Image Animation with Keypoint Mask

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

Abstract

Image animation is a task of synthesizing future video frames of a single source image by the motion from another video, the driving video. This task is challenging due to the complexity of motion representation and the unknown relations between the driving video and the source image. Despite its difficulty, this task has attracted great interests in machine learning at the recent years, with gradual improvements. The problem can be thought as decoupling of motion and appearance, which is often solved by extracting the motion from keypoint movement. In this work, we extract the structure by a keypoint heatmap map, without an explicit motion representation. Then, the structures from the image and the video are extracted to warp the image according to the video, by a deep generator. Our approach outperforms the state of the art in popular reconstruction benchmarks, and an improvement can be easily observed in animating videos. (Siarohin et al., 2020)

1. Introduction

Content that created in that task, known also as "deepfake". The word**deepfake* is a combination of the words "deep learning" and "fake" and regards genrally to "an image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said" (merriam-webster, "deepfake").

Although usually, this technology involve the generation and manipulation of human imagery, it's not necessary.

(Mirsky & Lee, 2021) in their survey about Creation and Detection of Deepfakes split this area into four categories of tasks: reenactment, replacement, editing, and synthesis.

our project focus involved at the "replacement" category.

The main challenges we need to deal with, as they defined in that survey are: paired training, identity leakage, occlusions and temporal coherence,

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.

our contribution it twofold: first, ignore the motion prior and better accuracy results, and second compact representation with better performance.

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.

1.1. Keypoints

In contrast to pixel-based approaches that were prevalent until a few years ago. Recently, keypoints-based approaches have been perceived as having the potential to achieve high performance in the field of video prediction. (relevant cites for that: (Kim et al., 2019), (Balakrishnan et al., 2018) (Ma et al., 2017) (Reed et al., 2017) (Chan et al., 2019) (Villegas et al., 2017) (Cai et al., 2018) (Wang et al., 2018) (Reed et al., 2015)

However, these works require frame-by-frame keypoints labeling, which limits the applicability of the methods. There are several of solutions for this problem. Basically, we need to use a pre-trained keypoints detector for our model.

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

(relevant cites for that:)

2. Methodology

Methods methods methods

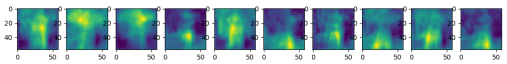

Figure 1. K channels of the keypoint detector network used in (Siarohin et al., 2020), before the softmax activation. Our main motion prior in this project.

Table 1. Images comparison

SOURCE IMAGE	DRIVING				
					
X2FACE					
MONKEY-NET					
FOMM					
PERTURBED MASK					
OURS					

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.

Table 2. Accuracy Metrics

METHOD	AKD	AED	L1
MONKEY-NET	10.798	0.228	0.077
FOMM	6.872	0.167	0.063
PERTURBED MASK	4.239	0.147	0.047
OURS SOFTMAX	14.760	0.245	0.077
OURS	5.551	0.141	0.045
IMPROVEMENT (FOMM)	19.2%	15.5%	28.5%

4. Conclusions

We constructed a novel method for image animation by moving the need for a strong motion prior (optical flow) to the assumption of a pre-trained keypoint detector/ keypoint heatmaps prior to activation. By doing so, we encapsulated motion to a motion mask, which is bottlenecked by the prior training which has the keypoint bottleneck. The motion masks are then fed into a generator, which combines the appearance of the source image and the mask which represents the structure, decoupled from any appearance naturally by the assumption that during the training of the keypoint detector, the heatmap mask went into a keypoint bottleneck. After evaluation, we can conclude that our method is competitive with the state of the art, and outperformes them in many cases.

Software and Data

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

References

Balakrishnan, G., Zhao, A., Dalca, A. V., Durand, F., and Guttag, J. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8340–8348, 2018.

Cai, H., Bai, C., Tai, Y.-W., and Tang, C.-K. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 366–382, 2018.

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.

Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5933–5942, 2019.

Kim, Y., Nam, S., Cho, I., and Kim, S. J. Unsupervised keypoint learning for guiding class-conditional video prediction. *arXiv preprint arXiv:1910.02027*, 2019.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Van Gool, L. Pose guided person image generation. *arXiv preprint arXiv:1705.09368*, 2017.

Mirsky, Y. and Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1), January 2021. ISSN 0360-0300. doi: 10.1145/3425780. URL https://doi.org/10.1145/3425780.

Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S. G., Wang, Z., Chen, Y., Belov, D., and Freitas, N. Parallel multiscale autoregressive density estimation. In *International Conference on Machine Learning*, pp. 2912–2921. PMLR, 2017.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. *Advances in neural information processing systems*, 28:1252–1260, 2015.

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.

Villegas, R., Yang, J., Zou, Y., Sohn, S., Lin, X., and Lee, H. Learning to generate long-term future via hierarchical prediction. In *international conference on machine learning*, pp. 3560–3569. PMLR, 2017.

Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., and Sebe, N. Every smile is unique: Landmark-guided diverse smile generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7083–7092, 2018.

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.