Microsoft's AI app VASA-1 makes photographs talk and sing with believable facial expressions

April 19 2024, by Bob Yirka

Given a single portrait image, a speech audio clip, and optionally a set of other control signals, our approach produces a high-quality lifelike talking face video of 512× 512 resolution at up to 40 FPS. The method is generic and robust, and the generated talking faces can faithfully mimic human facial expressions and head movements, reaching a high level of realism and liveliness. (All the photorealistic portrait images in this paper are virtual, non-existing identities.). Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2404.10667

A team of AI researchers at Microsoft Research Asia has developed an AI application that converts a still image of a person and an audio track into an animation that accurately portrays the individual speaking or singing the audio track with appropriate facial expressions.

The team has published a paper describing how they created the app on the *arXiv* preprint server; video samples are available on the research project page.

The research team sought to animate still images talking and singing using any provided backing audio track, while also displaying believable facial expressions. They clearly succeeded with the development of VASA-1, an AI system that turns static images, whether captured by a camera, drawn, or painted, into what they describe as "exquisitely synchronized" animations.

The group has proven the effectiveness of their system by posting short video clips of their test results. In one, a cartoon version of the Mona Lisa is performs a rap song; in another, a photograph of a woman has been transformed into a singing performance, and in yet another, a drawing of a man delivers a speech.

In each of the animations, the facial expressions change along with the words in a way that emphasizes what is being said. The researchers note also that despite the life-like nature of the videos, closer inspection can reveal flaws and evidence that they have been artificially generated.

The research team achieved their results by training their app on thousands of images with a wide variety of facial expressions. They also note that the system currently produces 512-by-512-pixel imagery running at 45 frames per second. Also, it took an average of two minutes to produce the videos using a desktop-grade Nvidia RTX 4090 GPU.

The research team suggests that VASA-1 could be used to generate extremely lifelike avatars for games or simulations. At the same time, they acknowledge the potential for abuse and are therefore not making the system available for general use.

More information: Sicheng Xu et al, VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time, *arXiv* (2024). DOI: 10.48550/arxiv.2404.10667

Project page: www.microsoft.com/en-us/research/project/vasa-1/