First Order Motion Model for Image Animation

Aliaksandr Siaorhin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, Nicu Sebe

https://aliaksandrsiarohin.github.io/first-order-model-website/

https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf

TL;DR

The paper's framework animates objects in a source image without using any annotation or prior information about the specific object. Once trained on a set of videos depicting objects of the same category, the method can be applied to any object of the class. Appearance and motion information is decoupled. A generator network models occlusions that arise in target motions and combines the appearance extracted from the source image and the motion derived from a driving video.

Introduction

Image animation refers to the task of automatically synthesizing videos by combining the appearance extracted from a source image with motion patterns derived from a driving video. GANs and VAEs have been used to transfer facial expressions or motion patterns between human subjects in videos. But, these approaches usually rely on pre trained models to extract specific representations. First-order motion, proposes using a set of self-learned keypoints together with local affine transformations to model complex motions. It also introduces an occlusion-aware generator, which adopts an occlusion mask that indicates objects parts that are not visible in the source image, to be inferred.

Method

We are interested in animating an object depicted in a source image S based on the motion of a similar object in a driving video D. Training is self-supervised. For training, a large collection of video sequences containing objects of the same object category is used. The model is trained to reconstruct the training videos by combining a single frame and a learned latent representation of the motion in the video.

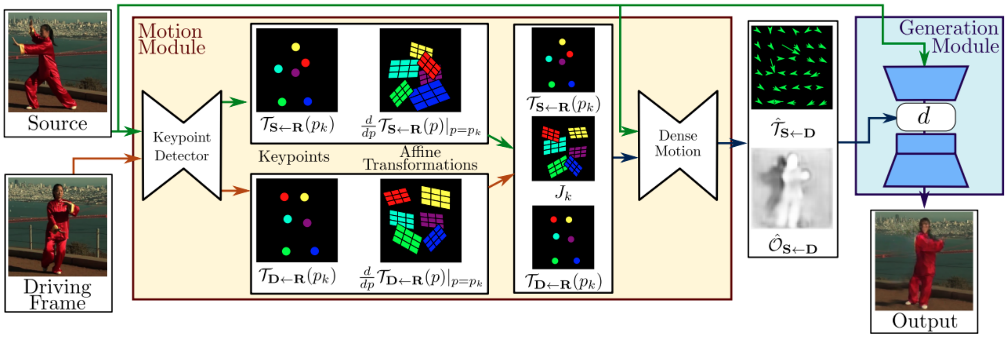
Source
Driving Frame
Motion Module
Keypoint Detector
$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
Keypoints
Affine Transformations
$\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)$
$\frac{d}{dp}\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p)|_{p=p_k}$
$\mathcal{T}_{\mathbf{S}\leftarrow\mathbf{R}}(p_k)$
J_k
$\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}(p_k)$
Dense Motion
$\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$
$\hat{\mathcal{O}}_{\mathbf{S}\leftarrow\mathbf{D}}$
Generation Module
d
Output

Figure 2: Overview of our approach. Our method assumes a source image $\mathbf{S}$ and a frame of a driving video frame $\mathbf{D}$ as inputs. The unsupervised keypoint detector extracts first order motion representation consisting of sparse keypoints and local affine transformations with respect to the reference frame $\mathbf{R}$. The dense motion network uses the motion representation to generate dense optical flow $\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$ from $\mathbf{D}$ to $\mathbf{S}$ and occlusion map $\hat{\mathcal{O}}_{\mathbf{S}\leftarrow\mathbf{D}}$. The source image and the outputs of the dense motion network are used by the generator to render the target image.

In the first step, transformations are approximated by using keypoints learned in a self-supervised way. The locations of the keypoints in D and S are separately predicted by an encoder-decoder network.

During the second step, a dense motion network combines the local approximations to obtain a dense motion field. The network also outputs an occlusion mask that indicates which image parts of D can be reconstructed by warping the source image and which parts should be inferred.

Finally, the generation module renders an image of the source object moving.

Experiments

First order motion was able to generate videos of much higher resolution compared to results in the paper *Animating arbitrary objects via deep motion transfer* in all experiments.