Using artificial intelligence to automate rotoscoping for digital artists

The AC model

by Jhem Murray

29/06/2019

Abstract

Rotoscoping is the necessary but tedious (subjective) task of removing people or objects from a segment of video footage, often to add in a new background and other digital set pieces.

The target object for removal will often have to be divided into several smaller groups, and each group will be manually tracked frame by frame. A 10 second piece of video footage will usually contain 240 frames (24 fps) [1] that will need to be 'treated' one by one. Semi-automated solutions do exist; the artist can often select a group of pixels, which the software will then track (follow). Still, a level of manual intervention will be necessary if the tracking software becomes confused, i.e. a new foreground object enters the frame, obfuscating the selected pixel group.

This paper will attempt to provide near-automatic solutions to this problem, using machine learning techniques like the technology displayed in 'Deepfakes'.

Deepfakes use an artificial neural network and local GPU computing power to automatically generate a video where the face of a target person is replaced with another person's face in a segment of video footage. The software uses Google's AI-Framework TensorFlow. [2]

The results tend to be convincing (subjective), considering the fully automated nature of the process. This opens the idea of whether such a technique can be used to aid digital artists in rotoscoping tasks.

How Deepfakes work

In simple terms, Deepfakes utilize generative adversarial networks called GANs. One machine learning model will study a large data set, and then attempt to generate the video 'forgery'. The other ML model with attempt to detect and discriminate against the forgery, i.e. blurriness or shimmering around the edges of the face which diminish the 'photoreal' effect. The forger ML model will then attempt to fix these issues in the next iteration. This iterative process produces results that become more and more photoreal over time.[3]

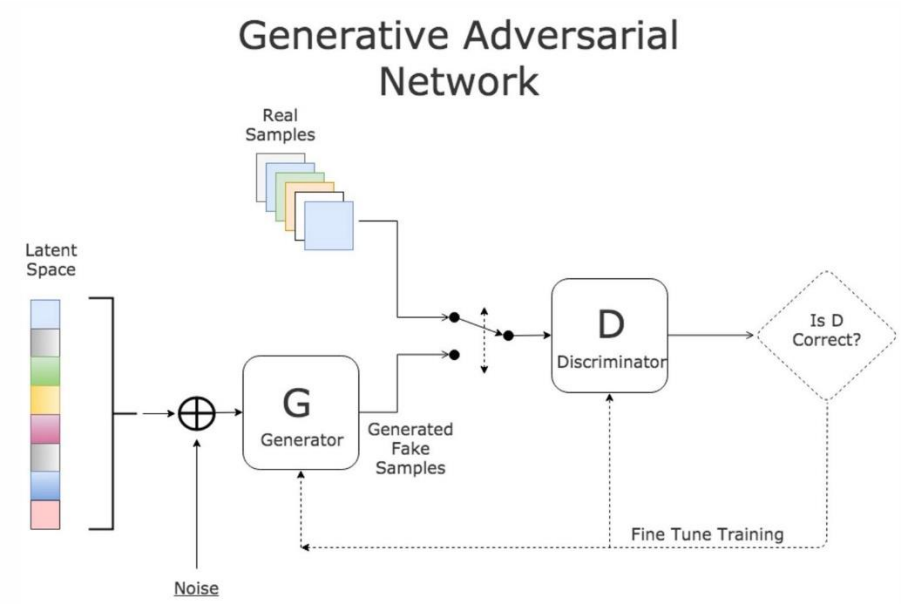


Diagram 1

Using machine learning techniques to automate rotoscoping

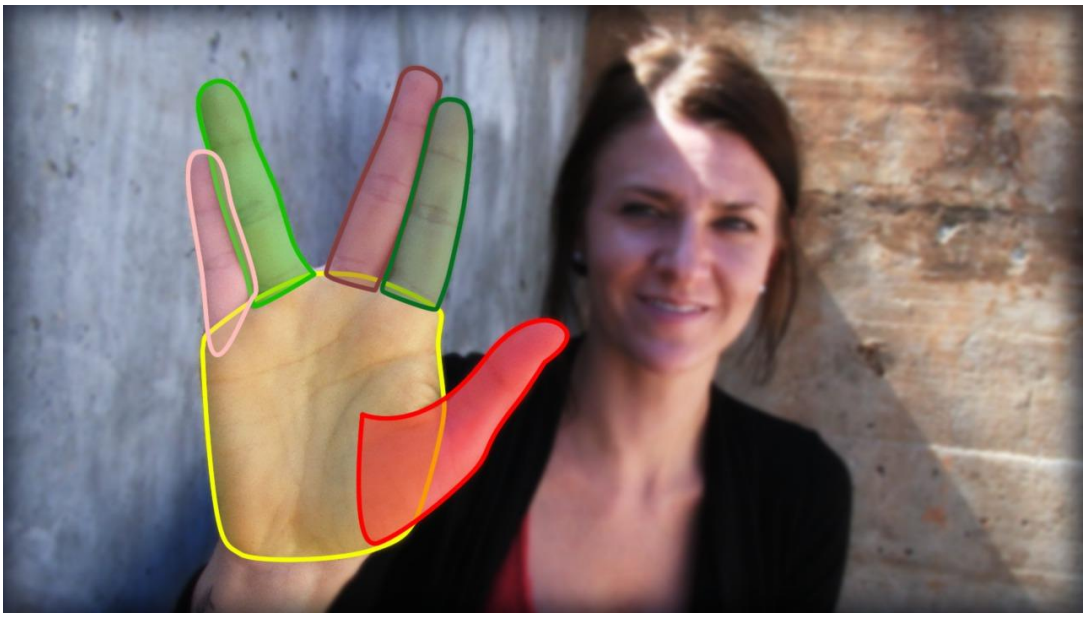
Deepfakes have produced many amusing videos of politicians saying funny things (subjective). [4] However, the machine learning models used have the advantage of having access to large data sets, in this instance, large amounts of publicly accessible videos of said politician, that can be utilised for model training.

Rotoscoping automation faces several challenges. The footage being rotoscoped can vary greatly in its content. Deepfakes have the advantage of focusing solely on face replacement, and do not have to worry about other aspects of the video footage, making the training process easier.

Digital artists can be provided with all kinds of different video and associated tasks for rotoscoping. Here are a few common examples:

1) Extract a person from the footage and replace background.
2) Remove trees from the background.
3) Remove lamp post in foreground as it passes the actor.

No single ML model will likely be able to deal with such a large range of tasks, but perhaps solutions can be provided for some of the most common types of rotoscoping work. This paper will focus on the first and arguably most common example; extracting a person from busy footage.



Extracting a person from footage automatically – The training model

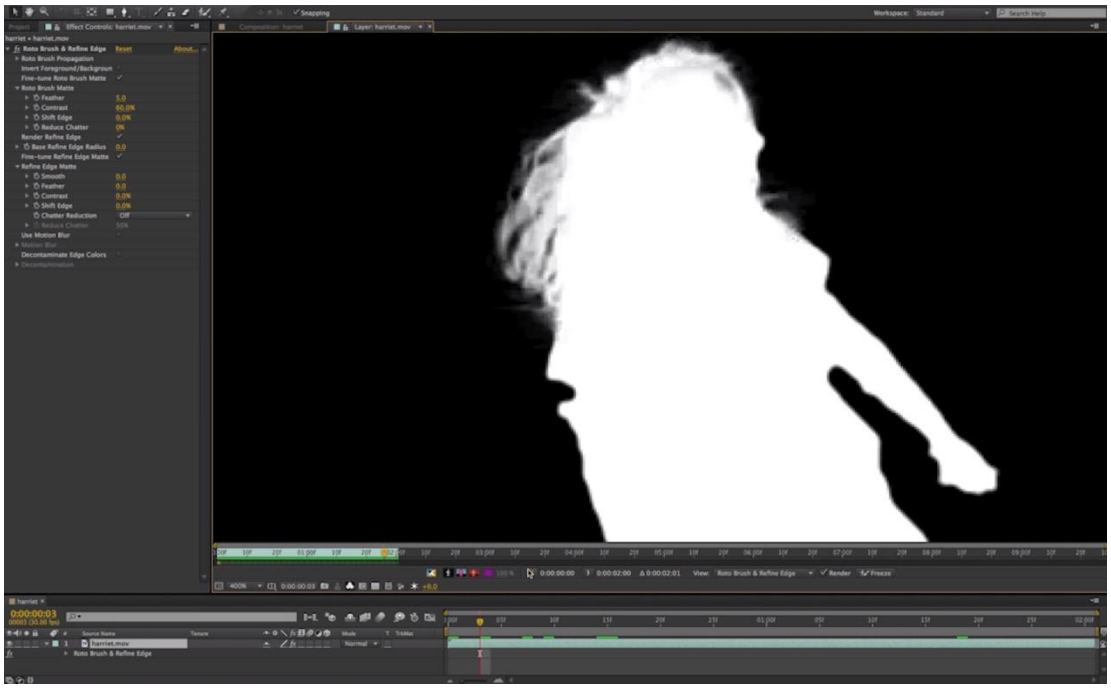
As mentioned earlier, Deepfake face replacement uses a large data set of videos to create the associated machine learning models. In the example of replacing the face of an actor with the face of a known politician, the data set required for training is readily available by searching the Internet. How can a data set be created for extracting a person from video footage?

The artist contribution training model (The AC model)

One approach to this problem is to utilise a cloud service such as the Adobe Creative Cloud. Adobe After Effects is a software tool commonly used for rotoscoping tasks. The Adobe Creative Cloud allows artists to collaborate by enabling easy file sharing, which increases efficiency and productivity in team projects. [5]

This solution proposes that artists within the Adobe Creative Cloud network submit pre (raw footage) and post processed (final and mask footage) examples to be included in the data set which will be used to train the AC model. The machine learning model will compare the raw video footage to the extraction mask and make attempts to automatically extract masks in the future.

A mask is a greyscale image resembling a silhouette, where the white pixels are fully opaque, and the black pixels are fully transparent. Extracting an accurate mask is usually the aim of the rotoscoping task. Once a clean mask has been generated, the artists can sample different backgrounds and effects, without having to redo the difficult and time-consuming rotoscoping work again. [6]



Once the AC data set contains enough examples of raw vs processed footage (the mask), future rotoscoping jobs are expected to become more and more automated, with the artist only having to intervene manually in certain problem areas, which will be fed into the Discriminator.

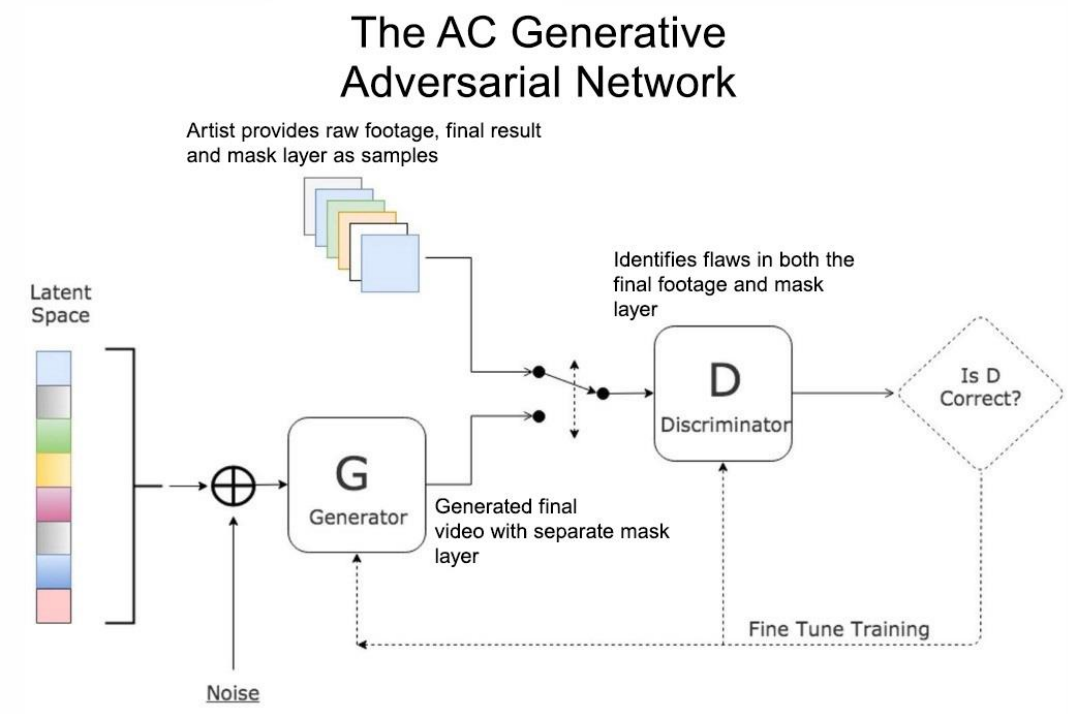
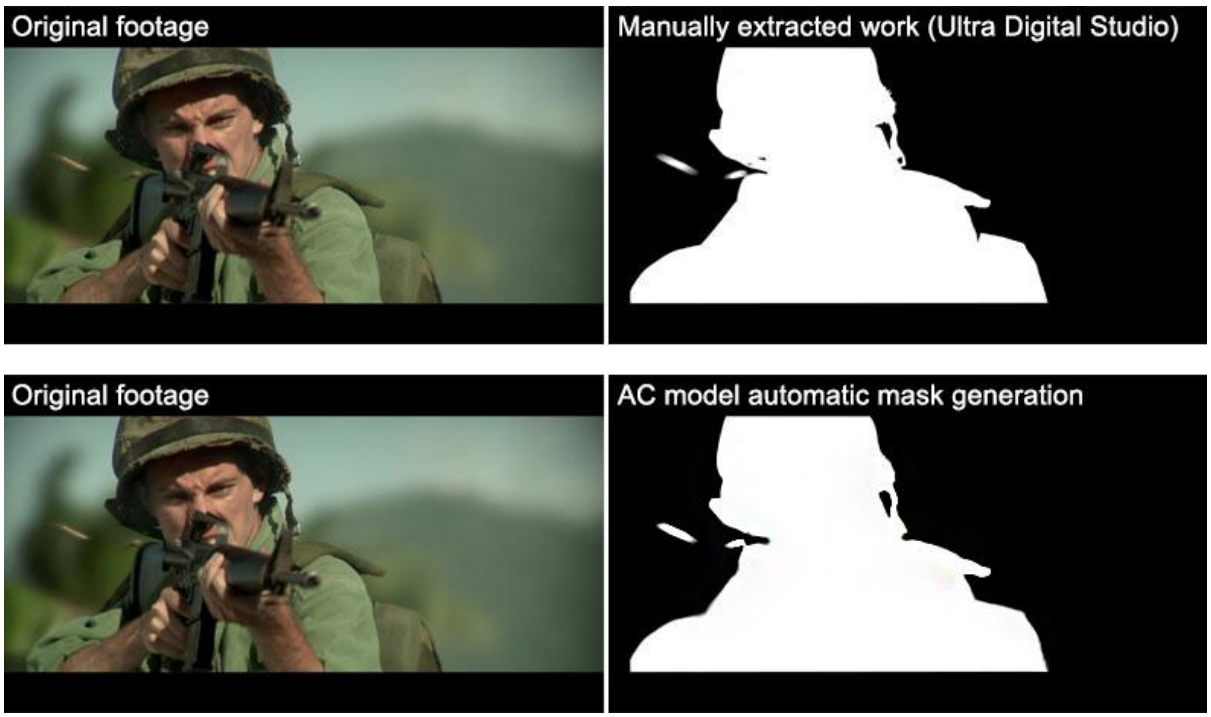


Diagram 2

Examples



The above example demonstrates real-time mask extraction, using the AC model. At the top, raw footage vs the manual rotoscoping work performed by Ultra Digital Studio. At the bottom, raw footage vs the Ocular Intelligence automatic AC model. This is a difficult shot to extract as the uniform of the soldier and background are very similar in colour, softening separation edges.

Ocular Intelligence

Ocular Intelligence is a UK Ltd company who specialise in machine learning and machine vision solutions. [7] Ocular Intelligence have developed an early model of the AC machine learning model, using a dataset created from 30 rotoscoping tasks (each task contained 100 frames). The raw footage and the manually extracted mask layers were fed into the AC ML model. Despite the relatively small data set used, the results have been close to the manually produced Ultra Digital Studio results.

In the previous example of extracting the soldier mask, the manual task performed by UDS took 74 hours in order to extract a clean mask layer from 97 frames. [8] The AC model extracted the mask layer in 2.2 seconds, with an estimated error range of 2.2% (assuming the manually performed work is 100% accurate). The 'error' range is calculated using a formula based on pixel deviation from the manual mask.

Accuracy in this instance can be subjective as one could argue that less 'accurate' results in certain portions of the actor's silhouette could happen to look more 'desirable' and pleasing to the eye.

One the data set becomes larger, i.e. via a large increase in submissions through the Adobe Creative Cloud network, it can be expected that the results will become increasingly desirable.

At this stage, a new method of error estimation can be developed, based on the subjective but likely accurate error perception of the artists, contributing to the Discriminator.

References:

1. (n.d.) in Wikipedia, retrieved June 29, 2019 from https://en.wikipedia.org/wiki/24p
2. (n.d.) in Wikipedia, retrieved June 29, 2019 from https://en.wikipedia.org/wiki/Deepfake
3. (n.d.) in Wikipedia, retrieved June 29, 2019 from https://en.wikipedia.org/wiki/Generative_adversarial_network
4. Feb 20, 2019 in YouTube, President Donald Trump on Alec Baldwin Deepfake: https://www.youtube.com/watch?v=XdBDouKV828
5. (n.d.) in Adobe.com, retrieved June 29, 2019 from https://www.adobe.com/uk/creativecloud.html
6. (n.d.) in Understanding Layer Masks In Photoshop in Photoshopessentials.com from https://www.photoshopessentials.com/basics/layers/layer-masks/
7. Ocular Intelligence Ltd from www.ocularintelligence.com
8. (n.d.) in Rotoscoping & Paint, retrieved June 29, 2019 from http://www.ultradigitalstudio.com/rotoscoping.php