Pixel-based Facial Expression Synthesis

Arbish Akram and Nazar Khan
Computer Vision and Machine Learning Group
Punjab University College of Information Technology (PUCIT)
Lahore, Pakistan

Overview

- Facial expression synthesis (FES) has achieved remarkable advances with the advent of Generative Adversarial Networks (GANs).
- While effective, these GAN-based FES models are limited in these aspects:
  1. They generate photo-realistic results as long as testing images are similar to training images.
  2. These methods require thousands of images for training.
  3. They require higher computational and storage resources at testing time.
- We propose a pixel-based ridge-regression (Pixel-RR) FES method in which each output pixel observes only one input pixel.
- Results demonstrate that Pixel-RR performs comparably well against state-of-the-art GANs on in-dataset images and significantly better on out-of-dataset images.
- Pixel-RR requires two order of magnitude fewer parameters compared to GAN-based FES models.

Pixel-based Regression

1. Pixel-based ridge regression (Pixel-RR)

$$E(w_p, b_p) = \frac{1}{2}\|w_p\mathbf{x}_p + b_p\mathbf{1} - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}(w_p^2 + b_p^2) \quad (1)$$

- Here scalars w_p and b_p are learnable weight and bias values.
- The unique global minimizers for Eq. 1 can be computed analytically as

$$\begin{bmatrix} w_p \\ b_p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_p\mathbf{x}_p^T + \lambda & \mathbf{1}\mathbf{x}_p^T \\ \mathbf{1}\mathbf{x}_p^T & N + \lambda \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{t}_p\mathbf{x}_p^T \\ \mathbf{t}_p\mathbf{1}^T \end{bmatrix} \quad (2)$$

2. Pixel-based kernel regression (Pixel-KR)

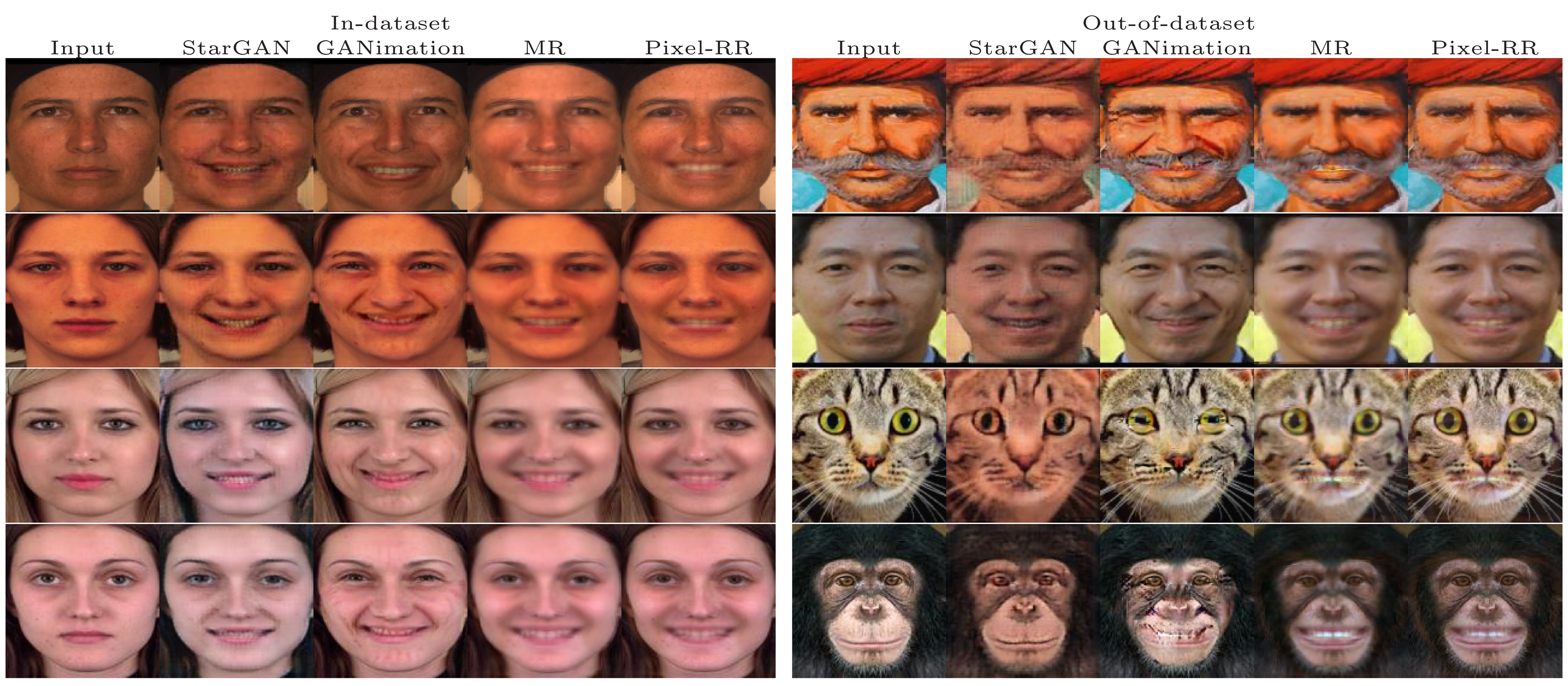
$$E(\mathbf{c}_p) = \frac{1}{2}\|\mathbf{c}_p\phi(\mathbf{x}_p)^T\phi(\mathbf{x}_p) - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}\|\mathbf{c}_p\phi(\mathbf{x}_p)^T\|_2^2 \quad (3)$$

$$= \frac{1}{2}\|\mathbf{c}_pK_p - \mathbf{t}_p\|_2^2 + \frac{\lambda}{2}\mathbf{c}_pK_p\mathbf{c}_p^T \quad (4)$$

- Here $K_p = \phi(\mathbf{x}_p)^T\phi(\mathbf{x}_p) \in \mathcal{R}^{N\times N}$ is the kernel matrix.
- The optimal projection matrix $\mathbf{c}_p$ can be computed as:

$$\mathbf{c}_p = \mathbf{t}_p(K_p + \lambda I)^{-1} \quad (5)$$

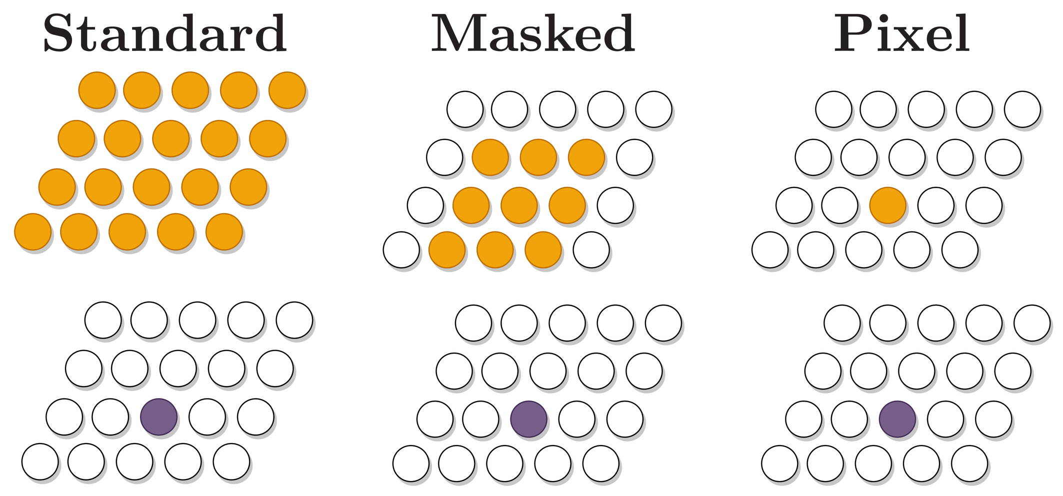
Qualitative Results



Motivation

- Recently, MR by Khan et al. [1] has shown that facial expressions usually constitute local instead of global changes in the input image.
- Motivated by this fact, our proposed method considers only one input pixel to produce an output pixel.
- To capture complex, non-linear characteristics of facial expression mappings, we also show how kernel regression can be exploited.

Receptive fields for Regression



Contributions

1. We have introduced the first pixel-based method to solve the FES problem.
2. Pixel-based idea can be extended using kernel regression.
3. The proposed method generalizes much better for a variety of out-of-dataset images.
4. The proposed model is two orders of magnitude smaller than GAN-based models.

Quantitative Results

Comparison of different FES models sizes

Parameters	$\times 10^4$
StarGAN [2]	850
GANimation [3]	850
MR [1]	16.2
Pixel-KR	655
Pixel-RR	**3.28**

User study to evaluate expressions

Model	Neutral $\rightarrow$ Happy
GANimation	26%
MR	17%
Pixel-RR	**57%**

Expression classification accuracy

Model	Accuracy
GANimation	68%
MR	84%
Pixel-RR	**85%**

References

[1] Nazar Khan et al. "Masked Linear Regression for Learning Local Receptive Fields for Facial Expression Synthesis". In: *International Journal of Computer Vision* 128.5 (2020), pp. 1433–1454.

[2] Yunjey Choi et al. "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8789–8797.

[3] Albert Pumarola et al. "GANimation: One-Shot Anatomically Consistent Facial Animation". In: *International Journal of Computer Vision* 128.3 (2020), pp. 698–713.