3D Face reconstruction using VRN

Rishav Chandra Varma [1]Prashant Mahanta [1] Kaustubh Pandey[1] Wasim Ishaq Khan[1]

Abstract— **Face reconstruction|(3D) is a basic computer vision problem. Most systems dealing with this issue assume that there are plenty of facial images are available to them as an input and also, they have to address many challenges like large facial poses,expressions,alignment, etc. Generally these methods requires complex pipelines for model building and fitting. Our method address many of these limitations by training a Convolutional Neural Network (CNN) on an appropriate dataset consisting of 2D images and 3D facial models (dataset: AFLW). The CNN does not require accurate alignment and works with a single 2D facial image to reconstruct the whole 3D facial geometry by bypassing the construction and fitting of a 3D Morphable Model. This is achieved by a simple CNN architecture that performs regression of a volumetric representation of the 3D facial geometry from a single 2D image. This also demonstrates how facial landmark localization can be included and into the framework and help improve or enhance reconstruction quality for the case of facial expressions.**

I. INTRODUCTION

3D face reconstruction is the problem of recovering the 3D facial geometry from 2D images. There are many variations and approaches to solve it. This work is on 3D face reconstruction using only a single image. Under this setting, the problem is considered far from being solved.In our approach we are directly learning a mapping from pixels to 3D coordinates using a Convolutional Neural Network (CNN). Our approach works with unconstrained images(with single face), including facial images of arbitrary poses , facial expressions or Occlusions.

Motivation. 3D face reconstruction requires in general complex pipelines and solving non-convex difficult optimization problems for both model building and fitting , no matter what the underlying assumptions are, what the input(s) and output(s) to the algorithm are. In the following paragraph, we provide examples from primary approaches:

- Using 3DMM: In the 3D Morphable Model (3DMM) [4, 5], the most popular approach for estimating the full 3D facial structure from a single image, training includes an iterative flow procedure for dense image correspondence which is prone to failure. Additionally, testing requires a careful initialization for solving a difficult highly non-convex optimization problem, which is slow.
- In [7], a state-of-the-art pipeline for reconstructing a highly detailed 2.5D facial shape for each video frame, an average shape and an illumination subspace for the specific person is firstly computed (offline), while testing is an iterative process requiring a sophisticated pose estimation algorithm, 3D flow computation between the model and the video frame, and finally shape refinement by solving a shape-from-shading optimization problem.
- The state-of-the-art method of [6] that produces the average 3D face from a collection of personal photos, firstly performs landmark detection, then fits a 3D Morphable Model (3DMM) using a sparse set of points, then solves an optimization problem, then performs surface normal estimation and finally performs surface reconstruction.

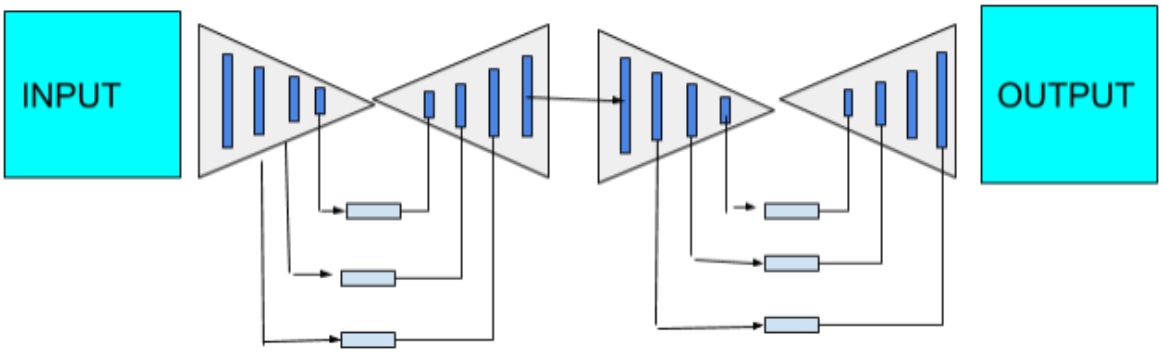
3D Morphable Model (3DMM): The most popular approach for estimating the full 3D facial structure from a single image (among others) is 3D Morphable Model (3DMM). The model has two components: (i) a mesh consisting of the mean face, and (ii) two matrices, one each for shape and texture that describe the various modes of variations from the mean. The number of modes of variation depends on the size of the mesh, and also is different for shape and texture. Hence the appearance of a given face can be summarized by a set of coefficients that describe how much there is of each mode of variation.

Problems with 3D Morphable Model are:

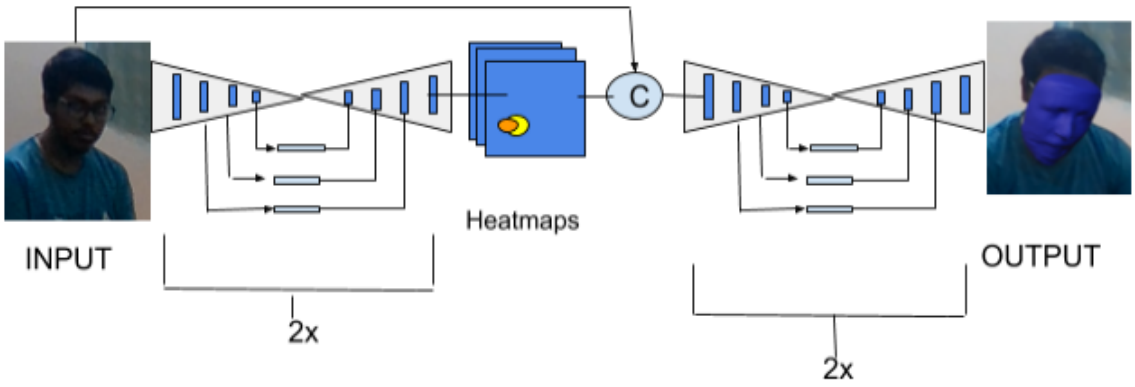
- Needs multiple facial images (of same or different subjects).
- Training includes an iterative flow procedure for dense image correspondence which is prone to failure.
- Testing requires a careful initialisation for solving a difficult highly non-convex optimization problem, which is slow.

Our method bypasses many of the difficulties which are faced during 3D face reconstruction by using volumetric representation of the 3D facial geometry, and an appropriate CNN architecture that is trained to regress directly from a 2D facial image to its corresponding 3D volume. An overview of the process is shown in [Fig 1]. In summary our contribution are:

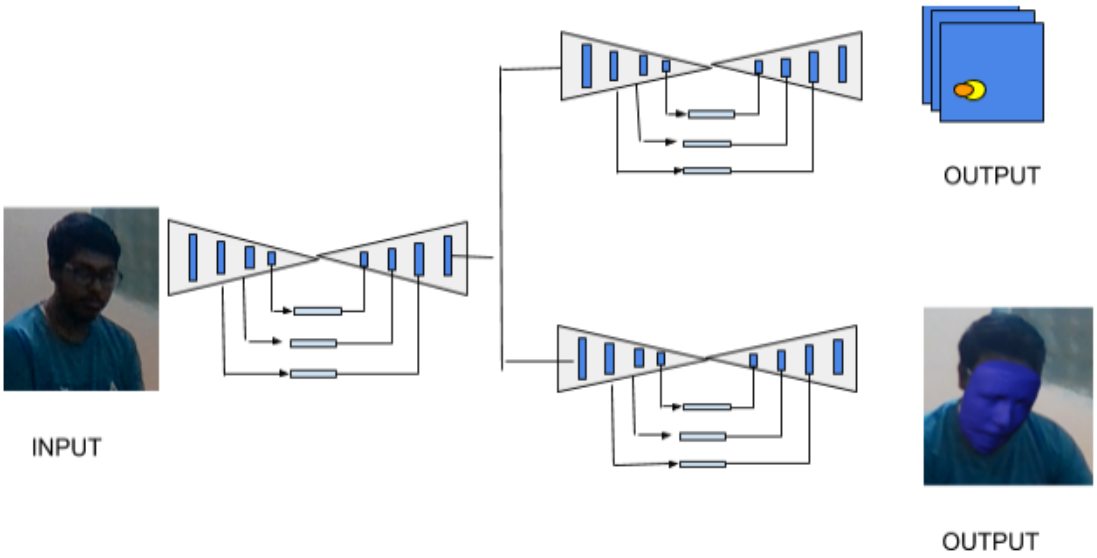
- Given a dataset consisting of 2D images and 3D face scans, we investigate whether a CNN can learn directly,in an end-to-end fashion, the mapping from image pixels to the full 3D facial structure geometry.
- Our CNN works with just a single 2D facial image, does not require accurate alignment, works for arbitrary facial poses and expressions, and can be used to reconstruct the whole 3D facial geometry bypassing the construction (training) and fitting (testing) of a 3DMM.
- We achieve this using a CNN that performs direct regression of a volumetric representation of the 3D facial geometry from a single 2D image.



(a) The proposed Volumetric Regression Network (VRN) accepts as input an RGB input and directly regresses a 3D volume completely bypassing the fitting of a 3DMM. Each rectangle is a residual module of 256 features.



(b)) The proposed VRN - Guided architecture firsts detects the 2D projection of the 3D landmarks, and stacks these with the original image. This stack is fed into the reconstruction network, which directly regresses the volume.



(c) The proposed VRN - Multitask architecture regresses both the 3D facial volume and a set of sparse facial landmarks.

Fig. 1. An overview of the proposed three architectures for Volumetric Regression: Volumetric Regression Network (VRN), VRN - Guided and VRN - Multitask.

II. METHOD

A. Dataset

Our method requires an appropriate dataset consisting of 2D images and 3D facial scans, to regress the full 3D facial structure from a 2D image. The dataset has been produced by fitting a 3DMM built from the combination of the Basel [8] and FaceWarehouse [9] models to the unconstrained images of the 300W dataset [10], careful initialisation and by constraining the solution using a sparse set of landmarks.Face profiling is then used to render each image to 10-15 different poses resulting in a large scale dataset (more than 60,000 2D facial images and 3D meshes) called 300W-LP. Note that because each mesh is produced by a 3DMM.

B. Volumetric representation

Our goal is to predict the coordinates of the 3D vertices of each facial scan from the corresponding 2D image using CNN regression.

III. Volumetric Regression Network

Our model has to learn a mapping from the 2D facial image to its corresponding 3D volume.

$$f : I \rightarrow V$$

This mapping can be learned by our CNN using the training set of 2D images and constructed volumes.This CNN architecture is based on the "hourglass network" of [1] which is an extension of the fully connected network of [2] using skip connections and residual learning [3]. Our architecture has two hourglass modules stacked together. Input is RGB image and output is a volume of 192 X 192 X 200 of real values.This network has a structure of encoder-decoder. A set of convolutional layers are firstly used to compute a feature representation of fixed dimension.This representation is further processed back to the spatial domain and spatial correspondence is re-established between input and output. The second hourglass does the task of refining and has identical structure as the first hourglass.

For training our VRN model, we use sigmoid cross entropy loss function:

$$l_1 = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D}[V_{whd} log\hat{V}_{whd} + (1 - V_{whd}) log\hat{V}_{whd}]$$

,

where $\hat{V}_{whd}$ is the corresponding sigmoid output at the voxel w,h,d of the regressed volume.

During testing, for an input 2D image, our VRN model regresses a 3D volume from which the outer 3D facial mesh is recovered. We are using soft sigmoid output instead of hard(binary) predictions as the former produces smoother results.

Finally from the 3D volume, by generating the iso-surface of this volume, a mesh can be formed.

A. VRN - Multitask and VRN - Guided

VRN - Multitask. We also propose to use the VRN - Multitask model (shown in fig.) which consists of three hourglass modules. The task of the first hourglass module is to provide features to a fork of 2 hourglasses (in the pipeline). The first of this fork regresses the 68iBUG landmarks [10] as 2D Gaussians, each on a separate channel. The second hourglass of this fork directly regresses the 3D structure of the face as a volume, as in the aforementioned unguided volumetric regression method. The goal of this multitask network being to learn more reliable features which are better suited to the two tasks.

VRN - Guided. We propose that reconstruction should benefit when simpler face analysis task is performed first; in particular we propose an architecture for volumetric regression guided by facial landmarks. To achieve this, We train a stacked hourglass network which accepts guidance from landmarks during training and inference. The architecture of this network is similar to the unguided volumetric regression method; the only difference between the two networks being the input. The input to this architecture is an RGB image stacked with 68 channels, each consisting a Gaussian (standard deviation = 1, approximate diameter of 6 pixels) centered on each of the 68 landmarks. This stacked representation and architecture is depicted in fig. While training the network, We used the ground truth landmarks while during testing we employed a stacked hourglass network trained for facial landmark localisation. We refer to this network as VRN - Guided.

IV. Training

Each of the above mentioned architectures was trained end-to-end using RMSProp with an initial learning rate of 10^{-4},which was lowered after 40 epochs to 10^{-5} .During training,random augmentation was applied to each input sample (face image) and its corresponding target (3D volume) : we applied in-plane rotation $r \ \epsilon \ [-45°, ..., 45°]$, translation $t_z, t_y \ \epsilon \ [-15, ..., 15]$ and scale $s \ \epsilon \ [0.85, ..., 1.15]$ jitter. In 20% of cases,the input and target were flipped horizontally. Finally, the input samples were adjusted with some colour scaling on each RGB channel.

In the case of the VRN-Guided,the landmark detection module was trained to regress Gaussians with standard deviation of approximately 3 pixels ($\sigma = 1$).

V. Results

We tested our VRN model on two datasets AFLW2000-3D and BU-4DFE.The performance of our proposed models are: From these results, we can conclude the following:

- Volumetric Regression Networks is able to outperform the state of the art methods 3DDFA and EOS on all datasets, showing that directly regressing the 3D facial structure is a much easier problem.
- All VRNs perform well for different types of facial poses, expressions and occlusions.
- VRN-guided outperforms other VRNs, however it has higher computational complexity.

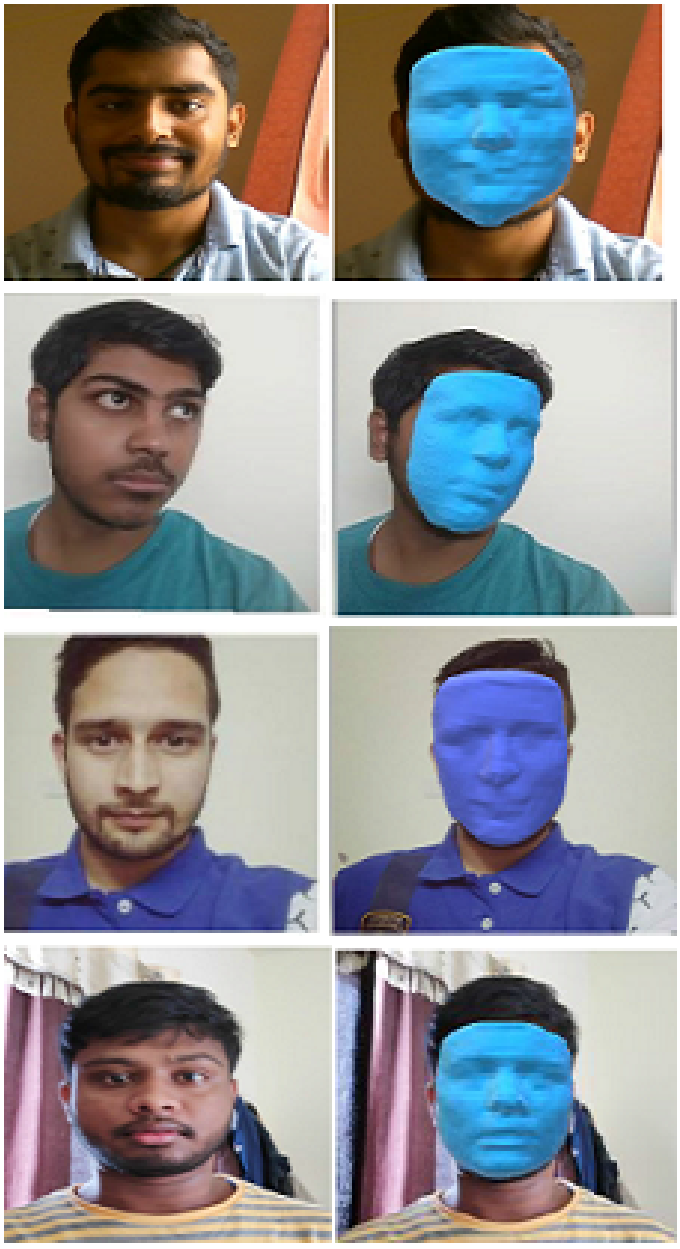


Fig. 2. Some visual results from the AFLW2000-3D dataset generated using our VRN - Guided method.

- VRN - Multitask does not always perform particularly better than the plain VRN. It doesn't justify the increase of network complexity.

Details about our experiment are as follows: **Datasets. (a) AFLW 2000-3D:** Our primary objective was to perform the testing of our network on totally unconstrained images, so we firstly worked on the ALFLW 2000-3D [11] dataset which contains 3D facial meshes for the first 2000 images from AFLW [12]. **(b) BU-4DFE:** We also worked on the rendered images from BU-4DFE [13]. We rendered each participant for both Happy and Surprised expressions with three different pitch rotations between -20 and 20 degrees. For each pitch, seven roll rotations from -80 to 80 degrees were also rendered. Large variations in lighting direction and colour were added randomly to make the images more challenging.

TABLE I
RECONSTRUCTION ACCURACY BASED ON NME

Method	AFLW2000-3D	BU-4DFE
VRN	0.0676	0.0600
VRN-Multitask	0.0698	0.0625
VRN -Guided	**0.0637**	**0.0555**
3DDFA	0.1012	0.1227
EOS	0.0971	0.1560

Error Metric. To measure the accuracy of reconstruction for each face, we used the Normalised Mean Error (NME) defined as the average per vertex Euclidean distance between the estimated and ground truth reconstruction normalised by the outer 3D interocular distance:

$$NME = \sum_{k=1}^{N} ||x_k - y_k||/d$$

where N is the number of vertices per facial mesh, d is the 3D interocular distance and xk, yk are vertices of the grouthtruth and predicted meshes. The error is calculated on the face region only on approximately 19,000 vertices per facial mesh.

VI. CONCLUSIONS

We proposed a direct approach to 3D facial reconstruction from a single 2D image using volumetric CNN regression. To this end, we proposed and exhaustively evaluated three different networks for volumetric regression, reporting results that show that the proposed networks perform well for the whole spectrum of facial pose, and can deal with facial expressions as well as occlusions. We also compared the performance of our networks against that of recent state-of-the-art methods based on 3DMM fitting reporting large performance improvement on three different datasets. Future work may include improving detail and establishing a fixed correspondence from the isosurface of the mesh

REFERENCES

[1] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.
[2] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016.
[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Computer graphics and interactive techniques, 1999.
[5] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In CVPR, 2005.
[6] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In CVPR, 2016.
[7] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In ECCV, 2014.
[8] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In AVSS, 2009.
[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. IEEE TVCG, 20(3), 2014.

[10] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In CVPR-W, 2013.
[11] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. 2016.
[12] M. Koestinger, P.Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, realworld database for facial landmark localization. In First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[13] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A highresolution 3d dynamic facial expression database. In Automatic Face Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, pages 1–6. IEEE, 2008.
[14] Jackson, Aaron S and Bulat, Adrian and Argyriou, Vasileios and Tzimiropoulos, Georgios.Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression. In International Conference on Computer Vision,2017