Paper Review: Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression

Antoni Mauricio
Research and Innovation Center in Computer Science
Universidad Católica San Pablo
manasses.mauricio@ucsp.edu.pe

Jorge López
Research and Innovation Center in Computer Science
Universidad Católica San Pablo
jorge.lopez.caceres@usp.pe

***Abstract*—3D reconstruction is one of the most recurrent problems in computer vision, its complexity goes through problems from objetc segmentation in a 3D environment to matching of points, most applications assume the availability of multiple images from different points of view as input and address a lot of propper issues. In 3D facial reconstruction, this issues include non-uniform illumination, expression and facial poses recognition in multiple facial images. The presented work proposes a nobel method based on Convolutional Neuronal Networks (CNN) using an 2D images as input dataset and voxel binary values (from 3D images and scans) as target dataset, but just a single 2D facial image in 3D reconstruction (using a trained model). The 3D output model works for arbitrary facial poses and expressions, and can be used to reconstruct whole 3D geometry, even non-visible parts of the face, bypassing the geometry model (in training) and fitting of a 3D Morphable Model (in testing). The method proposed include facial landmark localization task in CNN pipeline to improve reconstruction quality, especially for the cases of large poses and facial expressions, then CNN performs 2D input volumetric representation by direct regression of the 3D facial geometry output.**

1. Introduction

3D face reconstruction implies to build a 3D geometry model from 2D images. Many solutions and approaches have been used through years to solve it, depending on some assumptions. One of most studied assumptions is high-quality facial reconstruction from a single image, considering non-uniform illumination, hidden areas, expression and facial poses recognition. In this paper, a CCN architecture is used to correlate statistically a mapping from 2D pixels to voxels in 3D coordinates.

The most popular approach in one image techniques for 3D face reconstruction is the 3D Morphable Model (3DMM) [1], [2]. 3DMM describes the 3D face space with a Principal Component Analisys (PCA), but applied over an iterative flow procedure for dense image correspondence becomes prone to failures. Also, its application requires a careful parameters initialisation in order to solve a difficult highly non-convex optimization problem (too slow).

Many other papers [3], [4], [5] works using CNN architectures, and varing stack parameters, mostly based on 3DMM or 2.5DMM.

2. Method

The main idea is do a probabilistic match between 2D image input and 3D image/ scan output. FOr training the 3D image is represented as 3D binary matrix, which means that classification is done using logistic regression, in paper's case is sigmoid cross entropy loss function, which could be represented as the probability to choose a class while no the other. THis cost function is presented in equation 1:

$$l_1 = \sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{d=1}^{D} W_{whd} log(W'_{whd}) + (1 - W_{whd}) log(1 - W'_{whd}) \quad (1)$$

Where W'_{whd} is the corresponding sigmoid output at voxel w, d, h of the regressed volume.

Datasets are obteined from [4], and main volumetric regression model is based on hourglass network.

2.1. Hourglass Network

The design of the hourglass is motivated by the need to capture information at every scale. While local evidence is essential for identifying features like faces and hands, a final pose estimate requires a coherent understanding of the full body. The persons orientation, the arrangement of their limbs, and the relationships of adjacent joints are among the many cues that are best recognized at different scales in the image. The hourglass is a simple, minimal design that has the capacity to capture all of these features and bring them together to output pixel-wise predictions.

2.2. Residual Network

Due backpropagation learning impact reduction, some feedback from original source is needed, then a residual signal from original is added.

Acknowledgments

This work was supported by grant 234-2015-FONDECYT (Master Program) from Cienciactiva of the National Council for Science,Technology and Technological Innovation (CONCYTEC-PERU).

References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In Computer graphics and interactive techniques, 1999.

[2] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In CVPR, 2005.

[3] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In CVPR, 2016.

[4] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. 2016.

[5] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 34763483. IEEE, 2013.