Warp-Guided GANs for Single-Photo Facial Animation

- PDF
- Authors
  - Jiahao Geng
  - Tianjia Shao
  - Youyi Zheng
  - Yanlin Weng
  - Kun Zhou
- **Abstract**
  - Inputs
    - A single target image
    - A set of facial landmarks derived from a driving source.
  - Output
    - An image or video of the target person moving according to the movement of the landmark source.
  - The landmarks are used to warp the target image via lightweight 2D warps.
  - Then, a conditional GAN is used to fuse fine facial details (e.g., creases and wrinkles) to the warped image.

Introduction

- The authors points out that previous works often requires a driving video or video of a target person. They are needed because for either:
  - 3D reconstruction; for example, [Breuer et al. 2008].
  - Borrowing unseen parts of fine-scale details; for example, [Averbuch-Elor et al. 2017]
- They also observe that good results for the portrait animation problem can be obtained by:
  1. Perform image warping with lightweight 2D warps first.
  2. Add details via techniques such as ERI [Liu et al. 2001].
- The paper's approach:
  1. Perform image warping with lightweight 2D warps first.
  2. But, add fine details with a generative model, which they use a conditional GAN.
- Insight: 2D warps take care of the hard nonlinear transformations, allowing the network to focus on appearance synthesis.
- The algorithm in more details.
  1. Control points in the source face and the target face are extracted.

    2. Displacements of control points are used to generate a global 2D warp, which is a per-pixel displacement map.
    3. The displacement map is applied to the target image, giving a warped face image.
    4. The displacement map and the warped face image is fed to a conditional GAN to generate the final image with details added.
- The conditional GAN, called *wg-GAN*, is trained on warped face images and displacement maps extracted from public available video datasets.
- wg-GAN, however, cannot generate hidden regions, so the authors train another network to inpaint them.
    - They use [Iizuka et al. 2017]
- Having the generative model has several benefits.
    - We can relax some requirements on the input: we do not need the full driving video, just only the landmarks.
    - We can eliminate the ad-hoc algorithm in the Averbuch-Elor et al. paper that is used to add fine-scale details.
    - Using GPU allows for real-time animation.

Background

- Animation by fitting models to photos.
    - [Blanz and Vetter 1999] is the first paper that fits a 3D morphable model to a photo. This is the paper that people always cite.
    - [Breuer et al. 2008] describes an automatic pipeline to fit such a 3D morphable model to images or video.
    - [Piotraschke and Blanz 2016] describes another fitting algorithm. It also points that reconstructing from a single image requires manual initialization.
    - [Vlasic et al. 2005] takes the approach of extract movement → render face mesh → merge back to target image to animate face.
- Many papers assume the availability of videos of the source and/or target persons.
    - [Theis et al. 2016]
        - Track face meshes in source and target videos.
        - Transfer expression by mesh deformation.
        - Use target video of fine the appropriate frame to copy the mouth interior for inpainting.
    - [Li et al. 2012] assumes a facial performance database of a target person and selects the appropriate frame to use for each frame in the output video.
- The following two papers are a little different because they deal with replacing the face in the target video with another face.
    - [Dale et al. 2011] replace the face in a target video with another face from the source video. Requires tracking facial performance in both videos with morphable models.

    - [Garrido et al. 2014] is another face transfer system, but it does not rely on 3D models like Dale et al. Frames from the source videos are retrieved, warped, and merged into the target video. All using 2D approaches.
- Papers with specific focuses.
    - Viewpoint manipulation: [Fied et al. 2016]
    - Gaze manipulation: [Kuster et al. 2012], [Ganin et al. 2016]
    - Transferring lip motion: [Garrido et al. 2015]
    - Filling in the inside of the mouth: [Blanz et al. 2003], [Kawai et al. 2014]
- Song et al. [2017] is a network based method for facial expression generation.
- There are a rich literature on facial performance tracking. (TODO: fill this)
- There are also rich literature on using neural network for animation. (TODO: fill this)

Overview

- Input
    - A single target portrait photo with the face in the neutral-frontal pose with the mouth closed.
    - A set of landmarks for the face, the head, and the upper body, indicating the desired pose+expression.
- Output
    - Another image the subject having another expression with the upper body taking the pose specified by the landmarks.
- The algorithm has two stages.
    1. The *global stage* where structural transformation are carried out by a lightweight 2D warp of the whole image.
    2. The *local stage* where:
        - A conditinal GAN adds fine-scale details and remove artifacts brought by the 2D warp.
        - An inpainting network hallucinates the inner mouth region.
- Insights
    1. The global lightweight 2D warps captures the structural change of the face well enough
    2. The 2D warp exempts the GAN from learning hard non-linear geometric transformation, allowing it to focus on filling the local, per-pixel details.

The Method

- The paper detects the facial landmarks and the peripheral landmarks in pretty much the same way that the Averbuch-Elor et al. paper does. See the notes on that paper for more details.
- However, the 2D landmarks are generated differently.
    - The Averbuch-Elor paper uses Dlib.
    - The present paper uses the DDE algorithm by Cao et al. [2014].

        - The DDE algorithm tracks 2D landmarks and use them to recovers 3D blendshapes, the expression, and the 3D pose of the face.
        - The paper applies the DDE algorithm to both the target image and the source image.
        - Then, it uses transfers the expression and the pose of the source image to the mesh of the target image.
        - The landmarks for the transformed target mesh is then projected to get the displaced 2D landmarks.
- Once the 2D landmarks have been computed, the paper applies the confidence-awere warping in the Averbuch-Elor et al. paper to the target image.
    - This gets us the *coarse* face-warped image.
- The coarse face-warped image is still lacking in fine-scale details such wrinkles and creases and the inner region of the mouth.
- The Averbuch-Elor et al. paper handled these problems by (1) copything the mouth from the source video, and (2) use ERI to add fine-scale details.
- The present paper observed that the approach would not work well if the source face is too difference from the target face. As a result, it uses neural networks to do these instead.

Refining Warped Face with wg-GAN

- The paper uses a typical conditional GAN approach.
    - The generator network G is called the *face refinement network*.
    - The discriminator is denoted by D. It has no special name.
- Goals for network design.
    1. G must not alter the position of the organs.
    2. G should be able to add details such as wrinkles, creases, and illumination condition without modifying the identity of the face.
    3. G should be fast enough for real-time use.
- Input to G
    - A cropped face region which is warped and without the inner of the mouth region.
        - The face is rectified before being fed to the network, as in [Song et al. 2017].
    - A displacement map that was usesd to warp the face.
- Output of G
    - Face image with full details except the inner of the mouth.
- Displacement map generation
    - For each landmark in the cropped warped image I_w, compute the offset from the rest pose image I.
    - The displacements of different landmarks are multiplied by different multiplicative factors to "normalize" them.
        - If not normalized differently, small displacements (for example, eyebrow displacements) will be ignored by the network because of

larger displacements (for example, mouth displacement).
        - To normalize the displacement, the paper computes the standard deviation of the displacements for each landmarks and uses it to divide the dispalacements so that the standard deviation become 1.
    - The normalized displacements are then used to fill the whole image through the traingulation that used to warp the face image.
- The generator network G.
    - Standard encoder-decoder architecture.
    - It only downsamples the image to 1/4 of the original size to prevent too much information loss.
    - The bottleneck is 4 residual blocks. [He et al. 2016]
    - The upsampling layers are resize-convolution layers instead of deconvolution layers.
        - This is done to avoid the "uneven overlap" problem [Gauthier 2014].
    - Skips connections are added between the first and last two convolutional layers to preserve the image structure. Isola et al. 2017
- The discriminator network D.
    - Takes in a real face image or the output of G and the displacement mask.
    - Compresses the image through 6 convolutional layers into a 32768-dimensional vector feature vector.
    - The feature vector is then passed to a fully connected layer to generate a confidence value of the image being real.
- Both G and D use leaky ReLU as nonlinearity, except:
    - The last layer of G, which uses a sigmoid to force the output values in the $[0, 1]$ range.
    - The last layer of D does not have a non-linearity.
- Loss function.
    - Let x_w denote the input warped image.
    - Let M denote the displacement map.
    - The output of the generator is thus $G(x_w, M)$.
    - Let x_g denote the ground truth.
    - For the discriminator, the paper uses the Wasserstein GAN loss (the original one in [Arjovsky et al. 2017]):

$$\mathcal{L}_D := E_{x_w, M, x_g}[-D(x_g, M) + D(G(x_w, M), M)].$$

    - For the generator, the paper combines the Wasserstein GAN loss with an L1 loss:

$$\mathcal{L}_G := E_{x_w, M, x_g}[\alpha \|G(x_w, M) - x_g\|_1 - D(G(x_w, M), M)].$$

    - In the above equation, α is a hyperparameter. It is set to 0.004 in the paper.
    - To improve training stability, the discriminator is updated using a history of generaotr outputs along with the ones in the current minibatch [Shrivastava et al. 2017].

- Training data.
  - Collected from video sequences.
  - Each video must begin with the face in rest expression.
  - The paper detects facial landmarks for every 10th frame and generate warped image for the frame. This gives you the two inputs to G.
  - The ground truth is the real frame.
  - Data are gathered from public datasets:
    - MMI Pantic et al. 2005, Vlasic and Pantic 2010
    - MUG Aifanti et al. 2010
    - CFD Ma et al. 2015
  - Data is augmented by random cropping and flipping.

Hidden Region Hallucination with hrh-GAN

- The "hidden region hallucination network" (hrh-GAN) fills in the inner region of the mouth.
- Input: a cropped face without the inner mouth region.
- Output: a complete face with teeth and tongue inside the mouth.
- Training data
  - Data derived from MMI, MUG, and CFD + portrait images from the Internet.
  - 6211 images in total.
  - The mark of the inner mouth region is computed with the detected landmarks.
  - Data is augmented with flipping and random cropping.
- Architecture is the same as that of the Iizuka et al. 2017 paper. See this note for more details.
- The loss function has:
  - A mean squared error (MSE) term.
  - A global GAN loss term.
  - A local GAN loss term.
- Alleviating data scarcity problem.
  - The training set size is much smaller than the one used by Iizuka et al.
  - To solve this problem, the paper uses the progressive GAN technique [Karras et al. 2017].
  - It trains the GAN from 128×128 resolution first then changes to 256×256.
    - At first, the network receives a $128 \times 128 \times 3$ image and converts to $128 \times 128 \times 64$ feature, which is then fed to the rest of the network. The paper trains with this to completion.
    - Then, it replaces the first layer with the following convolution train:

      $$256 \times 256 \times 3 \rightarrow 256 \times 256 \times 16 \rightarrow 256 \times 256 \times 32 \rightarrow 128 \times 128 \times 64$$

    - The upsampling layers are modified similarly.

Evaluation

- In general, the paper performs better than the Averbuch-Elor et al. paper.
    - This is especially true when the mouth region in the source is not similar to the target.
    - The paper's algorithm also does not depend on the resolution of the source video, so it can generate the inner mouth that is has more resolution.
    - Details addition is also better. The Averbuch-Elor et al. paper sometimes misclassify beards as details to add to the target image.
    - It also performs better in user study.
- When comparing with Theis et al., the paper is better at adding facial details because, I think, Theis et al. does not deal with this.
- When comparing with Song et al., the paper generates faces with fewer artifacts.
    - Song et al.'s network has a tougher task of warping the face and adding details while the task is factored into three in the paper's approach.