Image Animation with Keypoint Mask

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

Abstract

Image animation is the task of synthesizing future video frames of a single source image according to the motion from a 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 this difficulty, this problem attracted great interests from researches 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 from a keypoint heatmap, 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. It is generic, unsupervised and can be applied to animation of any arbitrary object, without any domain specific model for the structure of the input.

1. Introduction

(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.

TODO: didn't you mean synthesis?

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 synthesize 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 2.1.

our contribution it twofold: first, show that reducing the motion prior can work and yield 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 representation (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 identity-less mask which might not even be optimal. That way, we can base more of the motion representation on the deep network, reduce our prior, and leave some space for our network to achieve better results.

1.1. Keypoints

In contrast to pixel-based approaches that were prevalent until a few years ago, recently, keypoint-based approaches have been perceived as having the potential to achieve high performance in the field of video prediction. (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. (Siarohin et al., 2019) (Thewlis et al., 2017) (Zhang et al., 2018) (Jakab et al., 2018) (Newell et al., 2016)

2. Methodology

The network can be split to to parts: obtaining the mask, and generating the synthesized frame. The generator archi-

[1]Tel Aviv University. Correspondence to: Or Toledano <ortoledano@protonmail.com>, Yanir Marmor <yanirmr@gmail.com>, Dov Gertz <dovgertz1@gmail.com>.

tecture is constant amongst all of our variations, and can be described as low resolution generation from the source image, source mask and driving mask; Followed by upscaling of the low scale synthesized prediction by passing it with the source frame to a high resolution generator. We created two different versions for the mask generator, and chose to show results for the first, absolute one, due to significant performance. The performance drop for our second, relative version was expected because the mask contained less structural information.

2.1. Architecture

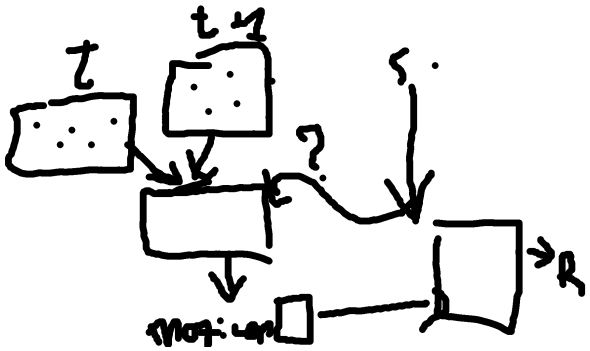

Figure 1. Architecture (TODO: someone create a real graphic out of this)

Our architecture follows (Siarohin et al., 2020) without the dense motion module, after changing its keypoint generation module to return our mask. After the mask is obtained, we follow the encoder decoder approach as (Shalev & Wolf, 2020). Namely, the encoder of the low resolution generator consists of $conv_{7\times7}$, *batch_norm*, *relu*, followed by six residual blocks of *batch_norm*, *relu*, $conv_{3\times3}$,*batch_norm*, *relu*, $conv_{3\times3}$, (and a sum with the source). The residual blocks help to maintain the identity of the source image (He et al., 2015). The decoder consists of two blocks, each is a sequence of $up\_sample_{2\times2}$, *batch_norm*, *relu*. The decoder is followed by a $conv_{7\times7}$ and a *sigmoid* activation. For the high resolution generator, use an encoder (decoder) with five encoding (decoding) blocks, where each block is a sequence of $conv_{3\times3}$, *batch_norm*, *relu* $avg\_pool_{2\times2}$, and each decoding block is a sequence of $up\_sample_{2\times2}$, $conv_{3\times3}$, *batch_norm*, *relu*. We add skip connections from each of the encoding layers to its corresponding encoding layer, to form a U-Net architecture (Ronneberger et al., 2015).

2.2. First mask version for absolute motion transfer, with warping

Our main mask for the project is obtained from carefully observing the keypoint module from (Siarohin et al., 2020). U-Net based keypoint modules of this form, work by extracting features, which pass through a *conv* layer to form a heatmap image K channels, where K is the number of keypoints. Then, a softmax is performed over the channels, and keypoints are extracted from the mean location of each heatmap channel. By carefully debugging the code, and the expectation for a segmentation map out of U-Net, we obtain:

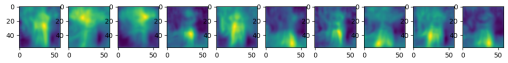

Figure 2. K channels of the keypoint detector network used in (Siarohin et al., 2020), before the softmax activation. Our main motion prior in this project.

Summing over the channels, we get

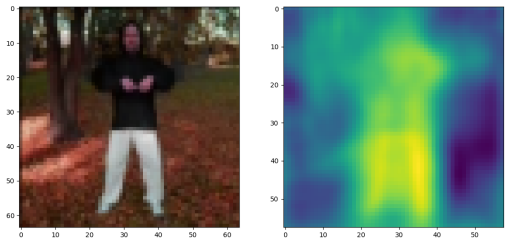

Figure 3. The sum of the K channels which is fed as a structural mask into the generator.

which is our output mask.

2.3. Second mask version for relative motion transfer

We 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 the previous heatmap mask. The mask catches the ge-

ometry representation (Wu et al., 2019) of the image, and by forcing it to be described as keypoints with a center, we can use the relative coordinates for the animation. However, this module didn't perform as well as our heatmap mask module (2). Relative motion transfer isn't always the wanted outcome, but this work indicates that a keypoint-only-prior based module is feasible for the task. It can be expected that we will get results which are more similar to (Siarohin et al., 2019), due to the fact that the only information that the pair of masks contain is the keypoint displacement, so our deep network can only try its best to simulate a zero order approximation.

By taking the softmax over the heatmap, we get:

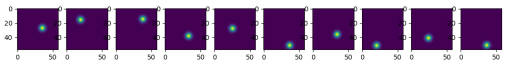

Figure 4. K channels of the keypoint detector network used in (Siarohin et al., 2020), after the softmax activation and Gaussian fit.

Summing over the channels, we get

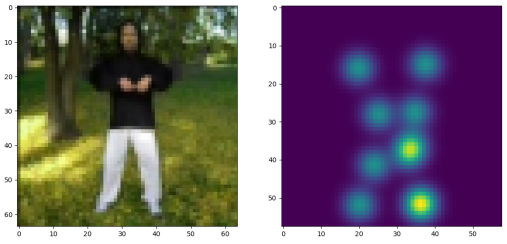

Figure 5. The sum of the K channels which is fed as a structural mask into the generator.

which is our output mask.

2.4. Losses

Perceptual Vgg19 (Simonyan & Zisserman, 2015) TODO

Table 1. Images comparison

SOURCE IMAGE	DRIVING				
					
X2FACE					
MONKEY-NET					
FOMM					
PERTURBE 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 compare 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. Our AED and AKD metrics were calculated using the following github: https://github.com/AliaksandrSiarohin/pose-evaluation.

When comparing our work visually to the previous state of the art models (Table 1) we argue that our results are visually preferable. The location of the limbs and their geometry are both more natural and more coherent with the driving frames compared to previous work. For instance, in the images in the 4th column, the geometry of the left leg in

Table 2. Accuracy Metrics

METHOD	AKD	AED	L1
X2FACE	17.654	0.272	0.080
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%

FOMM is very unnatural where as ours is more similar to the driving frame. Furthermore, in the 3rd column in the FOMM we can see that the left arm disappears where as in our work the left arm relatively follows the driving frame.

4. Future work

Test more datasets Remove the explicit sum of channels, maybe something deep in the features extracted in the keypoint detector Maybe feed all 10 channel masks into the generator Increase number of keypoints (probably won't because of memory)

5. 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 outperforms 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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Jakab, T., Gupta, A., Bilen, H., and Vedaldi, A. Unsupervised learning of object landmarks through conditional image generation. *arXiv preprint arXiv:1806.07823*, 2018.

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. *ACM Computing Surveys*, 54(1):1–41, Mar 2021. ISSN 1557-7341. doi: 10.1145/3425780. URL http://dx.doi.org/10.1145/3425780.

Newell, A., Yang, K., and Deng, J. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pp. 483–499. Springer, 2016.

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.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 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.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Thewlis, J., Bilen, H., and Vedaldi, A. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pp. 5916–5925, 2017.

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.

Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., and Lee, H. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2694–2703, 2018.

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.