Image Animation with Keypoint Mask

Or Toledano [1] **Yanir Marmor** [1] **Dov Gertz** [1]

Abstract

TODO OR important transgaga structure realtime (Siarohin et al., 2020)

1. Introduction

TODO Or

Our paper focuses on the motion transfer problem: given a source image S and a driving video D, the goal is to syntesize a video with the identity of S, and the motion from D. Some notable works (Siarohin et al., 2020), (Wiles et al., 2018), (Siarohin et al., 2019).

Our method does not rely on GANs - see Section 2.

Related Work: Our work doesn't rely directly on a strong motion prior, but uses a structure mask which was extracted from a keypoint detector of a motion based model, such as (Siarohin et al., 2020). The concept of using drawn keypoints as a geometry represantation (structural mask) was already used in the context of image-to-image translation, in works such as TransGaGa (Wu et al., 2019). The concept of using a structural mask in the context of image animation is demonstrated in (Shalev & Wolf, 2020). However, the current work differs by basing the mask off a motion related module, which saves us the hassle of perturbing the input hoping to achieve an identy-less mask. By doing so, we improve their results, and purposes an additional "circles" only mask which can be used in the context of relative motion transfer during animation, as in (Siarohin et al., 2020), which isn't possible with a mask.

2. Methodology

Methods methods methods

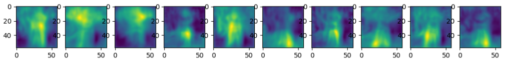


Figure 1. K channels of the keypoint detector network used in (Siarohin et al., 2020), before the softmax activation. Our main motion prior in this project.

3. Experiments

3.1. Datasets

The training and evaluation were done using Tai-Chi-HD dataset which containing short videos of people doing tai-chi exercises. Following (Siarohin et al., 2020), 3,141 tai-chi videos were downloaded from YouTube. The videos were cropped and resized to a resolution of 256^2, while preserving the aspect ratio. There are 3,016 training videos and 125 test videos.

3.2. Comparison with Previous Works

In order to compere our work to previous works(Table 2) we used metrics previously used in similar papers. Average Key-points Distance (Cao et al., 2017) (AKD) measures the average key-points distance between the generated video and the source video. Average Euclidean Distance (Zheng et al., 2019) (AED) measures the average euclidean distance between the representations of the ground-truth and generated videos in some embedding space. In addition, we added the L1 distance as well. The AED and AKD metrics were calculated using the following github: https://github.com/AliaksandrSiarohin/pose-evaluation.

TODO Dov

[1] Tel Aviv University. Correspondence to: Or Toledano <or-toledano@protonmail.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

Table 1. Images comparison

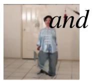
Source image	Driving				
X2Face					
Monkey-Net					
FOMM					
Yoav's work					
Ours					

Table 2. Accuracy Metrics

Method	AKD	AED	L1
Monkey-Net	10.798	0.228	0.077
FOMM	6.872	0.167	0.063
Yoav's work	4.239	0.147	0.047
Ours softmax	14.760	0.245	×
Ours	5.551	0.141	0.045
Improvement (FOMM)	19.2%	15.5%	28.5%

Software and Data

Detailed in our repository: https://github.com/or-toledano/animation-with-keypoint-mask

References

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shalev, Y. and Wolf, L. Image animation with perturbed masks, 2020.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. Animating arbitrary objects via deep motion transfer, 2019.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. First order motion model for image animation, 2020.

Wiles, O., Koepke, A. S., and Zisserman, A. X2face: A network for controlling face generation by using images, audio, and pose codes, 2018.

Wu, W., Cao, K., Li, C., Qian, C., and Loy, C. C. Transgaga: Geometry-aware unsupervised image-to-image translation, 2019.

Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., and Kautz, J. Joint discriminative and generative learning for person re-identification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.