Literature Review 1

Vishal Gadiraju 01672049

Automatic Cinemagraph Portraits

(Secondary Paper)

Abstract:

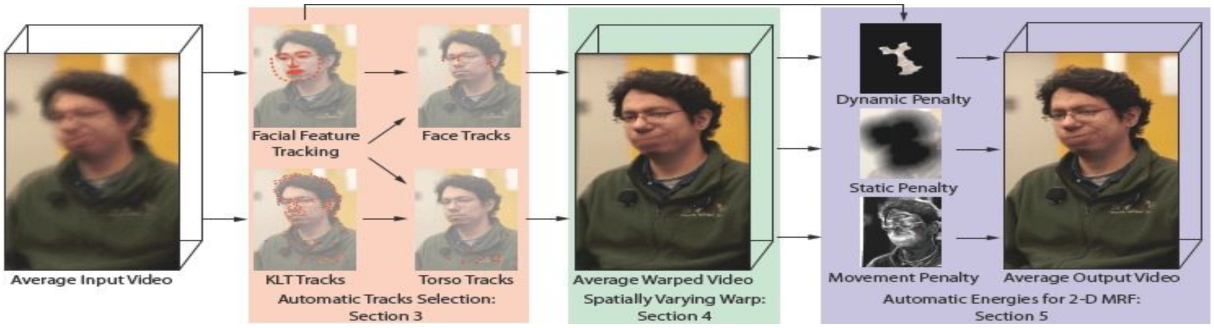
Cinemagraphs are a popular new type of visual media that lie in-between photos and video; some parts of the frame are animated and loop seamlessly, while other parts of the frame remain completely still. Cinemagraphs are especially effective for portraits because they capture the nuances of our dynamic facial expressions. We present a completely automatic algorithm for generating portrait cinemagraphs from a short video captured with a hand-held camera. Our algorithm uses a combination of face tracking and point tracking to segment face motions into two classes: gross, large-scale motions that should be removed from the video, and dynamic facial expressions that should be preserved.

This segmentation informs a spatially-varying warp that removes the large-scale motion, and a graph-cut segmentation of the frame into dynamic and still regions that preserves the finer-scale facial expression motions. We demonstrate the success of our method with a variety of results and a comparison to previous work.

Summary:

Cinemagraphs combine benefits of static images and videos. Portrait Cinemagraphs are hard to create and usually take a lot of time for them to be created. This paper creates a fully automated pipeline for portrait cinemagraphs.Most parts of the resulting cinemagraph are static whereas some animate in a seamless loop. The approach described in this paper uses Face-tracking and Kanade-Lucas-Tomasi to segment motion into two different classes i.e large-scale head torso motion which is to be removed and fine-scale facial expression motion which needed to be preserved.

Large-scale information is used to guide spatially varying warp which removes gross motion of portrait subject. 2D-graph cut technique is also used to compute a mask that combines dynamic facial expressions with still portrait. 2D approach is much faster than the 3D approach that was used in the earlier papers.
There were several assumptions about the input video. First that there is only one person with one facial pose in the video.Second, the face of the person must occupy a large portion of the image anda good rule of the face can be aligned using a 2D warp, meaning that there should not be large 3D rotations.



The figure above visualizes automatically generated portrait cinemagraph from an input video. Face and torso appear blurry as it a video i.e in motion. The average output video is blurry for facial parts that are animated and sharp for static regions.

Facial Movements Detection technique allows the algorithm to identify the location of the face and also to determine the movement of the major facial regions in the face. In this paper, four major facial regions; the eyes, eyebrows, mouth and lower jaw are considered. The mouth is further subdivided into three sub-regions; the left and right-most region of the lip and the bottom lip.

The location of each facial region is the center of its corresponding facial feature points. The position of each facial region is monitored across time by using a coordinate transform. Perpendicular distance from one of two local face orientation axes to detect finer-scale motions is also measured.

Similar technique is used to stabilize the torso, but center of the torso is set as 3r below the face and only consider tracks within a radius of 4r of that center. No torso-specific track selection is performed, but any tracks removed during facial track selection are not used for the torso even if they fall within the torso radius.

The key to this algorithm is the selection of KLT Tracks which lie on static regions allowing to immobilize the face after warping. Automatic energy values are also computed for graph-cut optimization. It composites the warped video with still image to create the final cinemagraphs.

Overall, 15 portrait videos were captured to produce cinemagraphs that demonstrated wide range of results. On comparisons with automated cinemagraphs against a user-directed cinemagraph, resulting video is on par with even as fewer tracks were selected. The compositing seam also produces similar results even as this method uses only 2D matte. This method produces significant results as far as timing is concerned. It produces shorter time i.e faster in all pipelines in comparison to all the previous works. For track-selection it averages 0.70 seconds and 41 seconds for compositing while the average compositing time in previous work was about 600 seconds.

However, the cases where there is large rotation of face could not handled since spatially-varying warp is used to align faces. This method heavily relies on accurate Facial Feature Tracker and its failure in some cases leads to negative results. It also fails when the facial deformation is too large and too fast.

A survey was also conducted on Amazon Mechanical Turk in order to understand how the general population would respond to portrait cinemagraphs. 7 sets of results (30 HITs, each HIT has 7 sets) were showed to 30 users. For each set, the still image, the input video and our cinemagraph result side-by-side and asked them to pick their favorite medium to represent each person. A total of 218 total votes were recorded and 80.3% of them were in favor of dynamic portraits (either in-put video or cinemagraph). Cinemagraphs gained 53% of the votes whereas only 19.7% were for a still image. The collected votes yield $\chi2$= 40.5229 when computed against an expected uniform distribution and exceeds the $\chi2$ value of 13.82 for p = 0.001 with 2 degrees of freedom. The survey supports the hypothesis that people prefer cinemagraphs for portraits.

The future work would be to automatically identify composite moving regions such as hair or cloth which are difficult to detect when in motion because of the ambiguity between background color and such complex moving regions.

Bringing Portraits to Life

(Primary Paper)

Abstract:

We present a technique to automatically animate a still portrait, making it possible for the subject in the photo to come to life and express various emotions.

We use a driving video (of a different subject) and develop means to transfer the expressiveness of the subject in the driving video to the target portrait. In contrast to previous work that requires an input video of the target face to reenact a facial performance, our technique uses only a *single* target image.

We animate the target image through 2D warps that imitate the facial transformations in the driving video. As warps alone do not carry the full expressiveness of the face, we add fine-scale dynamic details which are commonly associated with facial expressions such as creases and wrinkles.

Furthermore, we hallucinate regions that are hidden in the input target face, most notably in the inner mouth. Our technique gives rise to *reactive profiles*, where people in still images can automatically interact with their viewers. We demonstrate our technique operating on numerous still portraits from the internet.

Summary:

Facial expressions in humans convey major emotions and also give a deeper view into the emotional state of a person.This paper demonstrates techniques that produce animating faces in human portraits, and in particular controlling their expressions. Previously, facial animation techniques usually assumed the availability of a video of the target face exhibiting variation in both pose and expression. In contrast to previous work, the method described in this paper, uses only a single image of a target face to animate it. Animation of the single target face image from a driving video is done, allowing the target image to come alive and mimic the expressiveness of the subject in the driving video. While most previous work restrict themselves to only the face region, within limits, this method animates the full head and upper body. The target image is animated by a series of warps that imitate the facial transformations in the driving video. Like in previous works (e.g., [Fried et al. 2016; Leyvand et al. 2008; Yang et al. 2011]), the face is manipulated by lightweight 2D warps.

animate the target image through a series of 2D warps that imitate the facial transformations in the driving video.

The warps are controlled by a set of sparse correspondences between the target face and the face in the driving video frames.
There are methods that have been proposed for detecting fiducial points on faces at fixed standard locations but, to bring the full head of the target image to life, the entire head is allowed to move and change its pose. This technique also tracks points other than the face region to help guide the overall head. Since geometric warps alone cannot encode the full range of changes a face undergoes when the expression changes, other fine-scale changes, such as self-shadowing in wrinkles and creases help in conveying the complete expression of the face. Implicitly they have assumed closed mouth of the target face which requires the need for hallucinating the appearance, if the mouth opens in the driving video.
The main technical solutions described are:
Correspondence expansion: utilizes the high-fidelity of facial landmarks detection and tracking, and expand the correspondences in the facial region to correspondences that span the entire image and over time. These points allow tracking and changing the pose of the target head to follow and imitate the one in the driving video.

Confidence-aware warping: The sparse set of correspondences are extrapolated to a dense vector field over the entire.The highly-confident facial region and the rest of the image are distinguished, where there is no guarantee on the quality or quantity of corresponding points, and also smooth the vector field accordingly.

Hidden region transfer: When needed like in the case of open mouth in the driving video, the mouth interior is transferred to the animated target frame. The composite retains most of the details possible from the target image as only the mouth interior are transferred to the animated frame.

Detection of inlier wrinkles: Creases and wrinkles are transferred in the facial region to generate a realistic expression. Illumination changes caused by cast shadows or misalignments are detected between the warped video frames.
This technique was demonstrated on challenging casually-captured internet portraits like that of Monalisa. To drive them, videos from the MMI Facial Expression Database and some of their own videos were used.The limitations of this study were, significant changes to the head pose could not be made as only the visual information in the single target image was available and also it was assumed that the target image contains a neutral face, and when it was violated the output frames are not realistic.

Relation between two papers:

The secondary paper edits a facial video performance in an expression preserving manner, removing undesired large-scale motion, whereas the primary paper describes and demonstrates techniques that create new photo-realistic expressions which are significantly different from the input image.

References:

Primary Paper : Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F.Cohen. 2017. Bringing Portraits to Life. ACM Trans. Graph. 36, 4, Article 196 (November 2017)

Secondary Paper : Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. 2013. Automatic cinemagraph portraits. In Computer Graphics Forum, Vol. 32. Wiley Online Library, 17–25.