Laboratório VISGRAF

Instituto de Matemática Pura e Aplicada

Dance Rotoscope: AI Animation experiences

Bernardo Alevato, Luiz Velho, Jorge Lopes

Technical Report TR-24-03 Relatório Técnico

February - 2024 - Fevereiro

Dance Rotoscope: AI Animation experiences.

Abstract

This text describes the development of two animated dance sequences in which we used deep learning to simulate the rotoscoping technique. Both sequences are guided by copyright video footage. The first uses references from the artist Wassily Kandinsky triggered through the prompt to produce the images and the second uses a model with additional training based on watercolors to produce the animation frames. The description of the experiments is the basis for discussion where new possibilities are highlighted such as recognition of the human skeleton and poses. Then we make considerations regarding the coexistence between drawing, design and the use of artificial intelligence to generate animations.

Keywords: Artificial Intelligence, rotoscope, Stable Diffusion, watercolor, computer animation

1. Pseudo rotoscoping through the machine

In 1917 Max Fischer[1], animation director who created Betty Boop among other characters, patented the Rotoscope, a device that projected film images onto a clipboard. On this device it was possible to draw frame by frame using reference from the filmed image. This technique came to be called rotoscopy. Since then, despite all the changes in media, from film to digital, the applications for this technique have multiplied, and everything from visual effects to experimental films uses it to obtain reference to filmed movement in favor of some interference, visual creation or drawing. The visual effects market even uses the nomenclature rotoscoping to designate the removal of the background from the video image, separating it from a filmed person, generally for compositing. In our usage, the term is closer to the effect of drawing on the filmed image as a reference for the movement to create a new film.

From references, such as the conception of the music video “All Over” by the partnership Paula Toller and Donavon Frankenheiter [2], it is possible to see that the technique, despite opening up a series of visual possibilities in terms of freedom of graphic representation, depends on lots of human resources.

Therefore, it seems coherent to find the possibility of using artificial intelligence (AI), especially deep learning, to reduce difficulty. AI applied to images is going through a period of great development from the combination of computer vision resources, natural language and image generation.

The system chosen for image production was the Stable Diffusion model(SB)[3] version 1.5, an open source distributed image generation model created by an association of companies Stability.ai[4], RunawayML[5] and the University from Munich Compvis[6]. Stable Diffusion was distributed through an open license, which enabled the creation of tools, such as DreamBooth[7], img-to-img and Automatic 1111 [8], used in these experiments.

The fact that AI models, such as Stable Diffusion, produce images in series, using both training and external references, in addition to the possibility of estimating what a redesign would be like, brought the possibility of reviewing the rotoscoping process, of what it would be like to paint frame by frame over a filmic reference.

To generate animations with a well-defined style and coherence, other similar processes have been used with films and animations for SB itself, including using reference tools such as Controlnet[9], Deforum and Warpfusion. In our case the desired coherence is of the style, not between the frames, only the filmic

movement as a guide, the appearance of noisy and contrasting passage between the frames is desired in this specific case.

2. Development

The experiments presented seek to generate an expressive material, with an organic appearance, that resembles the technique of drawing or painting. The first, a film called *Animation1*, mixes dance footage with Kandinsky's painting style, this experiment answers whether it was possible to create an audiovisual piece. The second, called And the second is also a short film called *Animation2*, which uses model training from authorial drawings and watercolors and remixes them with video recorded by the author.
In the cases of *Animation1* and *Animation2*, the term pseudo rotoscoping can be used because "rotoscoping" determines the manual work on each frame of the film. In this process, we propose to work with the help of artificial intelligence (AI) in a generative way. These tools, in general, have the property of generating images or changing them from *prompts* of text and image guides, that is, descriptive lines of the image that generate and access a network trained on billions of data and images to form a new image from a seed.
The traditional process of artistic rotoscoping has beauty in its imperfections, making the moving image expressive and rich in details, but it tends to have problems with deadlines and budgets. It is still possible to reflect that the object produced is a film and not thousands of paintings. The idea of creating dance films seemed pertinent, as it contrasts human movement with technological movement and allows a wide space for creation between abstraction and figurative. The choreography and performance were performed by dancer *anonymous*.

2.1 Animation1

After some rehearsals, creating films through prompts, an attempt was made to make a dance film using AI as a frame generating tool, using video as a guide. A brief audiovisual piece as if it were material for a music video or graphics for an opening. The work of Kandinsky and the geometric movements proposed by the filmed choreography are taken as inspiration.

To create the sequence of images, Deforum was employed using Google Colab. The lack of consistency between frames was a problem for the literal representation of animations, but it is also a strong language element, as it is the result of continuous iteration between the prompt text, the video image and the AI model. The video was placed in slow motion to, as opposed to the rapidly changing images, bring consistency to the movement. For execution, an implementation on Google Colab Research was used, which is a platform for hosting AI models and scripts.

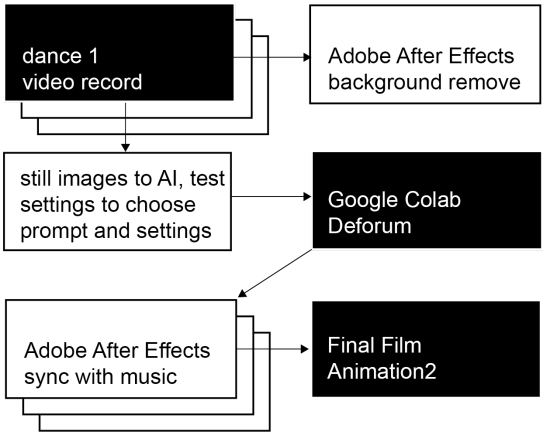

Figure 1: workflow schema

In this image generation, the data inputs were only the prompt text and the video. The video served as a flexible mask for the images provoked by the prompt. Prompts and generation parameters were tested, in this sense it is pertinent to mention that the *seed*, a random standard parameter responsible for the practically infinite variety of variation possibilities was kept fixed, after choosing the language in tests. Each frame generates a new interpretation of the image. The parameters of *steps* was used at 50 and guidance scale, which regulates how much the generation follows the prompt and how much freedom it has to find a solution, was set at 70. The tables were generated and edited in *Adobe After Effects*, synchronized with the track and finished with a texture in the background.

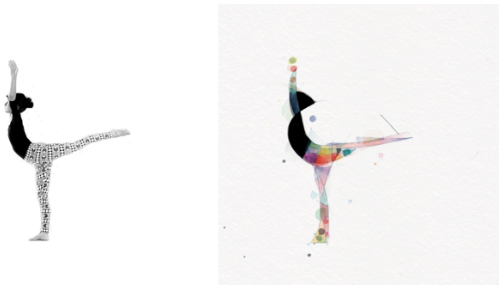

Figure 2: Comparison between the frame extracted from the video and the one resulting from generation using Deforum

In this sense, it is possible to think of this resulting object as a film within the audiovisual concept, that is, a film based on generative diffusion, where despite the guiding film, the result are synthetic images, far from the image index.

Figure 3: Montage of film frames produced from dance images and the AI with the prompt "Colorful Geometric shapes by Kandinsky" can compose illustrations

Prompt engineering was important in this case, as it is the way of communicating with the system to generate the shapes. In this case, the text assembly is the sum of the questions “What?” whose answer is colorful geometric shapes” and how, which reveals style and form “by Kandinsky".
The difference between the time of movement and the time of image changes would be unlikely in a traditional rotoscopy, because instead of repeating frames, it multiplies, which would be extremely laborious if carried out manually. The mix between video and AI referenced to the prompt in each frame produced a wide variety of drawings with the same identity, but with different shapes (Fig. 2)

2.2 Animation2

In this experiment, the mix of techniques used was filming a dance performance, training the Artificial Intelligence (AI) model with copyright images, and generating images that mix the characteristics of each image source and the AI model used. The film's linear guide is the dancer's performance filmed on DSLR.
In this path we do not exclude painting as a creative process, but it is used to create the style and thus, from this, produce the film frames generatively.

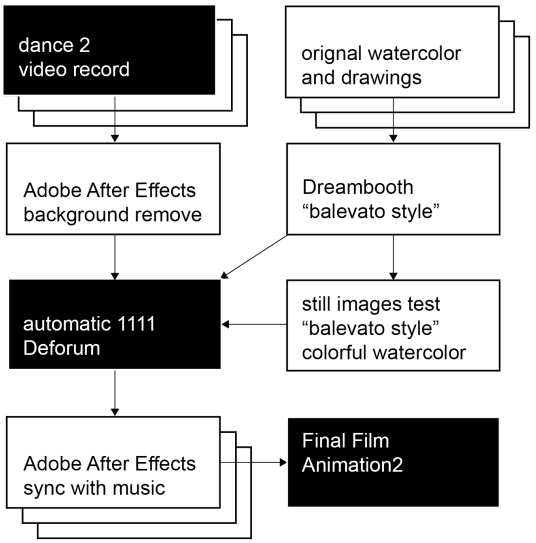

Figure 4: workflow schema

70 drawings and watercolors were selected for training, including old sketches and paintings. This material and its choice were intended to formulate a style to be replicated in the creation of the animation frames. It is important to emphasize that this process did not seek purity in results, but coherence in style. Here we present some of the images included in the creation of the style.

Figure 5: Some of the watercolors used for training

Further training an AI model using *DreamBooth,* it was powered by authorial drawings, especially watercolors and the model was personalized by the style created. For generation, the Automatic 1111 interface was used, a graphical interface that facilitates the operation of the model on a local computer.
After the training, some tests were carried out to observe the results of the training and, with generic prompts, specifying the technique. The results seemed, in some cases, consistent with the training, sometimes approaching a classic design, and perhaps saturating the colors, but seemed to indicate that the training had worked even if contaminated by other references in the model.

Figure 6: Test results for generating images from prompts

The video image treatment chosen was based on tests using static images of film frames, as a layout. The treatment chosen was colorful, with a cut-out background, in black. In a way, the error in cropping strengthened the organic result.
From the experiments it would be assumed that the result could be contaminated by the base, both in what a "watercolor" is and in its characteristics. But the desired watery look seemed to be working, even if randomly.
The images were generated following the following settings: *Prompt : "balevato style" watercolor colorful / Size: 1024 × 1024 / Guidance scale:9 / Strength: 0.66 / Steps: 45 / Seed: 2082233225 / Negative prompt: Disfigured, cartoon, blurry, photo, / Diffusion sampler: DDIM / Model: balevato / Created: June 4, 2023 2:05 AM*

Figure 7: Different styles generated with random seeds used to choose the final appearance

These images were generated on the same platform where the training was carried out, then the model was downloaded and taken to a local computer and through the Automatic 1111 interface a sequence of frames was generated using the img-to-img function, where it was also added the negative prompt for faces.

Figure 7: Final Look

3. Considerations

For processes such as animation, both two-dimensional and three-dimensional, the use of AI can be generous in reducing mechanical processes, increasing representation and producing at scale.
Training can enrich the possibilities of visual expression, avoid biases and make the use of AI for animation more authorial.
This text has a short delay between its writing and the experiment, the evolution towards perfection in representation is quite fast. At this moment, AI models for video production are emerging and images are beginning to reach hyper-realism. As well as video guides that, through controlnet, understand the pose, observing the skeleton through computer vision.
But when observing the use of models like Deforum, we conclude that this provides a new style of rotoscoping, with transformations and purposeful changes, which are the result of the options offered by the model and the iteration

between prompt and AI model. This ends up being more visible in the *Animation1* experiment. Thinking in this way, iteration, diffusion, transformation and lack of consistency are perhaps some of the characteristics of animations generated by AI. Just as Max Fleischer's rotoscope changed the animations of its time, the use of AI will probably change the animations of the coming years.

Acknowledgements

Agency, Lab, People,

References

[1] Pointer R. The Art and Inventions of Max Fleischer: American Animation Pioneer. McFarland; 2017.

[2] Paula Toller All Over (Feat. Donavon Frankenreiter). 2020.

[3] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models 2022. https://doi.org/10.48550/arXiv.2112.10752.

[4] Stability AI. Stability AI n.d. https://stability.ai (accessed January 19, 2024).

[5] Runway - Advancing creativity with artificial intelligence. n.d. https://runwayml.com/ (accessed January 19, 2024).

[6] CompVis - Computer Vision and Learning LMU Munich. GitHub n.d. https://github.com/CompVis (accessed January 19, 2024).

[7] Ruiz N, Li Y, Jampani V, Pritch Y, Rubinstein M, Aberman K. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation 2023. https://doi.org/10.48550/arXiv.2208.12242.

[8] AUTOMATIC1111. Stable Diffusion Web UI 2022.

[9] Zhang L, Rao A, Agrawala M. Adding Conditional Control to Text-to-Image Diffusion Models. arXivOrg 2023. https://arxiv.org/abs/2302.05543v3 (accessed January 18, 2024).