Towards an Unsupervised Spatiotemporal Representation Cilia Video Using A Modular Generative Pipeline

Meekail Zain, Sonia Rao, Nathan Safir, Quinn Wyner, Isabella Humphrey, Alex Eldridge, Chenxiao Li, BahaaEddin AlAila, Shannon Quinn

Department of Mathematics | Department of Computer Science | Comparative Biomedical Sciences

The University of Georgia

Abstract

Motile cilia are a highly conserved organelle found on the exterior of many human cells. Cilia beat in rhythmic patterns to transport substances or generate signaling gradients. Disruption of these patterns is often indicative of diseases known as ciliopathies, whose consequences can include dysfunction of macroscopic structures within the lungs, kidneys, brain, and other organs. Characterizing ciliary motion phenotypes as healthy or diseased is an essential step towards diagnosing and differentiating ciliopathies. We propose a modular generative pipeline for the analysis of cilia video data so that expert labor may be supplemented for this task. Our proposed model is divided into three modules: preprocessing, appearance, and dynamics. The preprocessing module augments the initial data, and its output is fed frame-by-frame into the generative appearance model which learns a compressed latent representation of the cilia. The frames are then embedded into the latent space as a low-dimensional path. This path is fed into the generative dynamics module, which focuses only on the motion of the cilia. Since both the appearance and dynamics modules are generative, the pipeline itself serves as an end-to-end generative model. This thorough and versatile model allows experts to spend less time caught in the minutiae of cilia biopsy analysis, while also enabling new insights by quantifying subtle patterns that would be otherwise difficult to categorize

Objective

We aim to develop an unsupervised representation for videos of cilia that accounts for both *spatial* and *temporal* patterns in an independent and disentangled manner. We split this problem into three core components:

1. Preprocessing
2. Spatial Modeling
3. Temporal Modeling

These tasks are handled by the preprocessing, appearance and dynamics modules, respectively. While these tasks are not *purely* independent, they can be considered independent for the sake of organization. This also allows for each module to be changed separately without impacting the development of the others. This results in greater extensibility and flexibility.

Pipeline

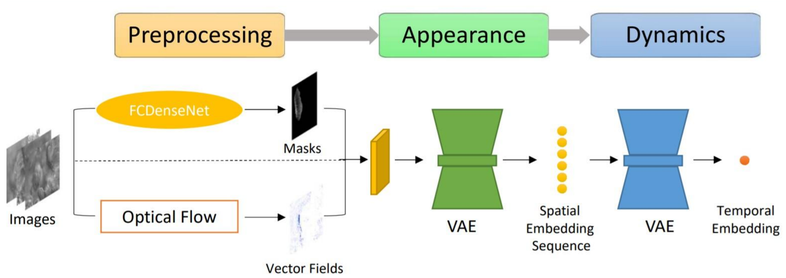

Figure 1. The pipeline starts with the preprocessing module which augments the dataset, then feeds into the appearance module to generate a spatial representation which is utilized by the dynamics module to generate a temporal representaiton.

We begin our pipeline with preprocessing module which is used to extract useful information and augment the raw video source. This module:

1. Extracts optical flow (OF) fields
2. Segments cilia
3. Augments starting videos/frames with the OF fields and segmentation maps to feed into the appearance module

The appearance module accepts the augmented frames as input and uses a Variational Autoencoder (VAE) [1] to learn a potent compressed spatial representation of cilia, or alternatively, a good spatial latent space. A well constructed spatial latent space means that:

1. The relative location of embedded points in the latent space has semantic significance
2. Most noise/unnecessary information in the source data is excluded in the compressed representation
3. Temporally coherent sequences of frames (videos) get mapped to well-behaved, continuous sequences of points in the latent space (paths or trajectories)

The final module is the dynamics module which also utilizes a VAE to learn a temporal representation of cilia based on how the spatial representation *changes over time* based on the sequences of the points embedded in the spatial latent space. This means that, since the latent space contains primarily semantically meaningful traits, the temporal representation is primarily based on semantically meaningful temporal patterns.

Preprocessing

The preprocessing module accomplishes two main goals: segmentation and optical flow extraction. Segmentation is needed as a precursor to the appearance module to ensure that the module learns a representation for *only* the cilia. In an average video from our dataset, the cilia occupy a tiny minority of the actual frame, whereas most of it is either the background, or a cell. Segmenting out the cilia themselves greatly reduces the bandwidth required by the appearance network, allowing it to focus on only the cilia.

An optical flow (OF) field is a vector map that attaches *displacement vectors* to pixels in a frame. Dense OF is a variant which assigns such a vector to *every* pixel. Since cilia are very small and often have non-linear movement, we utilize dense OF to capture as much information as possible. The OF is useful since often times cilia exhibit *significantly* more local movement than the cells they're attached to or the image backgrounds, making them stand out in the OF field. However, dyskinetic cilia tend to be borderline stationary, which makes them undetected utilizing OF alone. Thus the OF field can be appended to the frames as input to the appearance network to allow the network to learn a spatial representation from the motile cilia which can hopefully extend to immotile cilia as well. Further research is needed to determine the efficacy of this methodology.

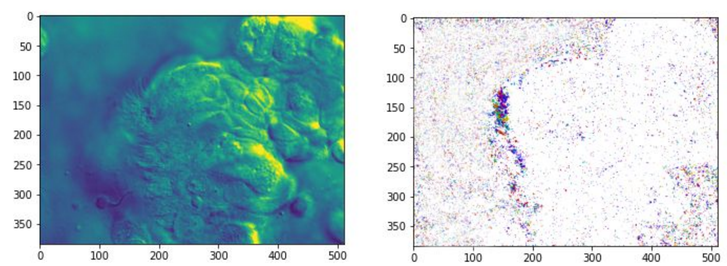

Figure 2. Frame of a cell and its cilia on the left, with the corresponding OF field on the right.

Appearance

The appearance module utilizes a convolutional VAE with ResNet like encoder and decoder. The normal unit Gaussian prior is replaced with a VampPrior [2] and a novel regularization term is added to the VAE loss to encourage the VampPrior pseudo-inputs to look like natural data.

The appearance module can accept any augmented frames from the preprocessing module in the form of multi-channel images, with each additional type of information concatenated as an additional channel. Segmentation maps, however, are processed separately. The entire augmented frame is then pixel-wise multiplied against the segmentation maps *before* being encoded, allowing the VAE to focus on only the cilia in the frame. The reconstruction objective of the VAE is thus also only the cilia portion of the original frame.

Dynamics

The dynamics module is trained and operated strictly on the *spatial latent space* generated by the appearance module, and thus is sensitive to the efficacy of the prior modules. Operating on the appearance module's compressed representation omits the vast amount of noise and redundancy intrinsic to video datasets, and in particular our cilia dataset. This allows the dynamics module to focus on the changes and patterns within the meaningful spatial latent space, narrowing the scope of its learning, allowing for a simpler architecture as compared to operating on the raw videos themselves.

The dynamics module takes as input a *sequence of points* in the *spatial latent space* and aims to represent the motion intrinsic to the entire sequence. A straightforward approach might be to simply attempt to compress and recreate the entirety of each sequence, to encourage the robustness of the representation, we assert that while the model encodes the entire sequence, it ought to be able to decode and reconstruct *arbitrary subsequences*. This mitigates the risk of having the temporal representation couple and encode the *length* of a sequence as well, allowing us to arbitrarily generalize sequences so that we may even *predict* unobserved future elements.

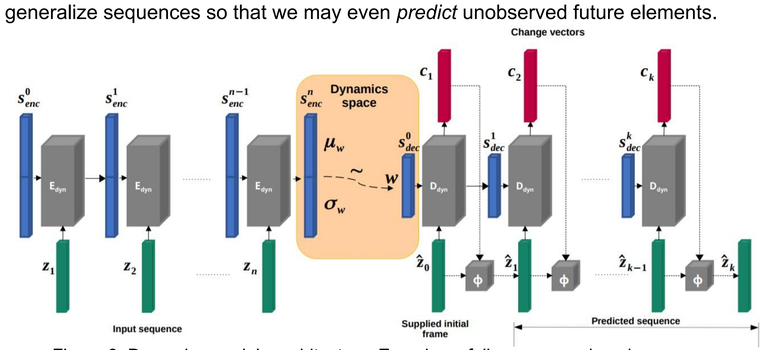

Figure 3. Dynamics module architecture. Encodes a full sequence, decodes an arbitrary subsequence

Conclusion

This project proposes a modular, generative pipeline for learning a factored spatiotemporal representation of cilia. The modular nature of the project facilitates exploration and ensures extensibility while also allowing for an efficient and powerful representation. This research is still currently at an early stage, with focus primarily on the preprocessing and appearance modules, with many plans for future research into refining this pipeline.

References

- [1] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. CoRR, abs/1906.02691, 2019. URL: http://arxiv.org/abs/1906.02691, arXiv:1906.02691.
- [2] Jakub M. Tomczak and Max Welling. VAE with a vampprior. CoRR, abs/1705.07120, 2017. URL: http://arxiv.org/abs/1705. 07120, arXiv:1705.07120