MimicME: A Large Scale Diverse 4D Database for Facial Expression Analysis

Athanasios Papaioannou[1,*], Baris Gecer[*], Shiyang Cheng[*], Grigorios Chrysos[*], Jiankang Deng[*], Eftychia Fotiadou[*], Christos Kampouris[*], Dimitrios Kollias[*], Stylianos Moschoglou[*], Kritaphat Songsri-In[*], Stylianos Ploumpis[*], George Trigeorgis[*], Panagiotis Tzirakis[*], Evangelos Ververas[*], Yuxiang Zhou[*], Allan Ponniah, Anastasios Roussos[2,3], and Stefanos Zafeiriou[4]

[1] apapaion@gmail.com
[2] University of Exeter, UK
[3] Institute of Computer Science, Foundation for Research and Technology Hellas
troussos@ics.forth.gr
[4] Imperial College London s.zafeiriou@imperial.ac.uk

Abstract. Recently, Deep Neural Networks (DNNs) have been shown to outperform traditional methods in many disciplines such as computer vision, speech recognition and natural language processing. A prerequisite for the successful application of DNNs is the big number of data. Even though various facial datasets exist for the case of 2D images, there is a remarkable absence of datasets when we have to deal with 3D faces. The available facial datasets are limited either in terms of expressions or in the number of subjects. This lack of large datasets hinders the exploitation of the great advances that DNNs can provide. In this paper, we overcome these limitations by introducing MimicMe, a novel large-scale database of dynamic high-resolution 3D faces. MimicMe contains recordings of $4,700$ subjects with a great diversity on age, gender and ethnicity. The recordings are in the form of 4D videos of subjects displaying a multitude of facial behaviours, resulting to over $280,000$ 3D meshes in total. We have also manually annotated a big portion of these meshes with 3D facial landmarks and they have been categorized in the corresponding expressions. We have also built very powerful blendshapes for parameterising facial behaviour. MimicMe will be made publicly available upon publication and we envision that it will be extremely valuable to researchers working in many problems of face modelling and analysis, including 3D/4D face and facial expression recognition[†]. We conduct several experiments and demonstrate the usefulness of the database for various applications.

1 Introduction

Arguably, 3D Morphable Models (3DMMs) have dominated the field of 3D statistical shape modelling the last 20 years since their introduction by the seminal

[*] Authors were with Imperial College London during this work.
[†] https://github.com/apapaion/mimicme

Fig. 1: Synthetic 3D faces with random identity, appearance and expression generated by the non-linear model that we built from the proposed MimicMe database. The collected large-scale 4D data (4,700 identities and over 280,000 high-resolution 3D meshes) alongside the proposed especially-designed processing framework yield models and results of unprecedented realism and quality.

work of Blanz and Vetter [5]. They have been used in a variety of applications and in different fields such as creative media, medical image analysis, biometrics, computer vision, human behavioral analysis, computer graphics, craniofacial surgery and large-scale facial phenotyping, see e.g. [2,18,44,30,58,19,39,38].

In essence, a face 3DMM is constructed by bringing all the facial 3D meshes of the training set into dense correspondence and then performing some form of dimensionality reduction, typically principal component analysis (PCA). When a set of 3D meshes is in dense correspondence, then vertices with the same index in different meshes represent the same part (e.g. a vertex with index i represents the nose tip in every mesh of the set). After the construction of a 3DMM, a new face shape is represented by a few parameters.

During the last years, the rise of Deep Neural Networks (DNNs), with Convolution Neural Networks (CNNs) as the catalyst, revolutionized computer vision. It was only natural that 3D shape modelling was among the fields that were significantly influenced by this trend. Two main lines of research have emerged. The first one has attempted to re-model the 3DMMs by introducing some form of non-

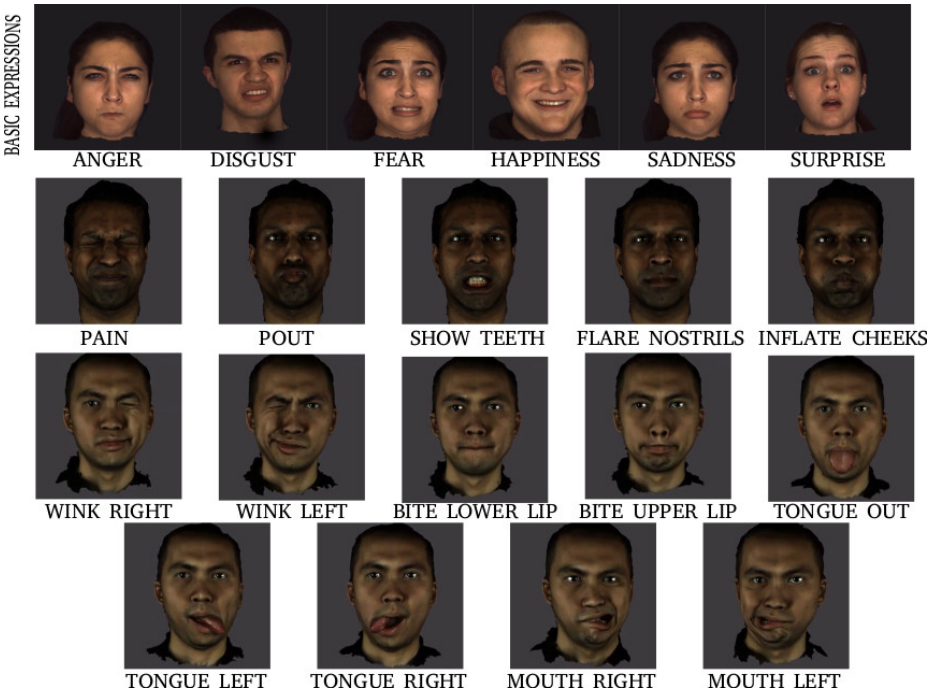

Fig. 2: Expressions included in the dataset. In the first row, the basic expressions are shown. These snapshots are taken from the video that subjects were watching in order to mimic the corresponding facial expressions.

linearity replacing the PCA part of 3DMMs with autoencoders [41,10,26,13]. The other one has attempted to leverage the success of DNNs in 2D images by applying these models in 2D representations of the 3D images like UV maps [35,23,22].

A common ground for both approaches is the need for large datasets. Even though this is trivial for the case of 2D images and videos, where numerous databases have been proposed for various applications, such as 2D-based face identification, in-the-wild face analysis, age estimation, facial emotion recognition, etc, the relevant advances in the field of 3D modelling are still rather limited. This is due to the fact that 3D acquisition devices are usually expensive, can be found only in specialised products or are used for medical purposes (i.e. CT, MRI). The majority of previous 3DMMs have been built using either neutral faces, or a small sample size of people under various expressions. Thus, the community is still lacking a database that combines large numbers in subjects, expressions and 3D images per subject.

In this work, we overcome the aforementioned limitations by constructing a large scale dynamic facial expression database, hereby coined as MimicMe. Our dataset includes $4,700$ subjects performing various facial expressions. It contains over $280,000$ high resolution 3D facial scans. Due to the demographic richness of

our dataset, we expect that MimicMe will become an invaluable asset for many different problems such as 3D/4D face and facial expression recognition, building high-quality expressive blendshapes, as well as synthesizing 3D faces for training deep learning systems. In addition, we revisit LSFM [9], a fully automated and robust Morphable Model construction pipeline, to incorporate the advances in landmarks' localization, and extend it to handle not only neutral faces but also faces with various expressions. As the number of 3D scans is huge and contains particularly large demographic variability (the capture process was performed in a museum with people from various countries), there were some flawed 3d registrations which we proposed to amend them by exploiting a StyleGan[29].

In summary, the contributions of this paper are the following:

- We introduce MimicMe, a database of 4,700 subjects collected over a period of three months with over 280,000 3D facial meshes, with various posed facial behaviours.
- We have a number of $55,000$ 3D facial landmarks manually corrected, which leads to a subspace of sparse representations for 3D facial expressions.
- We revisit LSFM pipeline for registration to take advantage of DNNs' for landmarks' localization and generalize for all facial poses.
- We propose a new method to correct distorted 3D scans by training a StyleGan to perform texture completion.
- We build novel expression blendshapes learned from our database that are more powerful than the off-the-shelf blendshapes provided by other datasets.
- We show that our dataset can be used to generate high-quality faces by capitalizing on the recent developments on Generative Adversarial Networks (GANs). We build a novel non-linear model that can synthesize shape, expression, texture and normals.

2 Related Work

2.1 3D/4D Face datasets

As the existence of a dataset of 3D faces is a crucial factor to build a 3D Morphable Model and an expression blendshape model, there have been attempts in the past to build as large and diverse datasets as possible. Even though there are datasets that used 3DMM fitting process to reconstruct 3D faces from images [57,27], a process called analysis-by-synthesis [19], these datasets are usually of limited quality. Thus, we focus on datasets that have been created using depth sensors, scanners or multi-view camera systems. These 4D face datasets can be categorized according to the kind of facial movements, namely, datasets that include a variety of expressions (e.g. happiness, sadness, disguise), and datasets that focus on speech.

One of the first 3D datasets with expressions is BU-3DFE database [53], which includes articulated facial expressions from 100 adults. The age of subjects ranges from 18 years to 70 years old and included the 6 prototypic expressions (happiness, disgust, fear, angry, surprise and sadness). Bosphorus database [42]

includes similar numbers of 105 individuals with a rich set of 34 expressions per subject, but it does not contain 3D meshes, as the data are in the form of depth maps and texture images. BU-4DFE [52] was an extension of BU-3DFE to dynamic 3D space using the same 6 facial expressions from 101 adults. FaceWarehouse [12] is one more database which has been used used to build 3D blendshapes. To this end, it includes both neutral 3D faces and 3D facial expressions. The capture process was implemented with a Kinect (RGBD camera) and the dataset size is of 150 subjects, aged 7-80 from various ethnic backgrounds. Expressions include the neutral expression and 19 others such as mouth-opening, smile, kiss, etc. 4DFAB [14] has been proposed recently and is one of the largest 4D datasets. It consists of over 1,800,000 3D meshes. 4DFAB contains recordings of 180 subjects captured in four different sessions spanned over a five-year period. Subjects performed not only the 6 prototypic expressions, but also spontaneous expressions and 9 words utterances. More recently, Ranjan et al. [41] propose the framework of convolutional mesh autoencoders (COMA) and introduce for the needs of their method a dataset of 12 different subjects with $20,466$ meshes of extreme expressions was also provided.

One other trend of 4D databases is to acquire databases which focus on speech and word utterance. In this category belongs VOCASET [17] which contains a collection of audio-4D scan pairs captured for 12 subjects. For each subject, 40 sequences of a sentence spoken in English, each of length three to five seconds have been collected. 4D Cardiff Conversation Database (4D CCDb) [34] contains 4 subjects captured while discussing topics of their own interest. In total 34 conversations, have been captured at a frame rate of 60 fps leading to 3500-4000 frames per sequence. [47] captured 4D sequences of 2 native and 2 non-native English speakers reading out the 500 words contained in the publicly-available Lipreading Words (LRW) in-the-wild dataset [15]. S3DFM [54] is a publicly-available dataset that focuses on speech-driven 3D facial dynamics across 77 subjects. Each subject read out 10 times a word and the whole process was captured with a high frame rate 3D video sensor (500 fps).

Table 1 provides a summary of the most recent 3D and 4D datasets, with details such as the the number of subjects, the year of acquisition and the quality of their meshes.

2.2 3DMM and blendshape models

The construction of a 3DMM, as introduced by the seminal work of Blanz and Vetter [5], consists of four main steps, namely data pre-processing, bringing the facial meshes into a common space by removing rotation, scale and translation, establishing group-wise dense correspondence between a training set of facial meshes, and finally performing some kind of statistical analysis, usually PCA, on the registered data to produce a low-dimensional model. Variations on steps of the aforementioned process led to different versions of 3DMMs. Lüthi [33] proposes to use Gaussian Process (GP) in order to construct 3DMMs, which was shown to exhibit better capacity in this respect. A different approach was suggested in [20] where a dictionary learning was used to form a 3D face shape

Dataset	Year	Type	Unique Participants	No vertices	Expressions	Public Available	Coverage
BU-4DFE [52]	2008	4D	101	35,000	6	Yes	Face
BOSPHORUS [42]	2008	3D	105	-	34	Yes	Face
BP4D-Spontaneous [56]	2014	4D	41	30,000-50,000	27 AU	Yes	Face
FaceWarehouse [12]	2014	4D	150	11,000	20	Yes	Face
4D CCDb [34]	2015	4D	4	30,000	Speech	Yes	Face
LSFM [9]	2016	3D	10,000	60,000	Neutral only	No	Face
LYHM [18]	2017	3D	1,200	180,000	Neutral only	Yes	Head
4D-FAB [14]	2018	4D	180	75,000	6	No	Face
COMA [41]	2018	4D	12	5023	11	Yes	Face
S3DFM [54]	2019	4D	77	-	Speech	Yes	Face
VOCASET [17]	2019	4D	127	5023	Speech	Yes	Face
SIAT-3DFE [51]	2020	4D	12	500K-1M	16	Yes	Face
FaceScape [50]	2020	4D	938	29,587	20	Yes	Face
MimicME (proposed)	2022	4D	4,700	60,000	20	Yes	Face

Table 1: 3D and 4D facial datasets.

model leading to better performance in terms of reconstruction and fitting accuracy than the PCA-based 3DMM. [31] introduced Gaussian mixture model in 3DMMs by assuming that the global population was a mixture of Gaussian sub-populations, each with its own mean and a shared covariance.

In a similar manner, blendshapes models (or expression models) are 3DMMs that consider the identity and expression as two distinguishable parts and create two different models for each of these parts. Depending on the way that these two parts are combined, the models in the literature can be classified to three categories: **additive**, **multiplicative**, and **nonlinear models** [19].

Additive models represent the expressions as the offset between a shape with expression and the neutral shape of a subject [4,3,45]. **Multiplicative models** combine identity and expression in a multiplicative manner, which in most of the cases is done by exploiting tensors. The main concept is to stack the 3D face data into a tensor and performing higher-order tensor decomposition (HOSVD) instead of PCA [48,6,11,7,49]. These models are characterized by their expressiveness and simplicity but require data with semantic correspondence, specified by expression labels. Finally, in **nonlinear models**, facial variations are modelled with nonlinear transformations, such as a physical simulation [28] or Gaussian mixture models to represent facial shape and texture [31]. Li et al. [32] introduce FLAME, a non-linear 3D expressive head model that combines explicit control over jaw articulation with expression blendshapes. They register the 3D meshes based on a non-rigid ICP method regularized by the face model. Many of the recent DNN methods belong to this category. Several recent deep learning approaches adopt autoencoder frameworks to build nonlinear 3DMMs by learning a relevant latent space, see e.g. [1,46].

2.3 Appearance models

In most cases, 3DMMs include also a modelling of the facial appearance. As in the case of the shape model, the appearance model of a 3DMM is built by performing statistics on the appearance information of the training shapes, which is represented either in terms of per-vertex values or as a texture in the so-called UV-space [19]. The latter is more popular as it does not require the shape and appearance to have the same resolution and the texture in the UV-space is treated as a 2D image, meaning that standard techniques for image processing and analysis are easily applied.

Appearance models of the facial texture in UV-space are grouped into **linear** and **nonlinear**. **Linear models** include the original work by Blanz and Vetter [5], the work of [18] for head and the work of [8] for face, to name a few. **Nonlinear models** include most of the recent deep learning-based approaches, which learn a joint shape and texture model, see e.g. [23,22]. For the successful training of these models, the existence of a large scale dataset of UV maps plays a crucial role. However, the solutions of the existing literature are not satisfactory and this is a gap that we fill with the proposed database.

3 MimicMe Database

3.1 Data Acquisition

The proposed database (MimicMe) was collected during a special exhibition in the Science Museum, London, over a period of 3 months. A large number of museums' visitors (4,700) volunteered to be recorded by a 3dMD* face capture system while performing various expressions. To avoid a biased dataset due to people's difficulties to act naturally, especially for the case of specific expressions [19], each subject had to watch twice a video of actors performing the expressions shown in Fig. 2. During the fist playback, the subject familiarized herself with the expressions she had to perform and in the second playback, when the capture process was happening, she had to mimic the actors' expressions. In that way, 20 expressions had been recorded. The frame rate of capture was between 2 and 4 frames per second. We also kept demographics information for each subject such as ethnicity, gender and age as it can be seen in Table 2 and Fig. 3. For each subject, a sequence of 70-120 3D images was created. The 3D triangular surface composed of approximately 120,000 vertices joined into approximately 250,000 triangles, along with a high-resolution texture map. The total number of captured 3D images is roughly $280,760$, which is larger than all previous expression-controlled 3D face datasets.

3.2 Annotation and registration process

In order to be able to use the 3D data in a statistically meaningful way, we need to register them into a common template such that all the meshes share the

*https://3dmd.com/

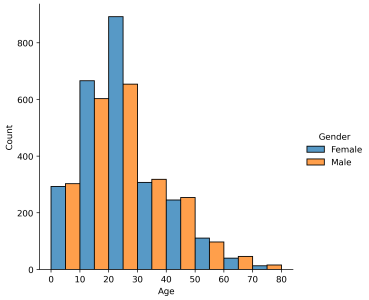

Fig. 3: Distribution of age and gender in MimicMe database.

Ethnicity	Number	Proportion
Caucasian	3,545	75.43%
Asian	612	13.02%
Black	112	2.38%
Mixed	326	6.94%
Other	105	2.23%

Table 2: Ethnicity distribution in MimicMe database.

same number of vertices joined into a common triangulation. In literature, the methods for performing 3D registration are classified into two categories based on the space where registration is done. The first category of methods exploits the advances of algorithms in 2D images and perform the dense registration in the 2D projections of the 3D meshes, namely their UV counterparts [40,16]. The second class of methods registers directly (i.e. in the 3D space) the mesh and the template [3,36].

As the former methods present some drawbacks like introducing non-linearities into the process and the need of rasterizing the UV image [8], we chose to register our 3D meshes using a method from the latter class of methods, namely performing the registration in the 3D space. There are plenty of methods and frameworks in this family like [59,25,8]. We opt for the framework used in building the LSFM [8], which is open source, publicly available and is based on Non-rigid Iterative Closest Point (NICP) for the 3D mesh registration. Since LSFM pipeline is used mainly on datasets with neutral faces, it is not suitable for datasets with faces with expressions. To address this problem, we need to adapt the template to handle each deformed face. We follow the updated approach of LSFM presented in [47] where the template used for registration is not the same for all the meshes to be registered but it is deformed driven by a set of landmarks of each mesh and the expression blendshape model built in [14] .

More precisely, a face detection and alignment model [55] is applied to each mesh and its corresponding color image. Then, a set of 68 sparse 2D landmarks is predicted which are easily projected in their corresponding 3D landmarks by exploiting the correspondence between the color image and the depth map. The predicted 3D landmarks are used to align the mesh to the template. To reduce effects from the identity difference between the template and the mesh, we register the neutral shape of each person, and we calculate the blendshape parameters $\mathbf{c}_b$ for a mesh through linear regression between the landmarks of the registered neutral shape and mesh's landmarks as follows:

$$||l_{\mathbf{x}_k} - \mathbf{A}(\mathbf{x}_n + \mathbf{U}_b\mathbf{c}_b)||_F^2 \tag{1}$$

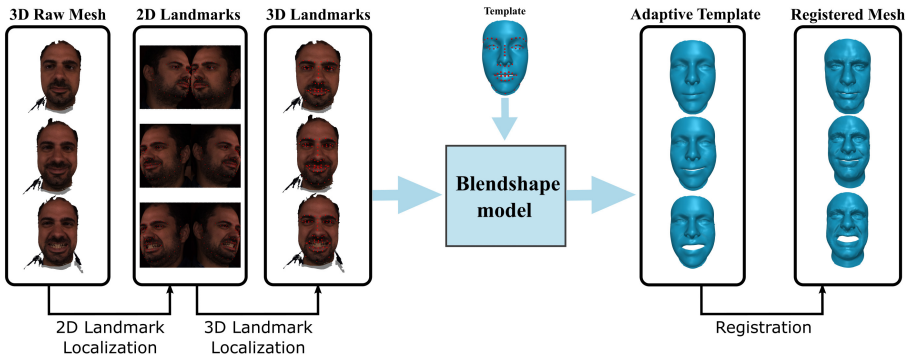

Fig. 4: Registration pipeline. The process starts by extracting 2D landmarks from the texture image, and then their corresponding position on the 3D mesh. By applying a regression between the 3D landmarks of the raw mesh and the neutral registered mesh, we create an adaptive template. Using this template and the 3D shape of the raw mesh, NICP can accurately register the raw mesh.

where $l_{\mathbf{x}_k} \in \mathbb{R}^{3m}$ is a vector with the m landmarks, $\mathbf{A} \in \mathbb{R}^{3m \times K}$ is an indicator matrix, $\mathbf{x}_n \in \mathbb{R}^{3n}$ is the neutral registered shape, $\mathbf{U}_b \in \mathbb{R}^{3n \times s}$ is a matrix with the blendshapes and $\mathbf{c}_b \in \mathbb{R}^s$ is a vector with parameters for the blendshapes.

Finally, we perform dense registration with NICP between the adaptive template and the mesh. In addition, we compute the corresponding texture for each registered mesh. We use a rectangular UV map for representation of the extracted texture as shown in Fig. 6 in order to make the training of GAN simpler.

As compared to [47], our pipeline has several advantages that increase the accuracy of model building in such dynamic scenarios. The most crucial one is the automatic correction of the predicted landmarks. After predicting the 3D landmarks for each mesh, we manually selected one neutral and the apex meshes (meshes with maximum facial change) of each expression per person and corrected the 3D landmarks (Fig. 5) using a 3D landmarking tool[¶]. In total, $55,000$ meshes were manually annotated. A statistical shape model for the 3D landmarks was built using the manually corrected landmarks. This model was used to correct the rest of the automatic predicted 3D landmarks, adding one more step in our pipeline 4.

3.3 Texture Completion

During the collection of such large-scale dataset, it is inevitable to obtain scan failures for some portion of the data. Such failures are often due to misplacement of the subject with respect to the camera or self-occlusions, e.g. occlusions due to hair, clothes, accessories. We estimated that a non-negligible portion of our

[¶]https://github.com/menpo/landmarker.io

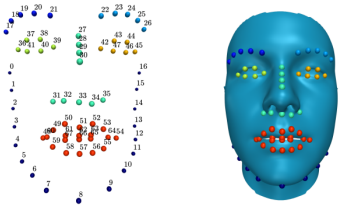

Fig. 5: Definition of the 68 3D facial landmarks (left) and an example of annotated template (right). Landmarks with the same color belong to the same semantic group (e.g. orange for the left eye).

Fig. 6: UV maps examples of registered meshes

Fig. 7: Texture Completion by projecting visible part of the texture to a GAN that is trained with complete textures, as explained in [21].

dataset is affected by such scanning failures and propose the following approach to inpaint missing parts of the texture maps.

We first gather a subset with samples where the subjects have been scanned properly and train a StyleGANv2 [29] from this subset dataset. Then, we project the remaining incomplete textures by using the method explained in [21], with a mask of visible texture, in order to hallucinate missing parts. Finally, we inpaint the texture by alpha blending between the original and hallucinated parts to achieve completed texture map. Some examples of this approach are shown in Fig. 7

3.4 Creating Expression Blendshapes

One of the ways that we have exploited our rich dynamic dataset was by using it to build a blendshape model following the standard process used in additive models, see e.g. [37]. In particular, we used the registered meshes from the previous step and the neutral shape of each sequence. For each of the sequences, we subtracted the neutral mesh of the sequence from each frame. After that, for each subject, we have a sequence of difference vectors, namely $\mathbf{d} \in \mathbb{R}^{3n}$ which were then stacked into a matrix $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_k] \in \mathbb{R}^{3n \times k}$, where n is the number of

Method	AN	DI	FE	HA	SA	SU	PA
FW	65.4	57.8	**44.4**	84.9	67.8	78.6	49.2
4DFAB	66.6	59.0	44.2	85.8	70.2	81.4	53.8
Ours	**69.4**	**61.7**	44.0	**86.1**	**70.6**	**81.6**	**62.9**

Table 3: Recognition Rates (RR) [%] obtained from facial expression experiments using 7 expressions (AN-Anger, DI-Disgust, FE-Fear, HA-Happiness, SA-Sadness, SU-Surprise, PA-Pain)

vertices in our mesh. Finally, incremental PCA was applied to our difference matrix $\mathbf{D}$ to identify the deformation components. We keep 28 blendshapes which correspond to 99.9% of the total variance.

4 Experiments

4.1 Facial expression recognition

We performed the standard FER experiments on all the expressions, namely the 6 prototypic expressions anger, fear, disgust, surprise, happiness, sadness, and some proposed expressions like pain, flaring of nostrils, inflating of cheeks, pout, showing the teeth, winking each eye, biting the upper and lower lip and moving the mouth left and right. We selected 700 subjects of our dataset that haven't been used for the creation of our blendshape model for training and testing. In total, almost 7000 meshes were utilised. We created a 10-fold partition, every time one fold was used for testing, the others were used for training.

We selected the apex frames of each expression sequence, so for each subject we had 20 meshes. For each mesh, the difference from the neutral mesh of the sequence was calculated and then the weights were found by projecting the difference to each blendshape model. In this way, every mesh was represented by a set of 30 parameters. After this, a multi-class SVM was employed to classify expressions. Radial Basis Function (RBF) kernel was selected, whose parameters were chosen by an empirical grid search. We achieved a recognition rate of 66.1%, 63.9%, and 68.3%, for 4DFAB, FaceWarehouse and our blendshape model, respectively. Table 3 shows the recognition rate for some expressions.

4.2 Evaluation of the expression blendshape model

We compare our blendshape model with FaceWarehouse (FW) [12] and 4DFAB [14] models in expression reconstruction. We randomly selected 2079 frames from 192 subjects that display various expressions, from which, we computed the facial deformation in the same way as described in [14] and reconstructed it using both blendshape models. We calculate the reconstruction error and plot the cumulative error curves for models with different number of expression components in

Fig 8. To provide a fair comparison with FW and 4DFAB, we report the performance of our model using the same number of components as the other models. It is clear that our blendshape model largely outperforms FW and 4DFAB models.

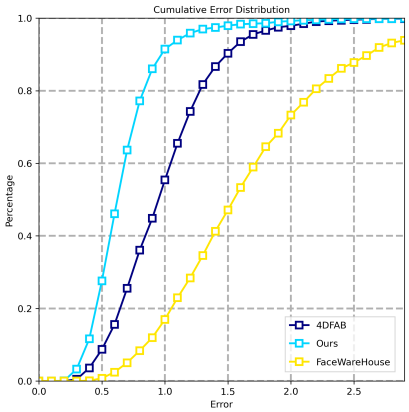

Fig. 8: Cumulative reconstruction errors achieved with different blendshapes over randomly selected expressions from our database.

Additionally, we plot some 3D expression transfer examples in Fig 9. For visualization reasons, we removed the ears and neck from the faces. For each expression transfer, the facial deformation from the corresponding neutral face was calculated and the blendshape parameters were computed using the aforementioned blendshape models. We then cast the reconstructed expression on a mean face for visualisation. Note that we fixed the number of our expression components to be identical for every model. Our blendshape model can faithfully reconstruct expressive faces with correct expression meaning.

4.3 Non-linear 3D Expression Model

Generating realistic faces in 3D is of high importance for many computer graphics and computer vision applications. In recent studies [23,24,43], Generative Adversarial Networks (GANs) have been trained by large-scale *private* datasets to generate high-quality textures of faces. However, since these datasets are private, it is difficult to explore the potential of such direction.

In order to demonstrate that MimicMe dataset is useful for training more sophisticated 3D face models such as GANs, we train a joint GAN model [22] that can synthesize shape, expression, texture and normals. Figures 1 and 10 shows some random generations from this generator. It is worth observing how the expression is reflected to both shape and texture and that the generator can synthesize wide range of expressions as well as identity.

Fig. 9: Comparison of FaceWarehouse[12] blendshape model, 4DFAB[14] and our expression blendshape model. The face in which the blendshapes are applied is the mean face, namely we do not transfer details of face identity

Generally, the deformation caused by expressions is often split between shape and texture in the current capture systems. However, the modelling of expression is isolated from the modelling of shape and texture, placing a fundamental limit to the synthesis of semantically meaningful 3D faces. Therefore, our dataset becomes particularly useful for exploiting the correlation between expression, texture and shape by deep generative models.

5 Conclusion

We have presented MimicMe, a large-scale detailed 3D facial dataset that can be used for biometric applications, facial expression analysis and generation of realistic faces in 3D for computer graphics. Compared to previous public large-scale 3D face datasets, MimicMe provides the largest diverse population, with high geometric quality. We demonstrate the usefulness of the database in a series of recognition experiments. Promising results are obtained with basic features and standard classifiers, thus we believe that even better results can be obtained using more recent deep methods. We built a powerful expression blendshape model from this database, which outperforms the state-of-the-art blendshape models and a Non-linear 3D Expression Model, which generated high-quality textures

(a) Normals (b) Texture (c) Rendered (d) Rend. 30° (e) Normals (f) Texture (g) Rendered (h) Rend. 30°

Fig. 10: Random synthetic faces generated by our model that is trained by our dataset. Please note the correlation between expression, texture and shape as well as the identity and expression diversity. Such correlation can be only achieved using a dataset that consists of large number of different identities under various expression, such as MimicMe.

of synthetic faces. We make MimicMe database publicly available for research purposes, which we anticipate to have a significant impact on the research in this field.

Acknowledgements S. Zafeiriou and part of research was funded by the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1).

References

1. Abrevaya, V.F., Wuhrer, S., Boyer, E.: Multilinear autoencoder for 3d face model learning. In: WACV 2018-IEEE Winter Conference on Applications of Computer Vision (2018) 6
2. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3d face recognition with a morphable model. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on. pp. 1–6. IEEE (2008) 2
3. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. pp. 1–8. IEEE (2007) 6, 8
4. Blanz, V., Basso, C., Poggio, T., Vetter, T.: Reanimating faces in images and video. In: Computer graphics forum. vol. 22, pp. 641–650. Wiley Online Library (2003) 6
5. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194. ACM Press/Addison-Wesley Publishing Co. (1999) 2, 5, 7
6. Bolkart, T., Wuhrer, S.: 3d faces in motion: Fully automatic registration and statistical analysis. Computer Vision and Image Understanding **131**, 100–115 (2015) 6
7. Bolkart, T., Wuhrer, S.: A robust multilinear model learning framework for 3d faces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4911–4919 (2016) 6
8. Booth, J., Roussos, A., Ponniah, A., Dunaway, D., Zafeiriou, S.: Large scale 3d morphable models. International Journal of Computer Vision **126**(2-4), 233–254 (2018) 7, 8
9. Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., Dunaway, D.: A 3d morphable model learnt from 10,000 faces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5543–5552 (2016) 4, 6
10. Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., Zafeiriou, S.: Neural 3d morphable models: Spiral convolutional networks for 3d shape representation learning and generation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7213–7222 (2019) 3
11. Brunton, A., Salazar, A., Bolkart, T., Wuhrer, S.: Review of statistical shape spaces for 3d data with comparative analysis for human faces. Computer Vision and Image Understanding **128**, 1–17 (2014) 6
12. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2014) 5, 6, 11, 13
13. Cheng, S., Bronstein, M., Zhou, Y., Kotsia, I., Pantic, M., Zafeiriou, S.: Meshgan: Non-linear 3d morphable models of faces. arXiv preprint arXiv:1903.10384 (2019) 3
14. Cheng, S., Kotsia, I., Pantic, M., Zafeiriou, S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 5, 6, 8, 11, 13
15. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3444–3453. IEEE (2017) 5
16. Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling.

In: 2011 International Conference on Computer Vision. pp. 2296–2303 (2011). https://doi.org/10.1109/ICCV.2011.6126510 8
17. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M.J.: Capture, learning, and synthesis of 3d speaking styles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10101–10111 (2019) 5, 6
18. Dai, H., Pears, N., Smith, W., Duncan, C.: A 3d morphable model of craniofacial shape and texture variation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3104–3112. IEEE (2017) 2, 6, 7
19. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020) 2, 4, 6, 7
20. Ferrari, C., Lisanti, G., Berretti, S., Del Bimbo, A.: Dictionary learning based 3d morphable model construction for face recognition with varying expression and pose. In: 3D Vision (3DV), 2015 International Conference on. pp. 509–517. IEEE (2015) 5
21. Gecer, B., Deng, J., Zafeiriou, S.: Ostec: one-shot texture completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7628–7638 (2021) 10
22. Gecer, B., Lattas, A., Ploumpis, S., Deng, J., Papaioannou, A., Moschoglou, S., Zafeiriou, S.: Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In: European Conference on Computer Vision. pp. 415–433. Springer (2020) 3, 7, 12
23. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1155–1164 (2019) 3, 7, 12
24. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.P.: Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 12
25. Gilani, S.Z., Mian, A., Shafait, F., Reid, I.: Dense 3d face correspondence. IEEE transactions on pattern analysis and machine intelligence **40**(7), 1584–1598 (2017) 8
26. Gong, S., Chen, L., Bronstein, M., Zafeiriou, S.: Spiralnet++: A fast and highly efficient mesh convolution operator. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019) 3
27. Guo, Y., Cai, J., Jiang, B., Zheng, J., et al.: Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. IEEE transactions on pattern analysis and machine intelligence **41**(6), 1294–1307 (2018) 4
28. Ichim, A.E., Kadleček, P., Kavan, L., Pauly, M.: Phace: physics-based face modeling and animation. ACM Transactions on Graphics (TOG) **36**(4), 1–14 (2017) 6
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 4, 10
30. Knoops, P.G., Papaioannou, A., Borghi, A., Breakey, R.W., Wilson, A.T., Jeelani, O., Zafeiriou, S., Steinbacher, D., Padwa, B.L., Dunaway, D.J., et al.: A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. Scientific reports **9**(1), 1–12 (2019) 2

31. Koppen, P., Feng, Z.H., Kittler, J., Awais, M., Christmas, W., Wu, X.J., Yin, H.F.: Gaussian mixture 3d morphable face model. Pattern Recognition **74**, 617–628 (2018) 6
32. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Transactions on Graphics (TOG) **36**(6), 194 (2017) 6
33. Lüthi, M., Gerig, T., Jud, C., Vetter, T.: Gaussian process morphable models. IEEE transactions on pattern analysis and machine intelligence (2017) 5
34. Marshall, A.D., Rosin, P.L., Vandeventer, J., Aubrey, A.: 4d cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations. Auditory-Visual Speech Processing,{AVSP} 2015 pp. 157–162 (2015) 5, 6
35. Moschoglou, S., Ploumpis, S., Nicolaou, M.A., Papaioannou, A., Zafeiriou, S.: 3dfacegan: Adversarial nets for 3d face representation, generation, and translation. International Journal of Computer Vision **128**, 2534–2551 (2020) 3
36. Myronenko, A., Song, X.: Point set registration: Coherent point drift. IEEE transactions on pattern analysis and machine intelligence **32**(12), 2262–2275 (2010) 8
37. Neumann, T., Varanasi, K., Wenger, S., Wacker, M., Magnor, M., Theobalt, C.: Sparse localized deformation components. ACM Transactions on Graphics (TOG) **32**(6), 179 (2013) 10
38. O'Sullivan, E., van de Lande, L., Oosting, A., Papaioannou, A., Jeelani, N., Koudstaal, M., Khonsari, R., Dunaway, D., Zafeiriou, S., Schievano, S.: The 3d skull 0–4 years: A validated, generative, statistical shape model. Bone Reports **15** (2021) 2
39. O'Sullivan, E., van de Lande, L.S., Papaioannou, A., Breakey, R.W., Jeelani, N.O., Ponniah, A., Duncan, C., Schievano, S., Khonsari, R.H., Zafeiriou, S., et al.: Convolutional mesh autoencoders for the 3-dimensional identification of fgfr-related craniosynostosis. Scientific reports **12**(1), 1–8 (2022) 2
40. Patel, A., Smith, W.A.: 3d morphable face models revisited. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1327–1334. IEEE (2009) 8
41. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018) 3, 5, 6
42. Savran, A., Alyuöz, N., Dibeklioglu, H., Celiktutan, O., Gökberk, B., Sankur, B., Akarun, L.: Bosphorus database for 3D face analysis. In: BIOID. pp. 47–56 (2008) 4, 6
43. Slossberg, R., Shamai, G., Kimmel, R.: High quality facial surface and texture synthesis via generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018) 12
44. Staal, F.C., Ponniah, A.J., Angullia, F., Ruff, C., Koudstaal, M.J., Dunaway, D.: Describing crouzon and pfeiffer syndrome based on principal component analysis. Journal of Cranio-Maxillofacial Surgery **43**(4), 528 – 536 (2015). https://doi.org/https://doi.org/10.1016/j.jcms.2015.02.005, http://www.sciencedirect.com/science/article/pii/S101051821500027X 2
45. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016) 6
46. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7346–7355 (2018) 6
47. Tzirakis, P., Papaioannou, A., Lattas, A., Tarasiou, M., Schuller, B., Zafeiriou, S.: Synthesising 3d facial motion from" in-the-wild" speech. In: 2020 15th IEEE Inter-

national Conference on Automatic Face and Gesture Recognition (FG 2020)(FG). pp. 627–634 (2020) 5, 8, 9
48. Vlasic, D., Brand, M., Pfister, H., Popović, J.: Face transfer with multilinear models. ACM transactions on graphics (TOG) **24**(3), 426–433 (2005) 6
49. Wang, M., Panagakis, Y., Snape, P., Zafeiriou, S.: Learning the multilinear structure of visual data. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4592–4600 (2017) 6
50. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 601–610 (2020) 6
51. Ye, Y., Song, Z., Guo, J., Qiao, Y.: Siat-3dfe: A high-resolution 3d facial expression dataset. IEEE Access **8**, 48205–48211 (2020) 6
52. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3d dynamic facial expression database. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. pp. 1–6 (2008). https://doi.org/10.1109/AFGR.2008.4813324 5, 6
53. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006) 4
54. Zhang, J., Fisher, R.B.: 3d visual passcode: Speech-driven 3d facial dynamics for behaviometrics. Signal processing **160**, 164–177 (2019) 5, 6
55. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016) 8
56. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P., Girard, J.M.: Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Computing **32**(10), 692–706 (2014) 6
57. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016) 4
58. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In: Computer Graphics Forum. vol. 37, pp. 523–550. Wiley Online Library (2018) 2
59. Zulqarnain Gilani, S., Shafait, F., Mian, A.: Shape-based automatic detection of a large number of 3d facial landmarks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4639–4648 (2015) 8