The Dawn of Deepfaked Emotions

Originally published at https://www.unite.ai/the-dawn-of-deepfaked-emotion-invertable-frowns/

Researchers have developed a new machine learning technique to arbitrarily impose new emotions on faces in video, adapting existing technologies that have recently emerged as solutions to match lip movements to foreign language dubbing.

The research is an equal collaboration between Northeastern University at Boston and the Media Lab at MIT, and is titled *Invertable Frowns: Video-to-Video Facial Emotion Translation*. Though the researchers concede that the initial quality of the results must be developed through further research, they claim that the technique, called Wav2Lip-Emotion, is the first of its kind to directly address full-video expression modification through neural network techniques.

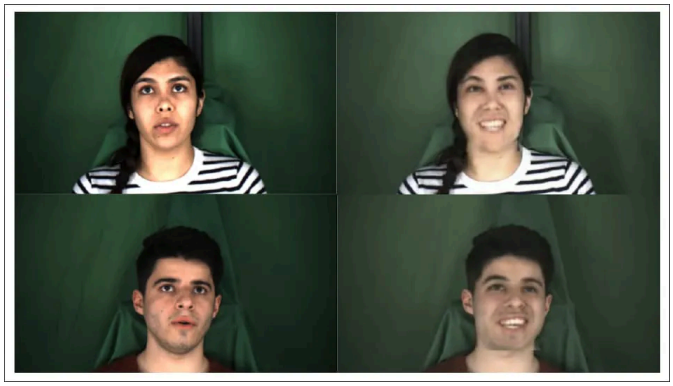
The base code has been released on GitHub, though model checkpoints will be added to the open source repository later, the authors promise.

On the left, a 'sad' frame of the source video. On the right, a 'happy' frame. In the center are two nascent approaches to synthesizing alternate emotions – top row: a fully-masked face where the entirety of the expression surface has been substituted; bottom row: a more traditional Wav2Lip method, which only substitutes the lower part of the face. Source: https://raw.githubusercontent.com/jagnusson/Wav2Lip-Emotion/main/literature/ADGD_2021_Wav2Lip-emotion.pdf

Single Video as Source Data

In theory, such manipulations are obtainable now through full-bodied training on traditional deepfake repositories such as DeepFaceLab or FaceSwap. However, the standard workflow would involve using an alternate identity to the 'target' identity, such as an actor impersonating the target, whose own expressions would be transferred to another individual, along with the rest of the performance. Additionally, deepfake voice cloning techniques would usually be necessary to complete the illusion.

Further, actually changing the expression of *target1>target1* in a sole source video under these popular frameworks would involve changing the facial alignment vectors in a way that these architectures do not currently facilitate.



Wav2Lip-Emotion maintains lips synchronization of the original video audio dialogue while transforming the associated expressions.

Instead, Wav2Lip-Emotion effectively seeks to 'copy and paste' emotion-related expressions from one part of a video and substitute them into other points, with a self-imposed frugality of source data that's intended eventually to offer a lower-effort method for expression manipulation.

Offline models could later be developed that are trained on alternate videos of the speaker, obviating the need for any one video to contain a 'palette' of expression states with which to manipulate the video.

Potential Purposes

The authors suggest a number of applications for expression modification, including a live video filter to compensate for the effects of PTSD and facial palsy sufferers. The paper observes:

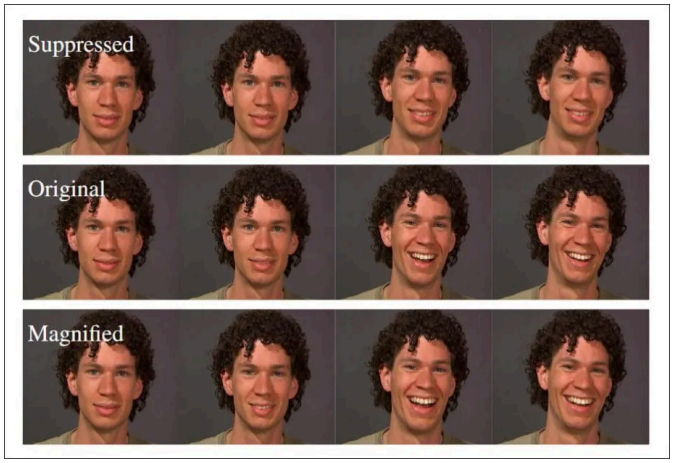
> *'Individuals with or without inhibited facial expressions may benefit from tuning their own expressions to better fit their social circumstances. One may want to alter the expressions in videos shown to them. Speakers might be yelling at each other during a video conference, but nevertheless want to gather the content in their exchange without the unpleasant expressions. Or a film director may want to augment or diminish the expressions of an actor.'*

Since facial expression is a key and core indicator of intent, even where it may abrade against the words being spoken, the ability to alter expression also offers, to an extent, the ability to change how the communication is received.

Prior Work

Interest in machine learning expression alteration goes back at least to 2012, when a collaboration between Adobe ADBE 0.82%, Facebook and Rutgers University proposed a method to alter expressions by using a Tensor-based 3D geometry reconstruction approach, which laboriously imposed a CGI mesh over each frame of

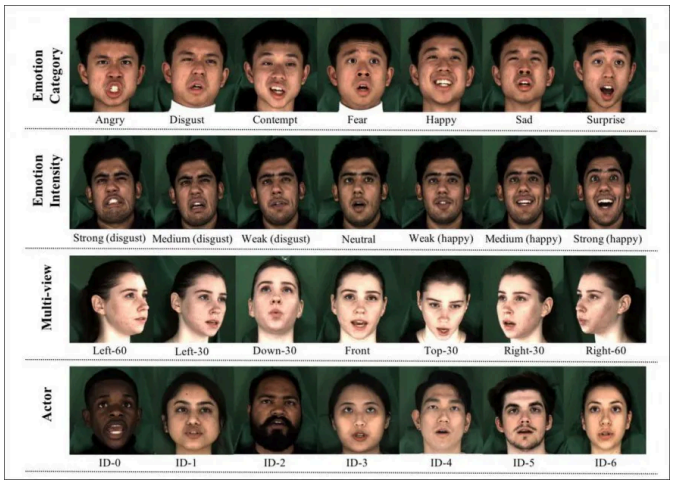
a target video in order to effect the change.



The 2012 Adobe/Facebook research manipulated expressions by imposing traditional, CGI-driven changes onto video footage. Expressions could be augmented or suppressed. Source: https://yfalan.github.io/files/papers/FeiYang_CVPR2012.pdf

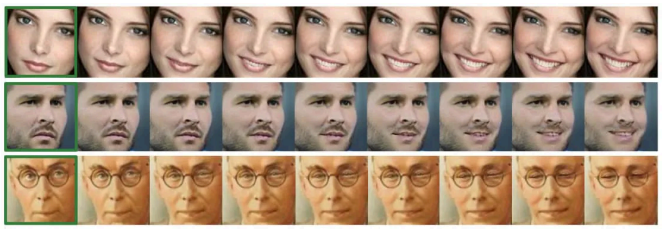
Though the results were promising, the technique was burdensome and the resources needed were considerable. At this point, CGI was far ahead of computer vision-based approaches to direct feature space and pixel manipulation.

More closely related to the new paper is MEAD, a dataset and expression-generation model released in 2020, capable of generating 'talking-head' videos, though without the level of sophistication that is potentially obtainable by modifying actual source video directly.



Expression generation with 2020's MEAD, a collaboration between SenseTime Research, Carnegie Mellon, and three Chinese universities. Source: https://wywu.github.io/projects/MEAD/MEAD.html

In 2018 another paper, entitled *GANimation: Anatomically-aware Facial Animation from a Single Image*, emerged as US/Spanish academic research collaboration, and used Generative Adversarial Networks to augment or change expressions in still images only.



Changing expressions in still images with GANimation. Source: https://arxiv.org/pdf/1807.09251.pdf

Wav2Lip-Emotion

Instead, the new project is based on Wav2Lip, which garnered publicity in 2020 by offering a potential method for re-synching lip movement to accommodate novel speech (or song) input that never featured in the original video.

The original Wav2Lip architecture was trained on a corpus of spoken sentences from the BBC archives. In order to adapt Wav2Lip towards the task of expression alteration, the researchers 'fine tuned' the architecture on the above-mentioned MEAD dataset.

MEAD consists of 40 hours of video featuring 60 actors reading out the same sentence whilst performing a variety of facial expressions. The actors hail from 15 different countries and offer a range of international characteristics aimed at helping the project (and derived projects) to produce applicable and well-generalized expression synthesis.

At the time of research, MEAD had only released the first part of the dataset, featuring 47 individuals performing expressions such as 'angry', 'disgust', 'fear', 'contempt', 'happy', 'sad' and 'surprise'. In this initial outing into a new approach, the researchers limited the scope of the project to superimposing or otherwise altering the perceived emotions 'happy' and 'sad', since these are the most easily recognized.

Method and Results

The original Wav2Lip architecture replaces only the lower section of the face, whereas Wav2Lip-Emotion also experiments with a full facial replacement mask and expression synthesis. Thus it was necessary for the researchers to additionally modify the in-built evaluation methods, since these were not designed for a full-face configuration.

The authors improve on the original code by retaining the original audio input, maintaining consistency of lip movement.

The generator element features an identity encoder, speech encoder and face decoder, in accordance with the earlier work. The speech element is encoded additionally as stacked 2D convolutions that are subsequently concatenated to their associated frame/s.

Besides the generative element, the modified architecture features three main discriminator components, targeting the quality of lip synchronization, an emotion objective element, and an adversarially trained visual quality objective.

For full face reconstruction, the original Wav2Lip work contained no precedent, and therefore the model was trained from scratch. For lower-face training (half-mask), the researchers proceeded from checkpoints included in the original Wav2Lip code.

Beside automatic evaluation, the researchers used crowd-sourced opinion supplied by a semi-automated service platform. The workers generally rated the output highly in terms of recognizing the superimposed emotions, while only reporting 'moderate' evaluations for image quality.

The authors suggest that, besides improving generated video quality with further refinements, future iterations of the work could encompass a wider range of emotions, and that the work could equally be applied in future to labeled or automatically inferred source data and datasets, leading, eventually, to an authentic system in which emotions could be dialed up or down at the whim of the user, or ultimately replaced by contrasting emotions in respect to the original source video.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More