Why Deepfakes Cannot Currently Convey Subtlety of Emotion

By Martin Anderson

First published ***February 3rd, 2022*** *at:*

https://www.unite.ai/why-deepfakes-cannot-currently-convey-subtlety-of-emotion/

Web-archived version

Yesterday's debut of episode 6 of the *Star Wars* spin-off *The Book of Boba Fett* seems to have divided fan opinion. Received to general approbation, there's a sweeping assumption across social networks that the much-improved recreation of a de-aged Mark Hamill (compared to the character's prior appearance in the season 2 finale of *The Mandalorian* in 2020) is a direct result of Industrial Light and Magic hiring the amateur deepfakes practitioner Shamook (who had radically improved on their work with open source software); and that the renderings of the character must be a combination of deepfake technology, perhaps tidied up with CGI.

There's currently limited confirmation of this, though Shamook has said little to the world since the ILM contractual NDA descended. Nonetheless, the work is an extraordinary improvement on the 2020 CGI; exhibits some of the 'glossiness' associated with deepfake models derived from archival works; and in general accords with the best current visual standard for deepfakes.

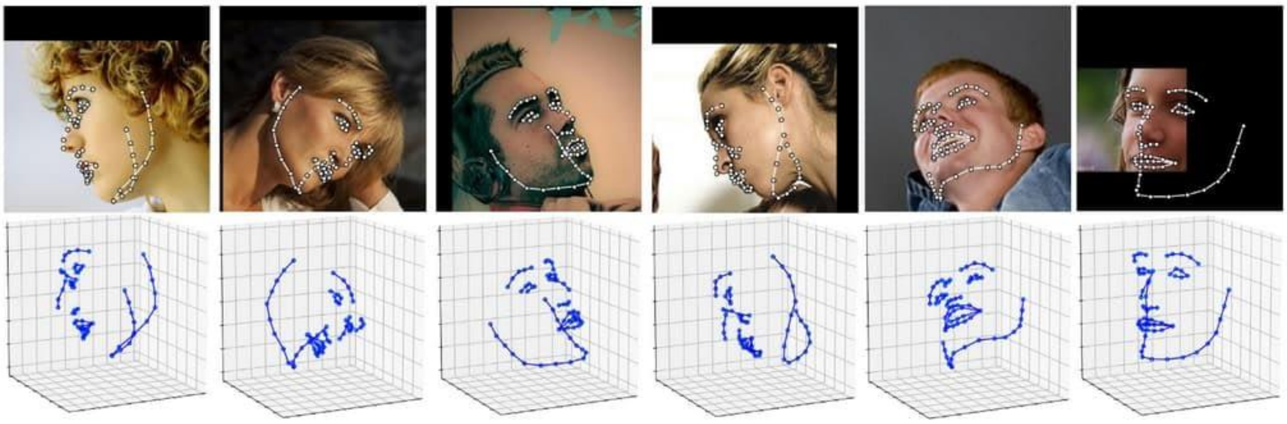
The other strand of fan opinion is that the new attempt at 'Young Luke' has a *different set of flaws* than the previous one. Perhaps most tellingly, the lack of expressiveness and subtle, apposite emotions in the very long sequences featuring the new Skywalker recreation are more typical of deepfakes than CGI; The Verge has described the *Boba Fett* simulation in terms of the *'uncanny, blank visage of Mark Hamill's frozen 1983 face'*.

Regardless of the technologies behind the new ILM recreation, deepfake transformations have a fundamental problem with subtlety of emotion that is difficult to address either by changes in the architecture or by improving the source training material, and which is typically evaded by the careful choices that viral deepfakers make when selecting a target video.

Facial Alignment Limitations

The two deepfake FOSS repositories most commonly used are DeepFaceLab (DFL) and FaceSwap, both derived from the anonymous and controversial 2017 source code, with DFL having an enormous lead in the VFX industry, despite its limited instrumentality.

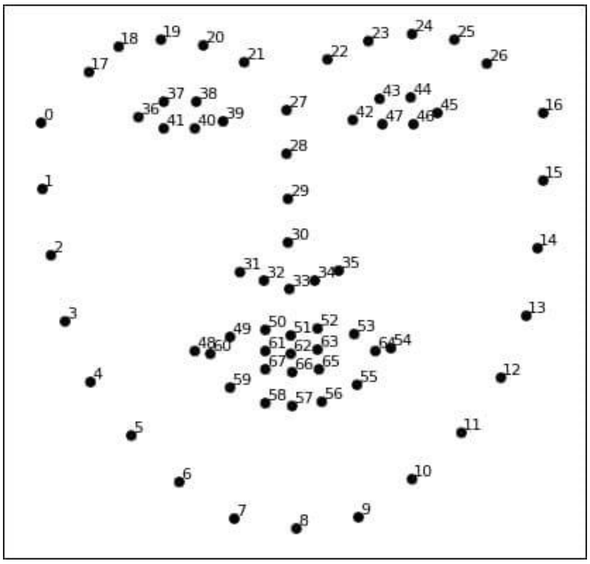
Each of these packages is tasked, initially, with extracting facial landmarks from the faces that it has been able to identify from the source material (i.e. frames of videos and/or still images).



The Facial Alignment Network (FAN) in action, from the official repository. Source: https://github.com/1adrianb/face-alignment

Both DFL and FaceSwap use the Facial Alignment Network (FAN) library. FAN can create 2D and 3D (see image above) landmarks for extracted faces. 3D landmarks can take extensive account of the perceived orientation of the face, up to extreme profiles and relatively acute angles.

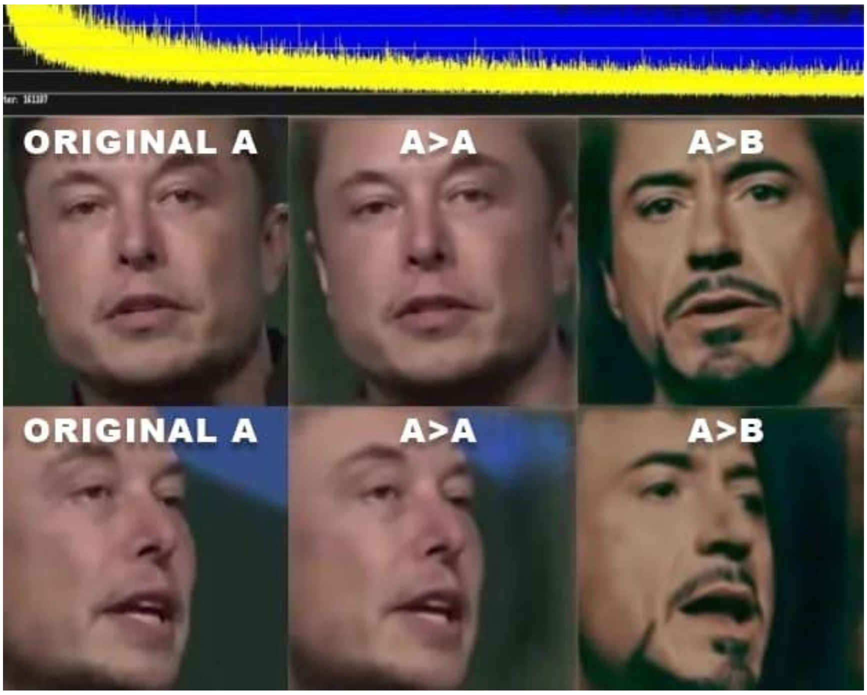
However, it's evident that these are very rudimentary guidelines for herding and evaluating pixels:



From the FaceSwap forum, a rough indicator of the available landmarks for facial lineaments. Source: https://forum.faceswap.dev/viewtopic.php?f=25&t=27

The most basic lineaments of the face are allowed for: eyes can widen and close, as can the jaw, while basic configurations of the mouth (such as smiling, scowling, etc.) can be traced and adapted. The face can rotate in any direction up to around 200 degrees from the camera's point of view.

Beyond that, these are pretty crude fences for the ways that pixels will behave within these boundaries, and represent the only truly mathematical and precise facial guidelines in the entire deepfake process. The training process itself simply compares the way pixels are disposed within or near these boundaries.

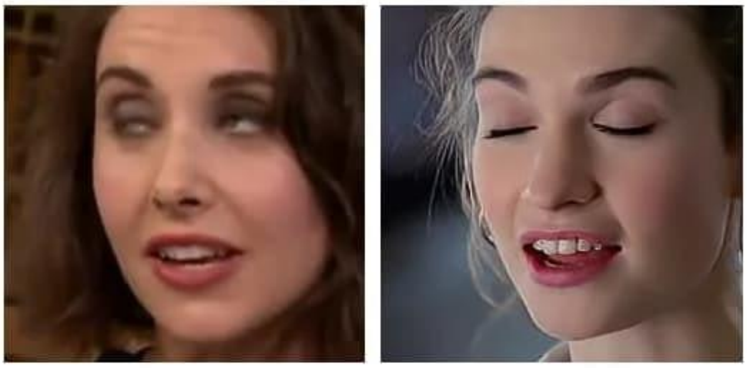


Training in DeepFaceLab. Source: https://medium.com/geekculture/realistic-deepfakes-with-deepfacelab-530e90bd29f2

Since there's no provision for topology of sub-parts of the face (convexity and concavity of cheeks, aging details, dimples, etc.), it's not even possible to *attempt* to match such 'subtle' sub-features between a source (*'face you want to write over'*) and a target (*'face you want to paste in'*) identity.

Making Do With Limited Data

Getting matched data between two identities for the purposes of training deepfakes is not easy. The more unusual the angle that you need to match, the more you may have to compromise on whether that (rare) angle match between identities A and B actually features *the same expression.*



Close, but not exactly a match.

In the example above, the two identities are fairly similar in disposition, but this is as near as this dataset can get to an exact match.

Clear differences remain: the angle and lens don't exactly match, and neither does the lighting; subject A does not have their eyes completely shut, unlike subject B; the image quality and compression is worse in subject A; and somehow subject B seems much *happier* than subject A.

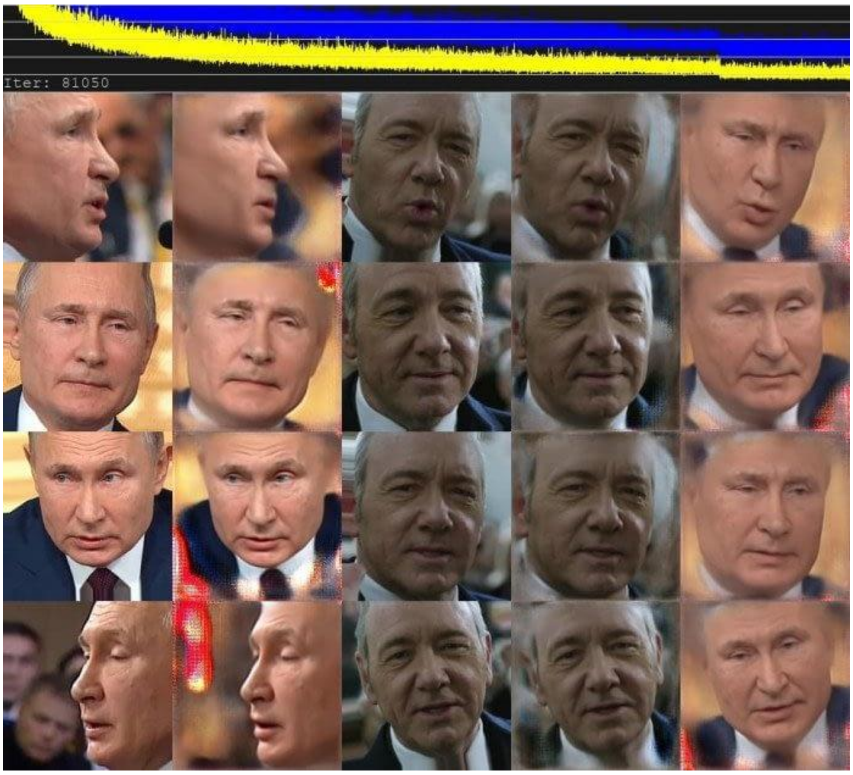
But, you know, it's all we've got, so we're going to have to train on it anyway.

Because this A><B match has so many unusual elements in it, you can be certain that there are few, if any, similar pairings in the set. Therefore the training is going to either *underfit* it or *overfit* it.

Underfit: If this match is a true minority (i.e. the parent dataset is quite large, and doesn't often feature the characteristics of these two photos), it's not going to get a lot of training time compared to more 'popular' (i.e. easy/neutral) pairings. Consequently this angle/expression isn't going to be well-represented in a deepfake made with the trained model.

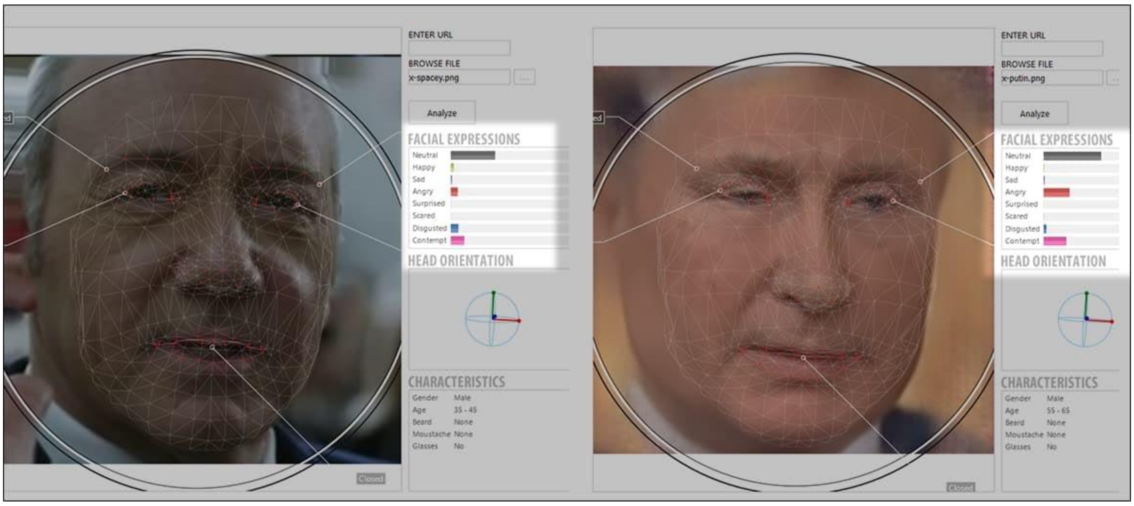
Overfit: In desperation at scant data-matches for such rare A><B pairings, deepfakers will sometimes *duplicate the pairing many times* in the dataset, so that it gets a better shot at becoming a feature in the final model. This will lead to overfitting, where deepfake videos made with the model are likely to *pedantically repeat the mismatches* that are evident between the two photos, such as the differing extent to which the eyes are shut.

In the image below, we see Vladimir Putin being trained in DeepFaceLab to perform a swap into Kevin Spacey. Here, the training is relatively advanced at 160,000 iterations.



Source: https://i.imgur.com/OdXHLhU.jpg

The casual observer might contend that Putin looks a little, well, *spacier* than Spacey in these test-swaps. Let's see what an online emotion recognition program makes of the mismatch in expressions:



BROWSE FILE x-spacey.png

Gender Male
Age 35 - 45
Beard None
Moustache None
Glasses No

BROWSE FILE x-putin.png

Gender Male
Age 55 - 65
Beard None
Moustache None
Glasses No

Source: https://www.noldus.com/facereader/measure-your-emotions

According to this particular oracle, which analyzes a much more detailed facial topography than DFL and Faceswap, Spacey is less *angry*, *disgusted*, and *contemptuous* than the resulting Putin deepfake in this pairing.

The unequal expressions come as part of an entangled package, since the popular deepfakes applications have no capacity to register or match expressions or emotions, except tacitly, as a raw pixel>pixel mapping.

For us, the differences are huge. We learn to read facial expressions as a basic survival technique from our earliest years, and continue to rely on this skill in adulthood for purposes of social integration and progression, mating, and as an ongoing threat assessment framework. Since we're so sensitized to micro-expressions, deepfake technologies will eventually need to account for this.

Against the Grain

Though the deepfake revolution has brought the promise of inserting 'classic' movie stars into modern movies and TV, AI cannot go back in time and shoot their classic works at a more compatible definition and quality, which is pivotal to this use case.

On the assumption (and for our purposes, it doesn't matter if it's wrong) that the *Boba Fett* Hamill reconstruction was largely the work of a trained deepfake model, the dataset for the model would have needed to exploit footage from the period near to the timeline of the show (i.e. Hamill as an early thirtysomething around the time of production for *Return of the Jedi*, 1981-83).

The movie was shot on Eastman Color Negative 250T 5293/7293 stock, a 250ASA emulsion that was considered medium to fine-grained at the time, but was surpassed in clarity, color range and fidelity even by the end of the 1980s. It's a stock of its time, and the operatic scope of *Jedi* afforded few close-ups even to its leading actors, making grain issues even more critical, since the source faces occupy only a part of the frame.

A range of scenes of Hamill in Return of the Jedi *(1983).*

Additionally, a lot of the VFX-laden footage featuring Hamill would have been run through an optical printer, increasing the film grain. However, access to the Lucasfilm archives – which have presumably taken good care of the master negatives and could offer hours of additional unused raw footage – could overcome this issue.

Sometimes it's possible to cover a range of years of an actor's output in order to increase and diversify the deepfakes dataset. In Hamill's case, deepfakers are hamstrung by his change in appearance after a car accident in 1977, and the fact that he almost immediately began his second career as an acclaimed voice actor after *Jedi*, making source material relatively scarce.

Limited Range of Emotions?

If you need your deepfaked actor to chew the scenery, you're going to need source footage that contains an unusually wide range of facial expressions. It may be that the only age-apposite footage available does not feature many of those expressions.

For instance, by the time the story arc of *Return of the Jedi* came round, Hamill's character had largely mastered his emotions, a development absolutely central to the original franchise mythology. Therefore if you make a Hamill deepfake model from *Jedi* data, you're going to have to work with the more limited range of emotions and uncommon facial composure that Hamill's role demanded of him at that time, compared to his earlier entries in the franchise.

Even if you consider that there are moments in *Return of the Jedi* where the Skywalker character is under stress, and could provide material for a greater range of expressions, face material in these scenes is nonetheless fleeting and subject to the motion blur and fast editing typical of action scenes; so the data is pretty unbalanced.

Generalization: The Merging of Emotions

If the *Boba Fett* Skywalker recreation is indeed a deepfake, the lack of expressive range that has been leveled against it from some quarters would not be entirely due to limited source material. The encoder-decoder training process of deepfakes is seeking a *generalized* model that successfully distills central features from thousands of images, and can at least *attempt* to deepfake an angle that was missing or rare in the dataset.

If not for this flexibility, a deepfake architecture would simply be copying and pasting base morphs on a per-frame basis, without considering either temporal adaptation or context.

However, the painful trade-off for this versatility is that expression fidelity is likely to be a casualty of the process, and any expressions which *are* 'subtle' may not be the right ones. We all play our faces like 100-piece orchestras, and are well-equipped to do so, whereas deepfake software is arguably missing at least the string section.

Disparity of Affect in Expressions

Facial movements and their effects on us are not a uniform language across all faces; the raised eyebrow that looks insouciant on Roger Moore might look less sophisticated on Seth Rogan, while the seductive allure of Marilyn Monroe might translate to a more negative emotion if deepfaked onto a person whose most data-available role is 'angry' or 'disaffected' (such as Aubrey Plaza's character across seven seasons of *Parks and Recreation*).

Therefore pixel><pixel equivalence across A/B face-sets is not necessarily helpful in this respect; but it's all that is on offer in state-of-the-art deepfake FOSS software.

What is arguably needed is a deepfake framework that not only can recognize expressions and infer emotions, but has the ability to embody high-level concepts such as *angry*, *seductive*, *bored*, *tired*, etc., and to categorize those emotions and their related expressions in each of the two face-set identities, rather than examining and replicating the disposition of a mouth or an eyelid.