Contents

1 Telescope

Kevin Poli, Phil Vitale, Connor O'Hara and Brendan Von Hofe Advisors: Hong Man (hman@stevens.edu), Jeff Thompson (JThomps4@stevens.edu)

2 Intro

Telescope is a machine learning assisted toolkit for digital video compositors with applications in visual effects, matte painting and diverse use cases accross the video post production pipeline. Tools from existing compositing packages will interact with a novel ML core to assist or completely automate the rotoscoping process. Rotoscoping is the process of masking and segmenting poritons of an image accross multiple moving frames, any feature length movie will often consist of hundreds of rotoscoped shots with multiple tracked mattes per image, and this is the primary job of thousands of roto artists accross the world.

We hope to make this process, fast, intuitive, and accesible to alleviate the manual and time consuming process that makes up a huge chunk of the man hours required to produce even low budget features. We believe that machine learning is in the process of revolutionizing image processing, and that user driven toolkits rather than black box command line workflows will bring our intelligent core into the hands of the artists where they can thrive.

2.1 Demonstration

Rotoscoping is the process of frame by frame selecting and isolating a given feature (usually an object or person) in a video, such that you can produce a video clip of exclusively that selection on a transparent background

Lets walk through this step by step:

- First, our source image at frame 1, of Marceu the Mime

- Lets start by creating a selection just of the Mime's face and hand - these are the features that are actually being "rotoscoped" out

- This purple selection represents a 'mask' which are the points and curves that make up the boundary of what we are looking to isolate. Traditionally, artists will digitally paint this selection in a software of their choice, by hand.

- This selection or mask is different from a matte, which is another important piece of terminology. A matte is a single channel image; meaning rather than pixels having red,green,blue values, they only contain 1 value from 0-255 called 'alpha'. 'Alpha' will often be displayed in software as white. The Matte of this selection is an image where only the pixels corresponding to the selection are white, and all other pixels are black.

  - this is so that, under the hood, all we need to do is pixel-wise 'multiply' the source image to the matte, meaning any pixels with a black 'zero value' in the matte will become transparent, and any pixels in the white '255 value' in the matte will remain.

  - Here is the result of that multiply, an image containing only the pixels we selected before

2.1.1 Frame By Frame

Much of the challenge and tedium of rotoscoping comes from repeating the above process for every frame, traditionally, artists will go frame by frame through the video and manually adjust their selections to match the feature they are isolating, here is the next frame of that video, with an adjusted selection for clarity

to see how the selections should move as features in the video move, check out this gif that displays the matte on the left, with the source on the right, and has selection lines on both `https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwj4poGei_bdAhVvTt8KHYSXBsOQjRx6BAgBEAU&url=https%3A%2F%2Ftaukeke.com%2F2014%2F07%2Frotoscoping-in-nuke%2F&psig=AOvVaw0rzBOnhBNxm_0WD1VdybtL&ust=1539062086451365`

2.1.2 Use Cases

With our selection isolated, we can start to play with the image accordingly

By layering the source footage and our rotoscoped hand and face, we can apply an effect, like the 'colorama' effect to only the pixels we roto'd previously

1. Compositing The most popular use case for rotoscoping is Compositing, which is the process of combining multiple images into one. Consider three layers to see how this is done.

   Say we want this red square video clip to appear 'behind' the Mime's face and hand (note what appears black is acutally transparent)

   We can grab our source clip and place the square image on top

Then grab our rotoscoped face and hand and place that on top

and here is the desired effect

3 Technical Plan

3.1 Components

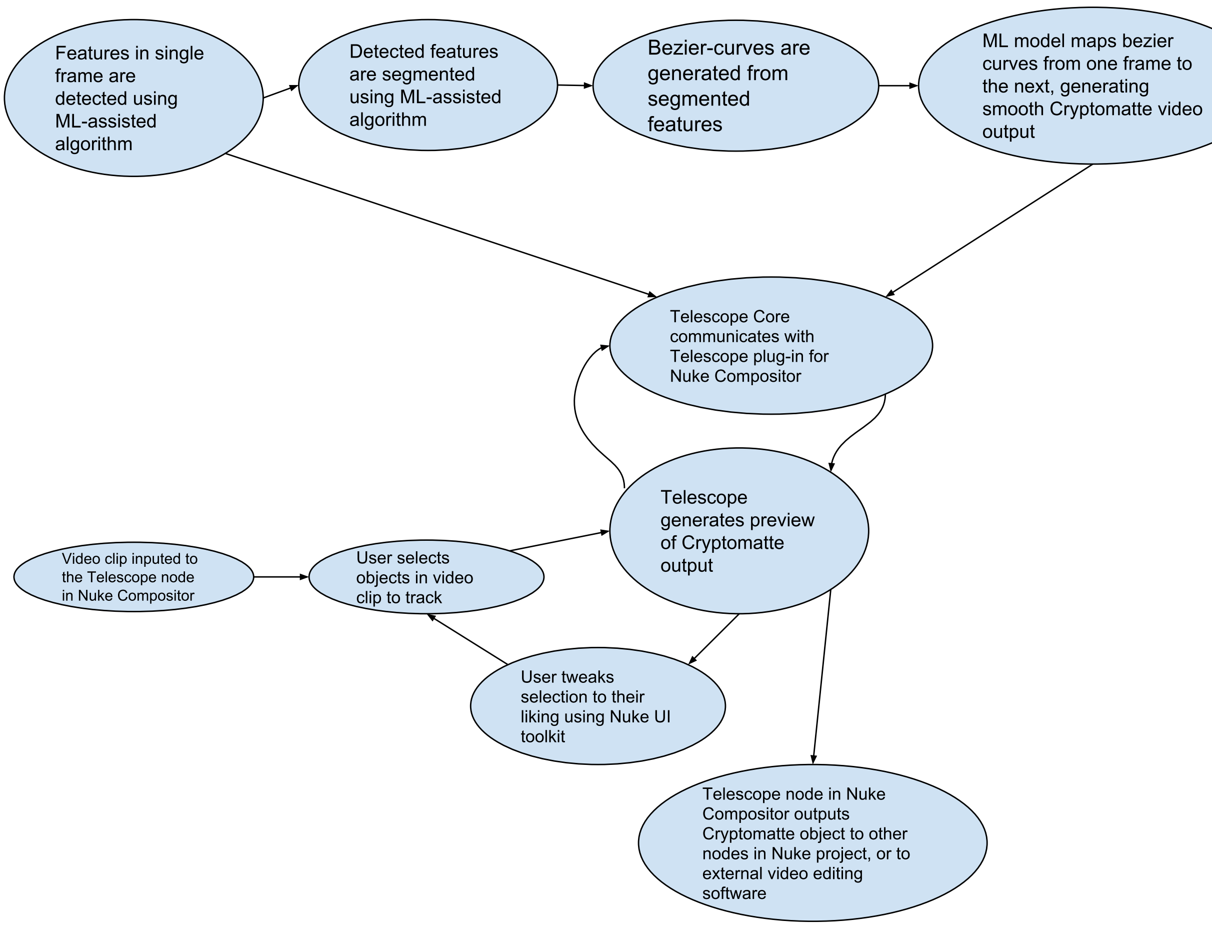
Telescope as a product will consist of two primary modules, the Telescope Core, which is a machine learning core assisted by traditional algorithmics that implements the novel functionality of Telescope, and an exchange plugin that allows existing professional compositing tools to interact with our proccesses. Telescope For Nuke is our chosen example exhange plugin, designed to demonstrate how the Telescope core can interact with existing artist workflows - but the separation of core and plugin is designed such that Telescope can be implemented into other software packages like Adobe After Effects or Blackmagic Design Fusion at a later date.

Category	What are we using?
Communication	
Email	Gmail
Web Conferencing	Facebook Video
Instant Messaging	GroupMe
Collaboration	
Document Collaboration	Google Drive
File Sharing/Data Tracking	GitHub
Plugin Development	
OS Supported	Windows, Mac OS, Linux
Host Application	Nuke
Development Language	C++
Machine Learning Development	
Development Language	Python
Packages	PyTorch

3.2 Algorithmics

The algorithmic core of our plugin will take images (frames of videos) as input and output segmentation masks (mattes) as output. The goal of the masks is to identify all the discrete objects in the image. It is class-agnostic and therefore does not need to determine what the objects are (e.g. cat or dog) but rather the fact that they are discrete. Our criteria for determining how well our model is accomplishing the task is the Intersection-over-Union metric (IoU). We have yet to determine what an acceptable IoU score is for industry applications. The model will be a convolutional neural network. Specifically, we will begin with the UNet model (`https://arxiv.org/abs/1505.04597`). Initially, our primary dataset to train the model with will be the Panoptic Detection COCO dataset, modified for a class-agnostic task. Further iterations of the model will take advantage of the additional information in EXR images to refine object mattes and the DAVIS video object segmentation dataset.

3.3 Dependency Model



4 Team

4.1 Roles

- Connor O'Hara: Image Processing (cohara1@stevens.edu)
- Kevin Poli: Application/ Artist Tools Developer (kpoli@stevens.edu)
- Philip Vitale: Application & Systems Developer (pvitale@stevens.edu)
- Brendan von Hofe: Machine Learning (bvonhofe@stevens.edu)

4.2 Delegation of Tasks

4.2.1 Connor O'Hara

1. Last Week
   - Research Generative Ladder Networks
2. Update
   - Research of Trimatte(explained later) has made this approach possibly obsolete
3. For Next week
   - Begin containerizing relevant binaries for Nuke, compilers and everything required to develop the project in Docker

4.2.2 Kevin Poli

1. Last Week
   - Follow along with Nuke developer tutorials, implement Nuke boilerplate
2. Update
   - Boilerplate requires additional mockups and ui paradigms as machine learning team is still developing their workflow
3. For Next Week
   - Design Mockups for every UI paradigm, and every node. I.e how the user will interact with our core via Nuke

4.2.3 Phil Vitale

1. Last Week
   - Follow along with Nuke developer tutorials, implement Nuke boilerplate
2. Update
   - Boilerplate requires additional mockups and ui paradigms as machine learning team is still developing their workflow
3. For Next Week
   - Research how the paradigms and mockups described above can be implemented in the Qt UI toolkit

4.2.4 Brendan Von Hofe

1. Last Week

   - Research DeepMask and SharpMask partial solutions

2. Update

   - Discovered newer research paper "Deep Image Matting" that shows very promising results with segmenting (matting) images. (This does not include tracking these objects throughout videos which will be a later stage.) `https://arxiv.org/pdf/1703.03872.pdf`

   The paper uses a similar encoder-decoder fully convolutional architecture originally specified (a.k.a. UNet). Differences include a second fully convolutional network that does not downsample or upsample the image, used for refining the image matte. They use novel loss functions dubbed the alpha-prediction loss and the compositional loss. Alpha-prediction loss is the difference in predicted alpha values at each picture of the image mask (matte) from the ground truth. Compositional loss is the difference in RGB values of the predicted composited photo (foreground, background, and alpha mask) and the ground truth composite. They also generate a novel dataset by carefully rotoscoping ~500 objects and then compositing them with thousands of images.

3. For Next Week Either procure the dataset from the authors of the paper or begin the process of recreating one. Implement baseline model.