Bringing Portraits to Life

- PDF link
- This is a SIGGRAPH Asia 2017 paper.
- Authors
    - Hadar Averbuch-Elor
    - Daniel Cohen-Or
    - Johannes Kopf
    - Michael F. Cohen
- **Abstract**
    - Input
        - A driving face video
        - A *single* target face image
    - Output
        - The video of the person in the target image acting with the motion of the driving video.
    - Technique
        - Animate by 2D warps that imitate the facial movement in the source video.
        - Add fine-scale dynamic details for creases and wrinkles.
        - Hallucinate the inner of the mouth.

Introduction

- The selling point of this paper is that it uses only a single target face image.
    - Previous papers assume a video or a collection of images of the target face.
- The paper uses "lightweight" 2D warps to transform face.
    - No construction of 3D model.
    - What is possible are only moderate facial movements.
        - I guess this means frontal to profile/side view transform.
    - Correspondences are established by facial landmarks.
        - There's a reliance on facial landmark detectors.
- Unique features.
    - Adding details such as wrinkles and creases.
    - Hallucinating hidden areas, especially the inside of the mouth.
- Novel application: *reactive profiles*
    - Think of the moving portraits from Harry Potter.

Previous Works

- Some previous papers that also use "lightweight" 2D morphs.
  - Perspective-aware Manipulation of Portrait Photos

- Manipulate camera viewpoint from a single image.
- Real-time facial reenactment.
- Data-driven enhancement of facial attractiveness
- Expression flow for 3D-aware face component transfer

- Papers that create use 3D models for creating animation from human photos.
  - A morphable model for the synthesis of 3D faces
    - The first paper that fits a morphable model to photo.
    - The model can then be manipulated to change pose and appearance.
  - Reanimating Faces in Images and Video
    - Fits model to photo and then manipulate the mouth region.
  - Automatic 3D Face Reconstruction from Single Images or Video
    - Automatic pipeline for fitting morphable model to a single image.
  - Automated 3D Face Reconstruction from Multiple Images using Quality Measures
    - This one uses multiple images.
    - It observes that, if one image is used, manual initialization is needed.
- Papers that require a target video as input.
  - Video Face Replacement
  - Face transfer with multilinear models
    - Edit 3D mesh of generated from the target video.
  - A Data-driven Approach for Facial Expression Synthesis in Video
    - Use facial performance database to generate output video.
  - Being John Malkovich
    - Uses image search for animation.
    - Only works if the target person has many images or videos to search from.
  - VDub: Modifying Face Video of Actors forPlausible Visual Alignment to a Dubbed Audio Track
    - Generating mouth movement from speech.
- Works that involve non-human avatars.
  - Real-time avatar animation from a single image
    - Local PDF
    - I should read and cite this paper.
  - Mood Swings : Expressive Speech Animation
    - Extract expression from video and transfer to avatar.
- Sophisticated capture methods
  - Real-Time High-Fidelity Facial Performance Capture
  - Real-Time Facial Segmentation and Performance Capture from RGB Input
- Some other cited papars.
  - Semantic Facial Expression Editing using Autoencoded Flow
  - FaceWarehouse: a 3D Facial Expression Database for Visual Computing
    - The Averbuch-Elor et al. paper says that this paper introduces video-to-image retargetting.

- [Gaze Correction for Home Video Conferencing]
- [DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation]
- [Data-Driven Speech Animation Synthesis Focusing onRealistic Inside of the Mouth]
- [Automatic Cinemagraph Portraits]
    - Remove camera shake and large movements to create relatively still moving portraits.
- [Facial Expression Editing in Video Using a Temporally-Smooth Factorization]
    - Exaggerating and attenuating expressions in some parts of a video.

Overview

- The algorithm
    1. The paper first extracts landmarks in the source video and the target images.
        - There are two types of landmarks:
            - Face landmarks
            - Non-face landmarks → in order to animate the whole moving head.
    2. After landmarks are extract, correspondences between source frames and target image are established.
        - A correspondence map is a vector field that tells that tells, for each pixel in the warped image, where in the original image should come from.
        - Hence, correspondence spans over the entire image, not just the face.
        - The paper expands correspondence from pairs of landmark positions to a vector field over the image.
        - When generating correspondence, the paper treat high-confidence regions (i.e., around face landmarks) differently from low-confidence regions (i.e., any other areas),
        - The paper also maintains correspondence over time.
    3. 2D warps are then generated from the correspondence and applied to the target image.
        - The output of this step is called the **coarse target video**.
    4. Hidden regions and fine-scale details are then added to the warped image.
        - Inner mouth region and wrinkles are transferred from the source video to the target video.

Coarse Target Video Generation

- Inputs
    - Target image t^*.
    - Driving video S.
        - The ith frame is called s_i.

- Output
  - A video T which maintains the identity of t^* but has the movement of S.
    - The ith frame is denoted by t_i.
- Assumptions
  - t^* has a neutral expression.
    - Mouth is closed.
  - S has a frame s^* with neutral expression.
    - This is generally assumed to be the first frame s_0, but this can be changed by manual choosing.
    - Since s_0 is defined in such a way, we shall assume that the 0th frame of the output video would be t^* itself. In other words, $t_0 = t^*$.
- The aligning transformation ϕ
  - t^* is not aligned to s^*.
  - So the paper generates a transformation ϕ that compensates for it.
  - This is done by first using [Dlib] to detect 68 facial landmarks in the two images.
  - The paper then finds a rotate-and-scale transformation (no translation?) that minimizes the square distance between the landmarks in the eye regions and the tip of the nose.
  - From reading the paper, it seems that ϕ aligns the s^* to t^*, not the other way around.
- For each source video frame and the target image, we extract landmarks.
  - Landmark positions are denoted by bold letter $\mathbf{p}$.
    - $\mathbf{p}_i^s$ = landmarks in the ith source frame.
    - $\mathbf{p}_i^s$ = landmarks in the ith target frame (which is the output).
  - There are two types of landmarks.
    - Face landmarks = the 68 face landmarks computed by [Dlib].
      - There's a paper for this if you want to cite it. [Link]
    - Peripheral landmarks.
      - The points are obtained by two means.
        - Tracking points in the source video. The paper uses a simple optical flow tracking algorithm. [Link]
        - Points on the image boundary that do not move throughout the video.
      - Note that the method above only generate peripheral landmarks in the frames of S. It does not generate peripheral landmarks in t^*.
      - To generate peripheral landmarks in t^*, the then *hallucianate* them by applying ϕ to the peripheral landmarks in a source target frame.
        - It is not clear whether ϕ is applied to $\mathbf{p}_0^s$ or $\mathbf{p}_i^s$ for each i separately.
        - IMHO, $\mathbf{p}_0^s$ makes more sense.

- The landmark positions in the target frame p_i^t are computed as follows.
  - We just previously discussed how to compute p_0^t, which are landmarks in the 0th frame.
  - For other frames, the paper computes:

$$\mathrm{p}_i^t = \mathrm{p}_0^t + \phi \cdot (\mathrm{p}_i^s - \mathrm{p}_0^s).$$

  - $p_i^s - p_0^s$ is the offset of the landmarks in the source frames relative to the neutral frame.
- Dense warp field computation
  - The paper computes a Delaunay triangulation using p_0^s.
  - The topology can then be imposed on p_0^t and p_i^t for all i.
  - By moving the points from p_0^t to p_i^t and moving the pixels inside the triangles as if the pixels are texture mapped onto the triangles, we have created a piecewise linear warp field from the original target image $t_0 = t^*$ to the target frame t_i.
- Confidence-based warping
  - The authors observed that the warping above works well only in the face region where the landmarks are reliable.
  - Outside the regions, weird things such as straight lines in the background might be warped incorrectly.
  - To alleviate this, the paper convolves the warp field with a blurring kernel whose radius increases away from the face region.
    - This blurs the warp field as we move away from the face region.
  - The paper uses blurring kernels with 10 radius values in the range $[0, 0.05 \times S_{diag}]$ where S_{diag} is the size of the image diagonal.

Transferring Mouth Interior

- Algorithm
  1. The paper first aligns the frame s_i with t_i using the warping procedure in the previous section.
  2. Then, it crops the mouth interior region from the source frame.
  3. To make sure that the crop strictly involves the inside of the mouth, the paper erodes by radius $0.1 \times h_{mouth}$ where h_{mouth} is the height (in pixels) of the mouth in the target frame.
  4. The crop is then alpha blended into the target frame.
  5. Poisson blending is then applied to merge the crop to the target frame.
- The mouth interior is transferred only when the mouth size in s_i is significantly bigger than in t^*.
  - Let a_{mouth}^* be the area (in pixels) of the mouth interior in t^*.
  - Let a_{mouth}^i be the same for t_i.
  - The mouth interior is transferred only when $a_{mouth}^i > 2a_{mouth}^*$.

- To make the change smooth, the paper linearly blends between the two mouths when the size is in the range $[a^*_{mouth}, 2a^*_{mouth}]$.

Transferring Fine-Scale Details

- The fine details considered include:
  - Shading changes induded by wrinkles around the eyes when smiling.
  - Creases alongside a smiling mouth.
- The transfer is based on the technique by Liu et al. [2001]
  - The paper deals with transferring appearance changes due to a change in a person's expression.
  - We have a neural face of a person in an image I_a.
  - To change the person's expression, we warp the face to obtain another image I_b with the facial parts in the right positions.
    - However, geometric warping does not include the fine-detail changes above.
  - We then need another image $\tilde{I}_a$, which will donate the appearance to I_b.
  - We first warp $\tilde{I}_a$ so that it aligns with I_a.
  - Then, we compute the *expression ratio image* (ERI):

$$R = \frac{f(\tilde{I}_a)}{f(I_a)}$$

    where f is some generic function such as computing the luminance channel.
  - The warped image I_b is then transformed to $\tilde{I}_b$ according to

$$\tilde{I}_b = R \times I_b.$$

    Here, R has been warped so that it now aligns with I_b.
- In the context of the Averbuch-Elor et al. paper, we have that:
  - The neural source frame s is I_a.
  - The frame s_i is the image $\tilde{I}_a$, which would donate the appearance.
  - The target frame t_i is the warped image I_b.
- The authors observed that applying the ERI everywhere causes the following problems.
  1. Certain areas of the resulting image may become saturated.
  2. The resulting image can include outliers.
     - Inappropriate shadowing caused by the nose or other misalignments.
  3. Temporal instability.
  4. Artifacts may appear in the eye or the background.
- The authors then discusses how to deal with these problems.
  1. For the saturation problem, they tune down pixels with $R > 1$ by multiplying it with a constant factor of 0.01.
  2. For the artifact problems outside the face, the paper estimate the face region and only apply the ERI there. The region is estimated by the following

two-step process.
    - Fitting an ellipse to the points along the chin.
    - Use the ellipse as an initial estimate for the Grab-Cut optimization to find a more accurate face region.
3. For the outlier problem, the paper detects and removes them.
    - The paper does not perform outlier detection for each frame separately.
    - Instead, it performs outlier detection in a *reference frame* s_{ref}, which is the most different from the neutral frame s^*.
        - To find the reference frame, the paper computer computes a transformation ϕ_i (consisting of a scale, a rotation, and perhaps a translation) that minimizes the distance between $\phi_i(\mathrm{p}_i^s)$ and p_0^s.
        - The paper then computes the L2 difference between $\phi_i(\mathrm{p}_i^s)$ and p_0^s.
        - The reference frame is the frame where the L2 difference above is maximum.
    - Once the reference frame has been determined, the paper detects outliers in it.
        - The paper first identify pixels with significant expression ratio values.
            - Significant values are those that are less than $1/1.1$ and more than 1.1.
        - The paper would then find connected components of these significant pixels (based on the 8 neighbors of each pixel).
        - For each pixel in the component:
            - The algorithm considers the 20×20 region around it.
            - For each pixel in this region, it considers the 3×3 neighborhood of each pixel and compute the maximum RGB pixel difference in the neighborhood.
            - Then, it computes the minimum of the maxima above in the 20×20 region.
        - With the above step, we have a minimum of maximum RGB pixel difference at each pixel in the component.
        - The algorithm then finds the average of the minima over the component.
        - If the average of the minima is less than 5, then the algorithm considers the compontent an outlier.
    - The footprints of the outlier components are then propagated to the other frames in the output video.
    - For the pixels close enough to the outlier component footprints (within 20 pixels radius), the ERI is set to 1, nullifying the expression ratio's effects.

- To increase the temporal stability of the ERI, the paper Gaussian blurs the aligned ERIs temporally over 21 frames.
- This is a convoluted and poorly described algorithm.