPS and Lucas Kanade mean Euclidian error values (in pixels) for tracked feature points versus ground truth landmark points. 8 dynamic AU sequences were tracked in this particular test. The result demonstrates the improved reliability of the AAM tracking method over the optical flow approach.

6. Conclusion and Future Work

In this paper we have presented the first dynamic 3D FACS data set (D3DFACS) for facial expression research. The corpus is fully FACS coded and contains 10 participants performing a total of 534 AU sequences. We also proposed a framework for building dynamic 3D morphable facial models and described an AAM based approach for non-rigid 3D mesh registration. Our experiments show that the approach has several advantages over optical flow based registration. For future work we wish to perform experiments comparing the performance of dynamic morphable models versus static ones in a series of benchmark tests such as tracking. We would also like to combine model based approaches such as AAMs with optical flow to improve dense feature point registration between the tracked feature points.

Acknowledgements

We would like to thank the Royal Academy of Engineering/EPSRC for funding this work. Also thanks to all the participants in the data set, particularly the FACS experts: Gwenda Simons, Kornelia Gentsch and Michaela Rohr.

References

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition and facial actions in spontaneous behavior. *Journal of Multimedia*, 2006. 1

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. of ACM Siggraph*, 1999. 1, 2, 3, 4, 5, 6

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25:1063–1074, 2003. 1, 2

[4] F. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11:567–585, 1989. 3, 5

[5] F. Bookstein. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge Uni. Press, 1991. 4, 6

[6] G. Borshukov, D. Piponi, O. Larsen, J. Lewis, and C. Tempelaar-Lietz. Universal capture - image based facial animation for the matrix reloaded. In *ACM SIGGRAPH Sketch*, 2003. 2, 5, 6

[7] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer. High resolution passive facial performance capture. *ACM Trans. Graph.*, 29:1–10, 2010. 2, 4, 5, 6

[8] F. R. G. Challenge. http://www.nist.gov/itl/iad/ig/frgc.cfm. 1

[9] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:681–685, 2001. 2, 5, 6

[10] D. Cosker, E. Krumhuber, and A. Hilton. Perception of linear and nonlinear motion properties using a facs validated 3d facial model. In *In Proc. of ACM Applied Perception in Graphics and Visualisation*, pages 101–108, 2010. 7

[11] J. Duncan. The unusual birth of benjamin button. *Cinefex*, 2009. 1

[12] P. Ekman, W. Friesen, and J. Hager. *Facial Action Coding System: Second Edition*. Salt Lake City: Research Nexus eBook, 2002. 1, 2, 3, 4

[13] Y. Furukawa and J. Ponce. Dense 3d motion capture for human faces. In *In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1674–1681, 2009. 1, 2

[14] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001. 1

[15] P. Gosselin, M. Perron, and M. Beaupre. The voluntary control of facial action units in adults. *Emotion*, 10(2):266–271, 2010. 2

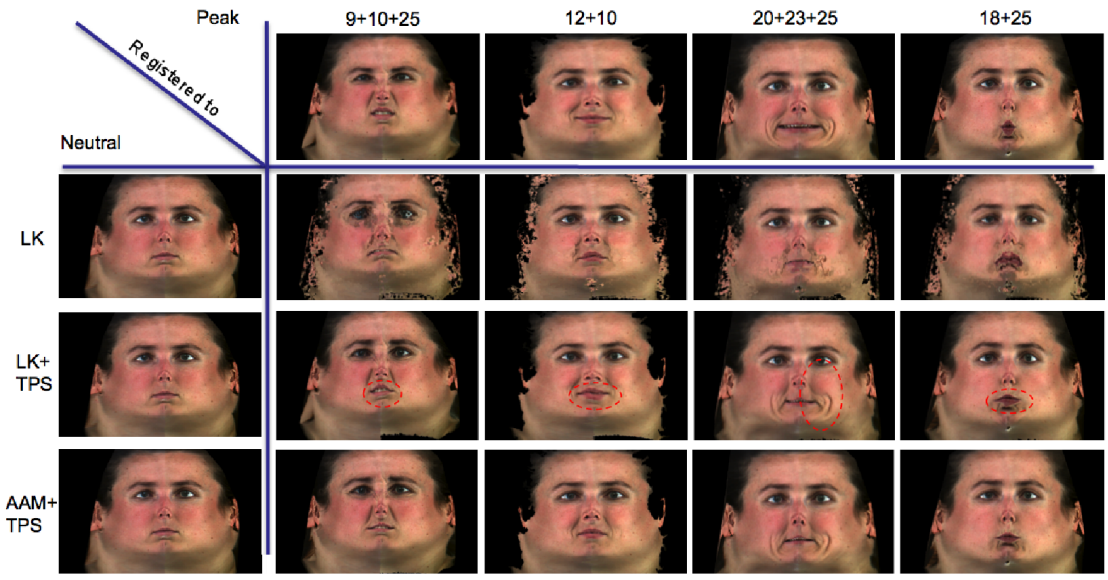

Figure 4. Peak frames registered to neutral frames using LK, LK+TPS and AAM+TPS (see Section 4.2). In each case the concatentated sequence information between the peak and neutral frame is used for registration. Red circles highlight drift errors in the LK+TPS approach.

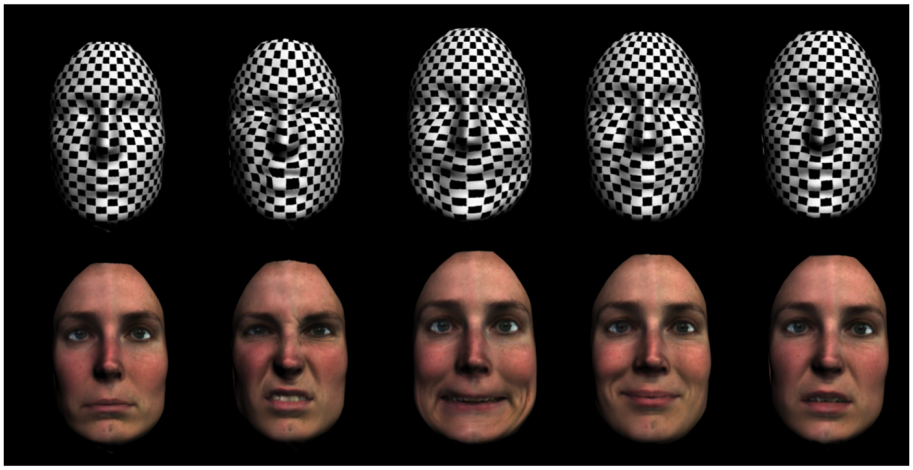

Figure 5. Outputs from a morphable model constructed using the AAM+TPS method: (left to right) Neutral, 9+10+25, 20+23+25, 12+10 and 16+10+25. The checker pattern highlights the underlying mesh deformation.

[16] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *In Proc. of Image Understanding Workshop*, pages 121–130, 1981. 5

[17] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 94–101, 2010. 1, 2, 6

[18] W. Ma, A. Jones, J. Chiang, T. Hawkins, S. Frederiksen, P. Peers, M. Vukovic, M. Ouhyoung, and P. Debevec. Facial performance synthesis using deformation-driven polynomial displacement maps. *ACM Tran. Graph.*, 27(5):1–10, 2008. 2

[19] M. Pantic, M. Valstar, R. Rademaker, and L. Matt. Fully automatic facial recognition in spontaneous behavior. In *In. Proc of International Conference on Multimedia and Expo*, pages 317–321, 2005. 1

[20] A. Patel and W. Smith. 3d morphable face models revisited. In *In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1334, 2009. 1, 2, 3, 5

[21] D. Systems. http://www.3dmd.com. 3

[22] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *In Proc. of Int. Conf. on Auto. Face and Gesture Recog.*, 2008. 1, 2

[23] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *In Proc. of Int. Conf. on Auto. Face and Gesture Recog.*, 2006. 1

[24] L. Zhang, N. Snavely, B. Curless, and S. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graph.*, 23(3):548–558, 2004. 2, 4, 5, 6