Joint Face Detection and Facial Motion Retargeting for Multiple Faces

Bindita Chaudhuri[1], Noranart Vesdapunt[2], Baoyuan Wang[2]
[1]University of Washington, [2]Microsoft Corporation
[1]bindita@cs.washington.edu, [2]{noves,baoyuanw}@microsoft.com

Supplementary Material

Our goal is to perform live facial motion retargeting from multiple faces in a frame to 3D characters on mobile devices. We propose a lightweight multi-scale network architecture to disentangle the 3DMM parameters so that only the expression and pose parameters can be seamlessly transferred to any 3D character. Furthermore, we avoid the performance overhead of running a separate face detector by integrating face detection with parameter estimation.

1. Network Topology

The architecture of our single scale single face retargeting network is shown in Fig. 1. The details of each block are given in our paper. The resolution (scale) of the image feature maps is reduced by 2 after every block, and the pose, expression and identity parameters are learned from the same feature map, hence the term single scale. We designed our multi-scale single face retargeting network to learn different groups of parameters from separate branches that represent image features at different scales.

2. Multi-face Retargeting Network Outputs

As mentioned in our paper, the multi-face retargeting network divides the input image into 9×9 grid and predicts 5 bounding boxes for each grid cell. Each bounding box b has the following co-ordinates: t_x, t_y, t_w, t_h, t_o and $t_{v_{1-104}}$. The final outputs (b_x, b_y - x, y co-ordinates of the box centroid, b_w, b_h - width and height of the box, b_o - objectness score, $b_{\text{id}}, b_{\text{exp}}, b_{\mathbf{R}}, b_{\mathbf{t}}, b_{\text{f}}$ - 3DMM parameters and b_{lm} - corresponding 2D landmarks) are then given by:

$$b_x = \sigma(t_x) + c_x;\ b_y = \sigma(t_y) + c_y;\ b_w = p_w * e^{t_w};\ b_h = p_h * e^{t_h}$$

$$b_o = Pr(\text{face}) * IOU(b, \text{face}) = \sigma(t_o)$$

$$b_{\text{id}} = t_{v_{1-50}};\ b_{\text{exp}} = \sigma(t_{v_{51-97}})$$

$$b_{\mathbf{R}} = t_{v_{98-101}};\ b_{\mathbf{t}} = t_{v_{102-104}};\ b_{\text{f}} = \sigma(t_{v_{105}})$$

$$b_{\text{lm}_x} = b_x + b_w * b_{\hat{\text{lm}}_x};\ b_{\text{lm}_y} = b_y + b_h * b_{\hat{\text{lm}}_y}$$

where σ denotes sigmoid function, (c_x, c_y) is the offset of the grid cell containing b from the top left corner of the image, (p_w, p_h) are the dimensions of the bounding box prior (anchor box), $b_{\hat{\text{lm}}}$ are the initial landmarks obtained using and IOU denotes intersection over union. As evident from the equations, the landmark loss puts additional constraints on the bounding box locations and dimensions, thereby improving the accuracy of face detection in the joint training compared to simple face detection.

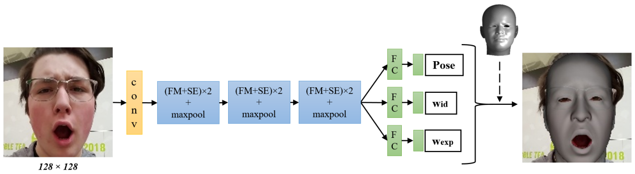

Figure 1: Our single scale single face retargeting network.

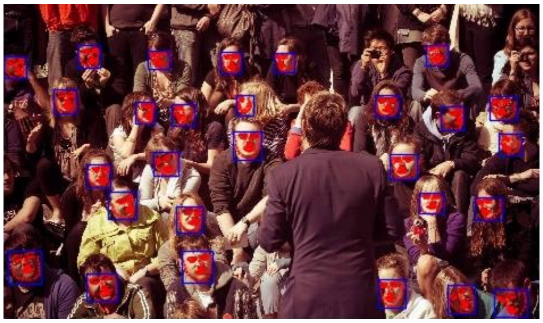

Figure 2: Network output for an image with multiple small faces from the AFW dataset.

3. Performance of Face Detection

Our network can detect multiple small faces of reasonable size even though it is not trained on images more than 20 faces. Figure 2 shows our network outputs for an image with more than 20 faces in the AFW dataset. The blue rectangles denote the predicted bounding boxes and the red points denote the 2D landmarks (for better viewing) projected from the predicted 3DMM parameters.

4. More Qualitative Results

Fig.3 and Fig.4 show more retargeting results for images with single or multiple faces using our method. The per-

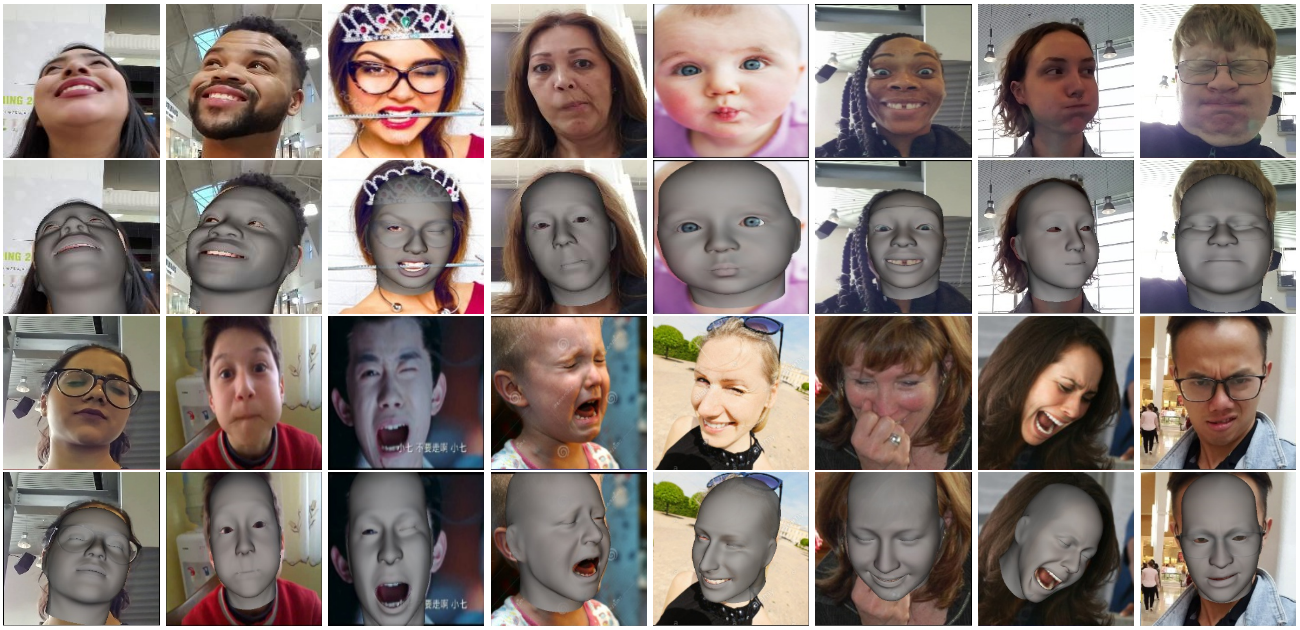

Figure 3: More results from our own expression test set using our single face retargeting network.

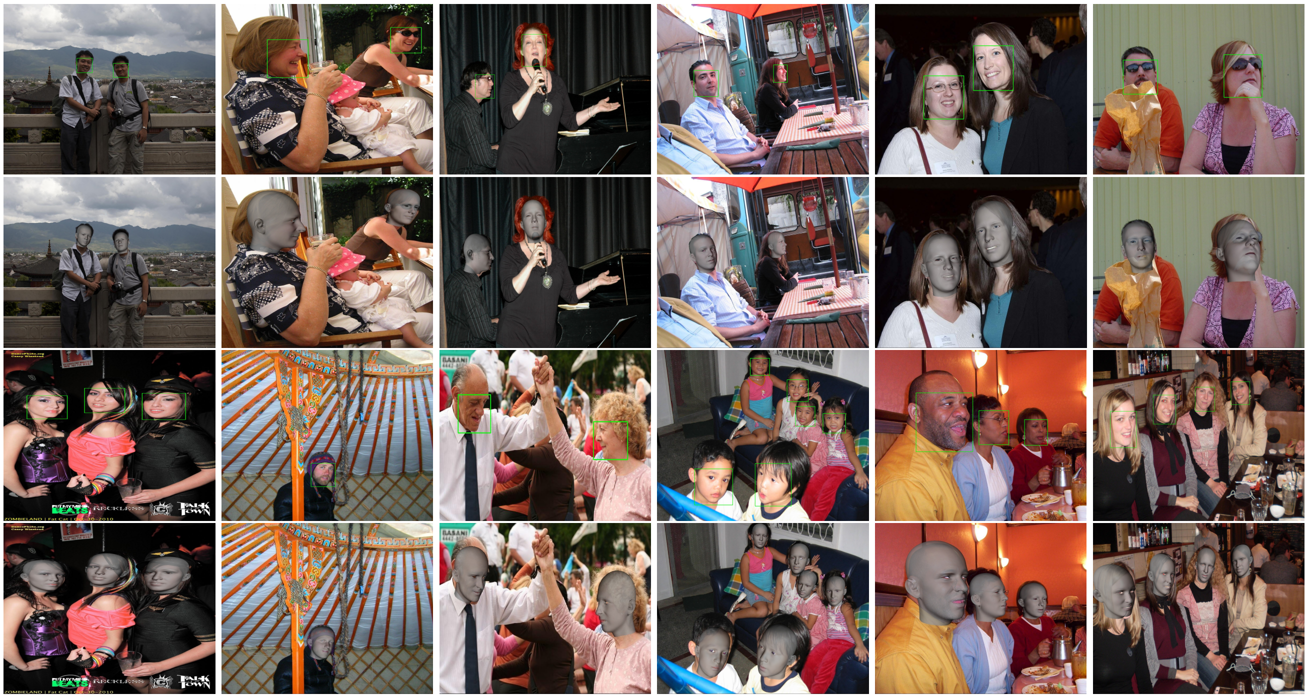

Figure 4: More results from AFW dataset using our joint detection and retargeting model.

formance of our networks on face videos is shown on the project webpage[1]. In the first half of the video, we show the results of retargeting from a single face video to a generic 3D human face model using our single face retargeting network. The face bounding box for the current frame is obtained from the boundaries of the 2D landmarks predicted in the previous frame. In the second half of the video, we show the retargeting results with videos having multiple faces using our multi-face retargeting network (only frames with multiple faces are shown).

[1] `https://homes.cs.washington.edu/~bindita/multifaceretargeting.html`