Photo-Realistic Hand Gesture Video Generation Using Generative Adversarial Networks

Varun Jain

Carnegie Mellon University

`jainvarun@cmu.edu`

Abstract

Hand gestures are a natural means of interaction in Augmented Reality and Virtual Reality (AR/VR) applications. Recently, there has been an increased focus on removing the dependence of accurate hand gesture recognition on complex sensor setup found in expensive proprietary devices such as the Microsoft HoloLens, Daqri and Meta Glasses. Most such solutions either rely on multi-modal sensor data or deep neural networks that can benefit greatly from abundance of labelled data. Datasets are an integral part of any deep learning based research. They have been the principal reason for the substantial progress in this field, both, in terms of providing enough data for the training of these models, and, for benchmarking competing algorithms. However, it is becoming increasingly difficult to generate enough labelled data for complex tasks such as hand gesture recognition. The goal of this work is to introduce a framework capable of generating photo-realistic videos that have labelled hand bounding box and fingertip that can help in designing, training, and benchmarking models for hand-gesture recognition in AR/VR applications. We demonstrate the efficacy of our framework [1] in generating videos with diverse backgrounds.

1. Introduction

In the past, researchers have proposed deep neural network architectures consisting of ensemble of models that solve specific sub-tasks; For instance, sub-tasks such as hand candidate detection, fingertip detection and classification are used to achieve a larger goal of hand gesture recognition in first-person view [3]. For accurate hand gesture classification sans the depth data, each of the constituent models have to be trained separately and demand extensive human labour in annotating the data. The authors [3] use a manually annotated dataset containing over $90,000$ frames to introduce enough variability in background and lighting conditions so as to make the models robust.

On the other hand, synthetically generated data is also being increasingly used of late to train and validate vision systems [9]. This is especially true of areas in which obtaining huge amounts of data with ground truth is tedious. However, existing literature states that the performance of systems that are trained only on synthetic data is not at par with systems that are trained on real-world data due to the issue of domain shift [5]. This problem arises since the probability distribution over the parameters resulting from the process of generating the synthetic videos may diverge from the parameters that describe the real-world data. Divergence in critical parameters such as lighting, scene geometry, and camera parameters often lead to poor generalisability in models that are trained solely on synthetic data.

Various works have derived or designed representations such as geometry and motion in synthetic domains that are quasi invariant to the problem of domain shift [1]. Ros et al. [5] have showed that augmenting large scale synthetic data with even a few real-world samples while training can relieve domain shift. Moreover, recent work in the field of generative adversarial learning [2, 8], has shown how unlabelled samples from a target domain can be used to iteratively obtain better point estimates of parameters in generative models by minimising the difference between the generative and target distributions. Taking cues from the two ideas, we generate photo-realistic videos with different backgrounds and gesture patterns and hypothesise that given a large-scale dataset, one can design simpler frameworks that implicitly learn the global task of gesture recognition without needing to explicitly localise hands and fingertips.

2. Proposed Framework

2.1. CycleGAN Based Approach

We adapt the architecture for our generative networks from Zhu et al. [8] who have shown impressive results for image-to-image translation.

The network contains two $stride-2$ convolutions, two fractionally strided convolutions with $stride1$, and several residual blocks. 9 blocks are used for 256×256 size input images. To detect whether overlapping image patches

---

are real or fake, the discriminator network uses 70×70 PatchGANs [4]. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion.
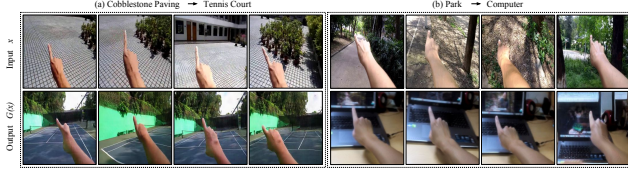

Figure 1. Results on two pairs of source and target domains. The upper row represents the real-world images and the lower row shows the synthesised image in the new domain.

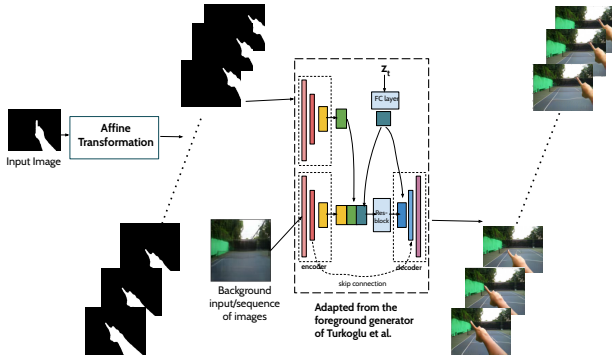## 2.2. Sequential Scene Generation with GAN

Figure 2. Proposed DNN: Given an input image, we apply gesture based affine transformation to generate a sequence of video frames. These masks passed in succession to the generator network results in a video sequence with given background image.

We used the ability of the model outlined by Turkoglu et al. [7] to generate video sequences with different backgrounds but same (or controlled) fingertip and hand as in the reference input image. The proposed framework sequentially composes a scene, breaking down the underlying problem into foreground and background separately. Our approach (figure 2) utilises the foreground generator as proposed by Turkoglu et al. [7] to superimpose elements over the given background.

3. Experiments and Results

3.1. Experiment 1

We use the Adam solver with a batch size of 1. All models were trained from scratch with a learning rate of 0.0002. The results were observed on varying number of epochs where the model was trained for 400 epochs with the same learning rate and linearly decaying the learning rate over

next 100 epochs. The model was trained on a Tesla V100 GPU for 24 hours.

We train our model on the SCUT-Ego-Finger dataset [3]. It has 93729 manually annotated frames for hand detection and fingertip detection in first-person view. The dataset includes videos from 24 different environments such as classroom, lake, canteen etc. We demonstrate our results in Figure 1 on two pairs of source and target domains: (a) $CobblestonePaving \rightarrow TennisCourt$, and (b) $Park \rightarrow Computer$.
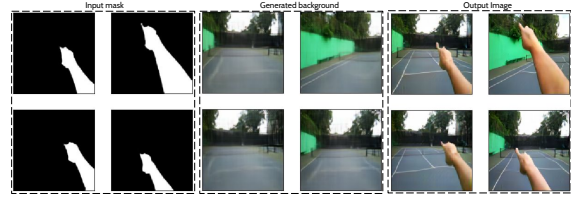
3.2. Experiment 2

Figure 3. Results obtained using the proposed method. Given a segmentation map as input, the network generates a background image and an output image taking cues from the semantic layout map. The generated images are highly photo-realistic with little or no distortion.

We ran our experiments on a subset of the SCUT-Ego-Finger dataset [3]. Since we did not have ground-truth labelled semantic maps for our dataset, skin pixels are detected from the images using the skin-colour segmentation. We applied the GrabCut algorithm [6] for foreground extraction followed by skin-thresholding in HSV colour format. Morphological erosion is also applied to remove some of the isolated blobs.

We trained the foreground and background generator (for extracting background images from the dataset [3]) for 100 and 200 epochs respectively, with a batch size of 4. Figure 3 demonstrates the complete use-case of the network. Because of the segmentation masks given as input to the model, the network is able to replicate hand and fingertip in the foreground fully.

Figure 4 shows images with different background domains but the same mask layout as input. We observe that the synthesised images do not suffer from any artefacts as compared to images generated by CycleGAN [8]. However, skin colour is a bit off in the fourth domain perhaps due to the texture of the background domain.

We extend this idea to generate egocentric gestures such as fingertip going down, up, left, and right. One such example has been demonstrated in Figure 5.

4. Future Work

The realisation of our end goal of generating photo-realistic videos with enough variability in background,
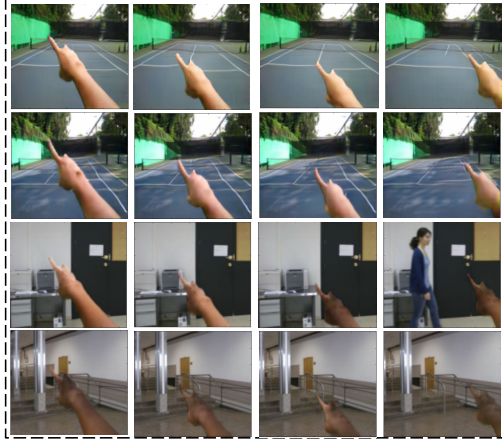
Figure 4. Results generated using same set of translated segmentation mask and fingertip location. Each row represents the synthesised images in the new domain.

lighting, and other such parameters that can help in designing, training, and benchmarking models for hand-gesture recognition would involve designing a model that introduces variations in the background features and some features present on the hand. We would like to experiment the inclusion of a recurrent network into the current framework which could generate photo-realistic hand movements corresponding to any given spatio-temporal sequence corresponding to an arbitrary input gesture. Finally, we observe that the background might change suddenly between consecutive frames leading to a jittery video and we would like to experiment with ways to make the background coherent across frames.
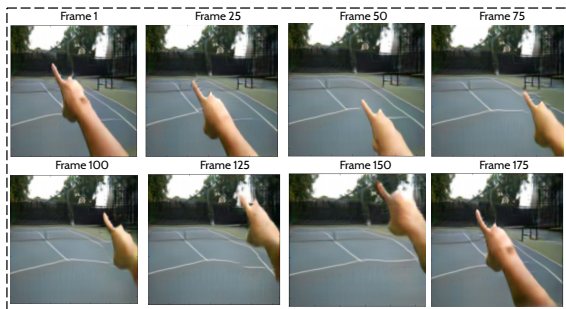

Figure 5. Generating a circle as an egocentric pointing gesture by applying orientation based mask affine transformation. Different frames depict gesture images as synthesised using the network. Note that complete gesture is obtained using a single layout mask as reference.

5. Conclusion

We have demonstrated a network capable of synthesising photo-realistic videos and show its efficacy by generating videos of hand gestures. We believe that this would help in the creation of large-scale annotated datasets, which, in turn, would encourage the development of novel neural network architectures that can recognise hand gestures from single RGB streams without the need of specialised hardware such as multiple cameras and depth sensors.

References

[1] Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] Yichao Huang, Xiaorui Liu, Xin Zhang, and Lianwen Jin. A pointing gesture based egocentric interaction system: Dataset, approach and application. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–23, 2016.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017.

[5] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[6] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, 2004.

[7] Mehmet Ozgur Turkoglu, William Thong, Luuk Spreeuwers, and Berkay Kicanaoglu. A layer-based sequential framework for scene generation with gans. In *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[9] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921. IEEE, 2017.