IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI ROTOSCOPING

[1]Mohammed Farhan M R, [2]Jayanth B S, [3]Kiran Kumara D G, [4]Faseeh Akbar, [5]Shilpa N Borkar

[1,2,3,4]Student Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka, India.

[5]Professor, Department of Computer Science and Engineering, AMC Engineering College, Bengaluru, Karnataka, India.

***Abstract**:* AI Rotoscoping is a transformative project that introduces cutting-edge artificial intelligence (AI) technologies to revolutionize the traditional rotoscoping process in video animation. Rotoscoping, a fundamental aspect of animation and visual effects, involves meticulously tracing objects frame by frame in live-action footage to create compelling animations. However, this process is notorious for its time-consuming and labor-intensive nature, often impeding the efficiency and creativity of animators. Our project addresses these challenges by integrating advanced AI algorithms to automate and enhance the rotoscoping workflow. The AI Rotoscoping system employs deep learning models to intelligently track and separate objects from the background in video frames, significantly reducing the manual effort required in the traditional rotoscoping process. This innovative approach empowers animators and visual effects artists to allocate more time and energy to creative aspects rather than getting bogged down in meticulous frame-by-frame detailing. Key features of the AI Rotoscoping project include precise object segmentation, leveraging state-of-the-art computer vision techniques. The system employs advanced motion tracking algorithms to ensure accurate object movement across frames, maintaining consistency and improving overall workflow efficiency. Semantic understanding is incorporated into the AI model to distinguish between various elements in the scene, reducing errors and enhancing the accuracy of the rotoscoping process.

Index Terms **- Rotoscoping, Video Segmentation, Manifold Models, Shape.**

I. INTRODUCTION

AI rotoscoping utilizes artificial intelligence algorithms to automate the process of creating matte images or masks that isolate objects in video footage. By leveraging machine learning techniques, AI rotoscoping accelerates the tedious manual tracing process, allowing for efficient and accurate segmentation of foreground elements from the background. This technology enhances productivity in visual effects and animation production, enabling artists to focus on creative tasks while achieving high-quality results in less time. AI rotoscoping offers a promising solution for streamlining post-production workflows and improving the efficiency of compositing tasks in the film, television, and animation industries. In the realm of visual effects and animation, the process of rotoscoping stands as a critical and intricate task, demanding precision and efficiency in isolating objects or characters from their background. As technology continues to evolve, the integration of artificial intelligence (AI) into rotoscoping workflows has emerged as a transformative frontier, promising to redefine the very essence of this intricate craft. This project embarks on the ambitious journey of developing an AI-driven rotoscoping system, leveraging advanced neural network architectures and innovative techniques to surpass the limitations of conventional approaches. The motivation for this endeavor stems from the recognition of the challenges inherent in traditional rotoscoping methods, which often entail labor-intensive manual processes and may struggle with the complexities of diverse scenes, intricate motion patterns, and varying lighting conditions. By harnessing the power of AI, specifically tailored for rotoscoping tasks.

AI rotoscoping revolutionizes the labor-intensive task of creating matte images or masks in video editing and animation by harnessing the power of artificial intelligence (AI). Traditionally, rotoscoping involves manually tracing over objects frame by frame to isolate them from the background. This process is time-consuming and requires significant human effort. However, with advancements in AI and machine learning, automated rotoscoping algorithms have emerged, promising to streamline the workflow and improve efficiency in visual effects and animation production. AI rotoscoping algorithms leverage deep learning techniques to analyze video footage and automatically generate matte images that accurately separate foreground elements from the background. These algorithms are trained on large datasets of annotated videos, learning to recognize and segment objects based on their visual characteristics such as color, texture, and motion.

By learning from examples, AI rotoscoping models can generalize their understanding of object boundaries and produce precise matte images with minimal human intervention. In addition to its efficiency and versatility, AI rotoscoping seamlessly integrates into existing post-production workflows, offering a user-friendly interface that streamlines the adoption process for industry professionals. Many AI rotoscoping tools are available as standalone applications or integrated modules within popular video editing and animation software, facilitating effortless incorporation into established pipelines.

However, despite its transformative potential, AI rotoscoping is not without its challenges. Complex scenes featuring overlapping elements or intricate motion patterns may pose difficulties for AI algorithms, necessitating human intervention to ensure optimal results. Furthermore, the acquisition of annotated training data presents a significant hurdle, as it requires considerable time and resources to compile datasets of sufficient quality and diversity.

Nonetheless, as AI rotoscoping continues to evolve, these challenges are expected to diminish, paving the way for even greater advancements in the field of video editing and animation. In summary, AI rotoscoping represents a monumental leap forward in the quest for efficiency, accuracy, and creativity in visual effects and animation production, promising to revolutionize the way artists approach the creation of matte images and masks in the digital age

Delving into the intricacies of data preparation, model architecture, and real-time processing optimization, we aim to provide a holistic understanding of the project's multifaceted approach. The integration of contour detection algorithms, such as the renowned Canny edge detector, adds an extra layer of sophistication to the AI rotoscoping system, enhancing its ability to capture detailed object boundaries with finesse. As we navigate through the technical intricacies, the user-centric design philosophy of the project takes center stage. A user interface that seamlessly integrates the prowess of AI with the creativity of human artists is envisioned, fostering a collaborative workflow that empowers users to interact, provide feedback, and actively participate in the refinement of the rotoscoping process. The project's commitment to ongoing testing, validation, and user support underscores its dedication to delivering not just a technological solution but a practical and user-friendly tool that resonates with the diverse needs of industry professionals. Against the backdrop of the ever-evolving landscape of visual content creation, this AI-driven rotoscoping project emerges as a beacon of innovation, seeking to push the boundaries of what is achievable in the intricate realm of object isolation and motion tracking. The project's significance lies not only in its technical advancements but in its potential to liberate artists and content creators from the constraints of time-consuming manual processes, unlocking new dimensions of creative expression. As we venture deeper into the project's core methodologies, the fusion of artificial intelligence with contour detection algorithms takes center stage. The inclusion of the renowned Canny edge detector is not merely an enhancement but a strategic choice to imbue the AI system with an acute awareness of object boundaries, breathing life into the minutiae of visual elements. This synergy of technologies is poised to redefine the very essence of how we perceive and execute rotoscoping, promising a leap forward in terms of precision, adaptability, and the seamless integration of AI into the creative process. Furthermore, this project stands at the intersection of technological prowess and human intuition. It is not merely about automating a process but about crafting a tool that resonates with the artistic sensibilities of those who wield it. The envisaged user interface becomes a canvas where the marriage of AI sophistication and human creativity is celebrated, offering an intuitive space for collaboration and iterative refinement. The collaborative workflow, empowered by the project, transforms the traditional paradigm, placing the artist in the driver's seat with AI as a capable and adaptable co-pilot. This AI-driven rotoscoping project aims to revolutionize the visual effects and animation industries by seamlessly integrating advanced artificial intelligence techniques, including contour detection algorithms like the Canny edge detector, into the rotoscoping workflow. The primary goal is to address the limitations of traditional rotoscoping methods, such as labor-intensive manual processes and challenges in handling diverse scenes and complex motion patterns. The project's foundation lies in a robust dataset that encompasses a wide spectrum of scenes, lighting conditions, and motion patterns. By leveraging deep learning architectures tailored for semantic segmentation, such as U-Net or DeepLab, and incorporating contour detection algorithms, the project seeks to achieve a new level of precision in separating foreground elements from their backgrounds.

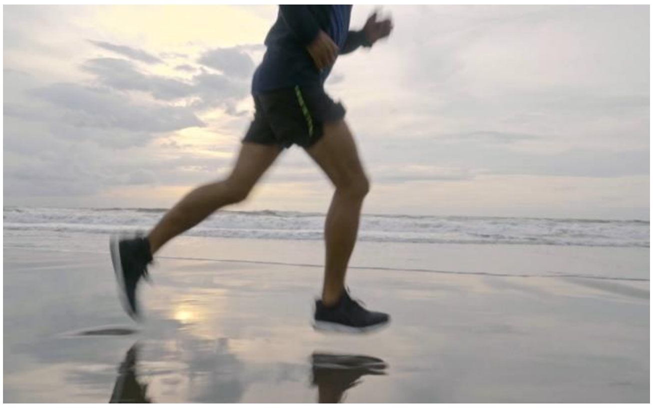

Figure1. Sample image

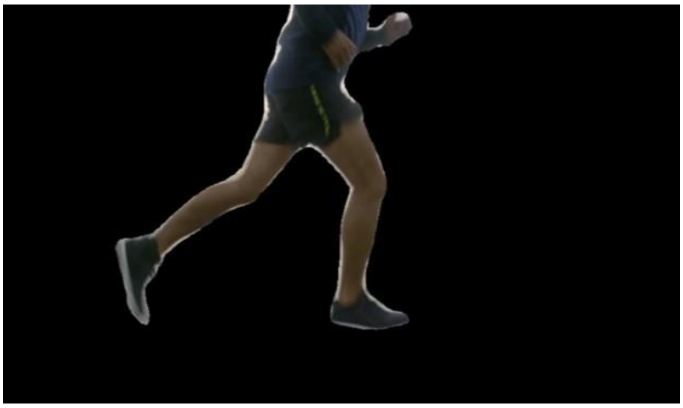

Figure2. Rotocoped image

Rotoscoping is a technique used in animation and visual effects to trace over live-action footage frame by frame. It involves manually creating a matte or mask around objects or characters in a video to separate them from the background. This process allows for the integration of animated elements or special effects into live-action scenes. Rotoscoping can be used for various purposes, including character animation, adding visual effects like explosions or creatures, or even for stylistic effect in filmmaking. Traditionally done by hand, rotoscoping is labor-intensive and time-consuming, requiring skilled artists to ensure accuracy and quality. However, advancements in technology, such as AI and machine learning, are being utilized to automate or assist in the rotoscoping process, speeding up production while maintaining precision. Despite technological advancements, rotoscoping often requires a combination of automated tools and manual refinement to achieve the desired results, blending the efficiency of automation with the artistic touch of human intervention.

Matte images and color grading Rotoscoping is a technique used in animation and visual effects to create matte images or masks that isolate specific elements within a scene. Matte images produced by rotoscoping serve as a fundamental component in compositing, allowing artists to separate foreground objects from the background and manipulate them independently. In rotoscoping, artists trace over live-action footage frame by frame to create precise outlines of objects or characters. This manual process ensures accurate delineation of complex shapes and movements, resulting in high-quality matte images. These mattes define the areas where visual effects or animations will be applied, enabling seamless integration of computer-generated elements with live-action footage. Matte images generated through rotoscoping are typically black and white, with the foreground object represented in white and the background in black. This binary representation simplifies the compositing process, as it provides a clear distinction between the foreground and background elements.

The accuracy and detail of matte images produced by rotoscoping are crucial for achieving realistic visual effects. By precisely defining the boundaries of objects, rotoscoped mattes allow for precise adjustments in lighting, color grading, and special effects. They also facilitate complex visual manipulations, such as object removal, replacement, or enhancement, while maintaining the integrity of the original scene. Rotoscoping is a labor-intensive process that requires skilled artists to ensure smooth and accurate outlines, especially for scenes with intricate motion or fine details. While advancements in software tools have automated certain aspects of rotoscoping, the human touch remains essential for achieving high-quality results.Overall, matte images produced by rotoscoping play a vital role in the post-production pipeline, enabling filmmakers and animators to seamlessly blend live-action footage with digital elements to create compelling visual experiences. Through meticulous tracing and refinement, rotoscoped mattes provide the foundation for realistic and immersive storytelling in film, television, and other visual media.

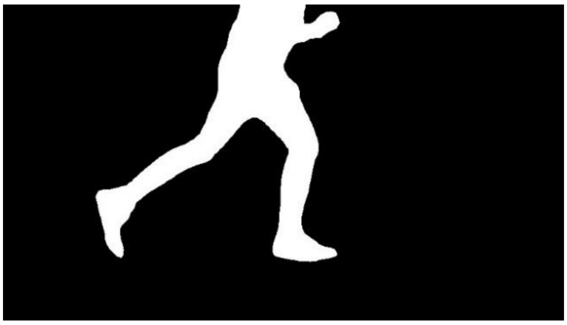

Figure 3.Matte image

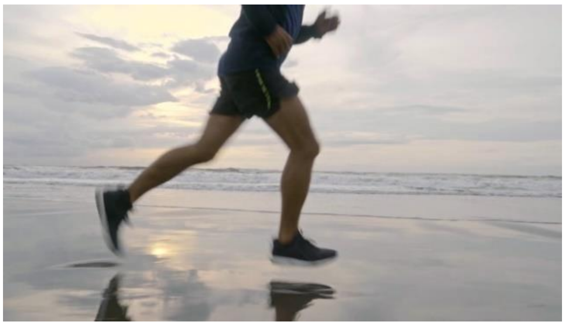

Figure 4.color grading

Color grading through rotoscoping is a technique used to selectively apply color adjustments to specific elements within a scene. Unlike traditional color grading methods that affect the entire image uniformly, rotoscoping allows for precise control over color correction and enhancement by isolating individual objects or characters.In the context of color grading, rotoscoping involves creating matte images or masks that define the areas where color adjustments will be applied. Artists manually trace over the elements of interest frame by frame, ensuring accurate delineation and preserving fine details. These matte images serve as guides for applying targeted color grading effects, allowing for nuanced adjustments tailored to specific parts of the scene. Rotoscoping-based color grading offers several advantages over conventional methods. Firstly, it enables selective color manipulation, allowing artists to enhance or alter the appearance of particular objects or characters without affecting the entire image. This level of control is particularly useful for emphasizing focal points, enhancing visual storytelling, or achieving stylistic effects. Additionally, rotoscoping facilitates complex color grading tasks that may be challenging or impractical to achieve with traditional techniques. For example, artists can isolate moving objects against dynamic backgrounds and apply customized color treatments to enhance their visibility or integrate them seamlessly into the scene.

Furthermore, rotoscoping-based color grading allows for precise adjustments in situations where automated or global color grading algorithms may produce undesirable results. By manually defining the boundaries of objects, artists can ensure that color corrections are applied only to the intended areas, minimizing artifacts and preserving visual coherence. However, it's important to note that rotoscoping for color grading is a labor-intensive process that requires time, skill, and attention to detail. Artists must meticulously trace each frame to achieve smooth outlines and accurate matte images, which can be particularly challenging for scenes with complex motion or fine details.In summary, color grading through rotoscoping offers a powerful method for selectively enhancing and manipulating the colors of specific elements within a scene. By providing precise control over color adjustments, rotoscoping-based color grading enables artists to achieve tailored visual effects and enhance the overall impact of their work. This AI-driven rotoscoping project aims to revolutionize the visual effects and animation industries by seamlessly integrating advanced artificial intelligence techniques, including contour detection algorithms like the Canny edge detector, into the rotoscoping workflow.

II. LITERATURE SURVEY

The research paper by Luis Bermudez and team [1] proposes an innovative method for enhancing the rotoscoping process in multi-shape systems through a learning-based approach. The focus of the research is on developing parametric models that can adapt to the complexities of multiple shapes within a given scene. The approach incorporates machine learning techniques to automatically learn and refine parameters, improving the accuracy and efficiency of the rotoscoping process. By leveraging this learning-based strategy, the paper introduces a more adaptive and automated solution for handling intricate shapes in visual content, contributing to advancements in the field of rotoscoping.

The research paper by Wenbin Li and team [2] introduces an innovative framework, Roto++, designed to expedite and enhance the professional rotoscoping process. The key focus is on leveraging shape manifolds, mathematical representations of shapes, to accelerate the rotoscoping workflow. The proposed method aims to reduce the manual effort involved in tracing complex shapes by providing automated tools that exploit the underlying structure of the shapes. Through the integration of shape manifolds and advanced optimization techniques, Roto++ offers a more efficient and accurate solution for professional rotoscoping, ultimately contributing to increased productivity and improved results in visual content production.

The research paper by John Canny [3] explores methods and techniques for edge detection in images using a computational approach. The focus is on identifying boundaries within images by detecting abrupt changes in intensity. The paper delves into various computational algorithms, such as the Canny edge detector, and evaluates their effectiveness in accurately locating edges. Additionally, it discusses the significance of parameter tuning and the trade-offs involved in edge detection algorithms. The research contributes to the understanding of computational strategies for edge detection, providing insights into the challenges and considerations in implementing these techniques for image analysis and computer vision applications.

The research paper by Aseem Agarwala, Aaron Hertzmann, David H. Salesin, and Steven M. Seitz [4] explore keyframe-based tracking techniques specifically tailored for rotoscoping and animation tasks. Rotoscoping involves tracing over live-action film footage to create animated sequences, while animation involves creating movement and actions in digital environments. Keyframe-based tracking is crucial for accurately capturing the motion and nuances of objects or characters in these processes. Keyframes are pivotal frames selected by animators to represent significant moments in motion sequences.

The research paper by Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang [5] explores methods and techniques to propose a deep learning framework for human pose estimation that achieves high-resolution representation learning. Human pose estimation involves predicting the skeletal pose of a person from an image or video, which is crucial for various applications such as action recognition, human-computer interaction, and surveillance. It discusses the challenges associated with accurately estimating human poses from images or videos, including occlusions, complex poses, and variations in scale and viewpoint.

The research paper by Xuehan Xiong and Fernando De la Torre [6] explores methods and techniques to introduce and demonstrate the effectiveness of the Supervised Descent Method (SDM) for face alignment. Face alignment involves the process of locating facial landmarks, such as eyes, nose, and mouth, in images, which is a fundamental task in computer vision with applications in face recognition, facial expression analysis, and augmented reality. It discusses the challenges associated with accurately localizing facial landmarks in images, including variations in pose, illumination, and facial expressions. SDM is a cascade regression technique that learns a sequence of descent directions from annotated training data to iteratively refine the initial estimate of facial landmarks.

The research paper by Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang[7] explores methods and techniques to address the challenge of unconstrained face alignment, which involves accurately locating facial landmarks in images with varying poses, expressions, and lighting conditions. Face alignment is a fundamental task in computer vision with applications in face recognition, facial expression analysis, and augmented reality. It discusses the challenges associated with accurately localizing facial landmarks in unconstrained settings, such as large pose variations and occlusions. This hierarchical approach allows the model to capture both global and local variations in facial appearance. The research contributes to the understanding of computational strategies for edge detection, providing insights into the challenges and considerations in implementing these techniques for image analysis and computer vision applications.

III. PROPOSED METHOD

3.1 DESIGN

We designed a new tool to work with artists to accelerate the rotoscoping workflow while also overcoming some of the main limitations and respecting the artists' requirements.

Tracker Drift Tracking is known to reduce required keyframe count when it works well but the most common failure mode is for the tracker to drift in difficult situations. This yields roto-shapes which depart significantly in shape from the edited keyframes. The strong regularization of Agarwala et al. [2004] can help to prevent drift but limits the output to smooth interpolations of the keyframes.

Manifold Shape In order to get over this restriction, we suggest combining the tracker's output with a strong prior on the potential roto-shapes. From the artist's existing keyframes, we derive a statistical shape model. We then constrain the intermediate frames to be as close to the tracker output as possible while still being valid shapes from the shape model. This implies that the user will still see satisfactory results even if the tracker output deviates from the range of acceptable roto-shapes. Our shape model gets more and more accurate as the user adds more keyframes, which leads to increasingly better estimations for the intermediate roto-shapes. This model is represented as a low-dimensional, generative manifold. The original keyframes are points in this space and our hypothesis is that other regions of the manifold will generate distinct but similar shapes that are likely to include the correct shapes for the intermediate frames. Figure 3 provides an illustrative example of our shape manifold. A 2D manifold has been embedded with the roto-curve defining the lower arm, and the highlighted spots correspond to the locations that produce the distinct roto-shape in each frame. The line connecting the points together denotes the passage of time from frame to frame.

Choice of Manifold Model We use a Gaussian Process Latent Variable Model (GP-LVM) [Lawrence 2005] as part of our global model of the joint probability between the control points both within and between frames since it is generative, Bayesian and non-linear. Previous approaches only regularise trackers with local proxies such as smoothing over neighboring keypoints in time and space. Linear subspace models, such as ASMs, are trained with a large dataset. Since non-linearity is known to capture more variance in fewer dimensions, our strategy, which benefits from being Bayesian, is aimed at a small number of training samples (the keyframes) [Prisacariu and Reid 2011]. Our model actually subsumes linear models and is more general. Other models, such as [Agarwala et al. 2004], are neither Bayesian nor generative and therefore cannot be used to provide user suggestions or the intelligent drag tool.

Keyframe Recommendation Another advantage of constraining our predicted roto-shapes to come from our shape model is that we can identify frames when the tracking result departs heavily from our shape model. This could mean one of two things, either the tracker has drifted and needs to be reinitialized, or the shape model is not sufficiently well-defined to include all the valid shapes. Thankfully, both of these situations are remedied by the user labelling the frame as a keyframe. To make use of this result, we provide the artist with helpful feedback. We suggest which keyframe will most help the most to improve the tracker, the shape model, or both to produce better interim shapes.

To update the shape manifold it is necessary to know when a new keyframe is added. We therefore asked the user to formally specify when they have finished editing a keyframe. When a new keyframe is added, the shape model is updated and the tracker is recalculated, all in realtime. This improves the unedited curves and updates the frame recommendation.

3.2 User Interface

We implemented an interactive tool, Roto++, to evaluate our approach, as shown in Figure 4. The interface aims to provide a familiar look and feel to existing tools while adding our advanced functionality behind the scenes Our Bezier curve based tool is made up of ´ of a Timeline, a Design View, and a number of Roto and Design Tools combined.

- Design Tools: a common subset of curve-drawing operations of leading commercial software.

- Timeline: a thumbnail-sized view of the shot. Color coding on a thumbnail's border shows the status of that frame. We also use this view to indicate to the artist our recommendation for the next frame to edit to improve the result.

Instrumentation To provide a detailed evaluation of our method, the tool is highly instrumented to maintain a detailed log of all the user operations performed down to the level of individual mouse operations. Accurate timestamps were recorded allowing us to determine the time spent performing different operations. We also logged the current state of the roto-curves periodically to allow us to determine the accuracy of the roto output as a function of artist time expended.

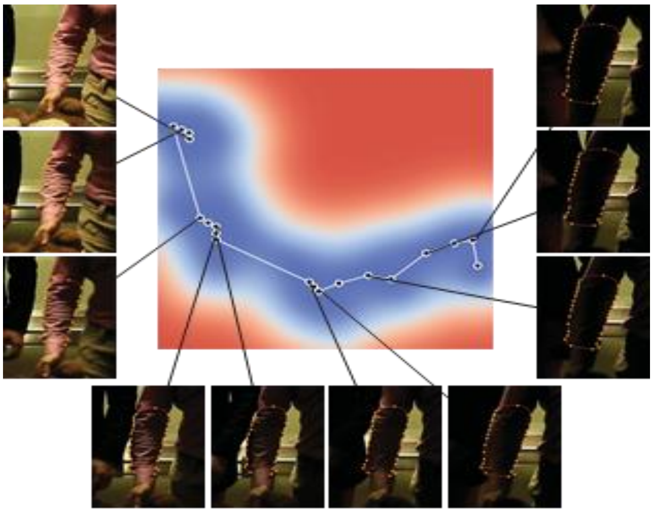

Figure3: *An example manifold for the arm roto sequence. As we move across the manifold the shape of the arm changes smoothly. Even though the roto-curve contains 87 control points (to account for the ripples in the shirt) the sequence can be perfectly recoveredfrom a 2D shape manifold. (Note: we observe a complete change in object appearance as the character moves from light to shadow; this sort of sequence represents a significant challenge to techniques that track edges or make use of color models).*

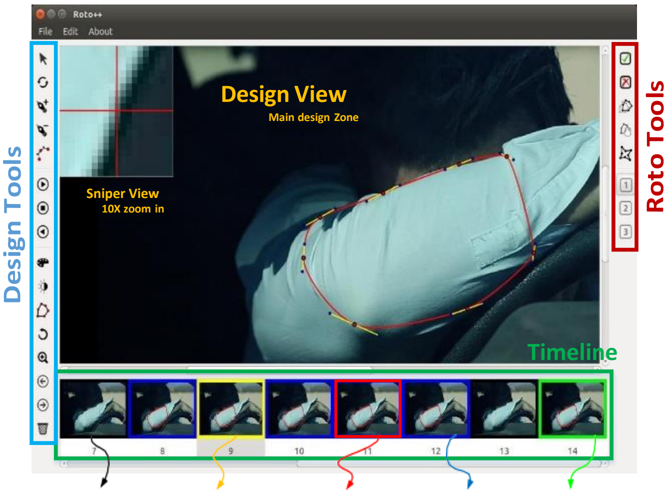

Figure 4: *A screenshot of our interactive tool* ***Roto++****. This user interface consists of a design view, design tools, roto tools and a timeline. This timeline encodes our frame selection feedback by using colored borders. A full user guide to our tool is included in the supplemental material*

3.3 Interactions

The features presented in Section 3.1 aim to reduce the number of key frames that an artist needs to supply to produce an accurate output. While this is clearly advantageous, it is only half the story. Furthermore, we would like to spend less time labeling each key frame in order to increase efficiency even more. Beating the baseline is a challenging task since artists are highly trained and have, in many cases, years of experience using these techniques. Furthermore, they have exacting requirements on interaction the most relevant are the need for intuition and instantaneous feedback combined with predictable results that do not corrupt previously edited results. The limiting factor on the editing time of a key frame for baseline methods occurs when deformable shape changes occur that cannot be accurately tracked. In most cases, the artist is then forces to edit the roto-curve control points or tangent vectors in small groups or individually. This can require a very large number of mouse operations to select the points in turn and move them to their correct locations.

Intelligent Drag Mode To assist with this editing, we can once again exploit our shape manifold. Once the manifold has been trained, new shape proposals can be generated very efficiently from locations on the manifold. This means that we can generate new sets of plausible shapes to match any input from the user. As the user selects a point on the curve and drags it, our solver uses the manifold to suggest the new best fitting shape in real time.

We give the artist additional freedom when using Roto++ by allowing them to first choose which control points on the curve they want to update. The other points will remain fixed ensuring the requirement that if a user is happy with part of the outline, another editing operation will not corrupt it. Once the points to move are selected, they can be dragged to a new location. In real time we run a cut-down version of our tracking solver which replaces the tracker result with the new curve location under the drag operation and then solves for the new shape recommended by the manifold. Figure 1 shows the same result achieved by a large number of control point moves being performed in a single operation with our intelligent drag tool. The operation of the tool is perhaps more clearly demonstrated in the video included in the supplemental material.

IV. Technical Approach

In this section we describe how we implement the new features in our Roto++ tool described in Section 3. We first describe the notation used throughout this section. Subsequently we detail how we obtain our shape manifold from a set of keyframes edited by the artist. We then provide details on how to combine our shape manifold with an existing tracker to create the Roto++ solver that estimates the roto-shapes for the unlabeled frames. Finally we show how this solver may be used to provide the intelligent drag tool.

4.1 Notation

In Appendix A details the notation used in this section. Throughout, we assume that we are operating on a single, closed roto-curve in a single shot. Our results can be applied in parallel to multiple roto-curves in a straight forward fashion. We also assume that there is no distinction between an interpolation control point and an tangent control point on a Bezier curve. Our closed roto-curves ´ are made up of a closed sequence of cubic Bezier curves, each of ´ which contains 4 control points with the last control point of the curve forming the first control point of the subsequent curve. The k^th^ keyframe spline is denoted U_k where

$$U_k = \begin{bmatrix} U^k_{x1,} U^k_{x2,} U^k_{x3,} \dots \dots U^k_{xn,} \\ U^k_{y1}, U^k_{y2}, U^k_{y3} \dots \dots U^k_{yn} \end{bmatrix}$$

This is seen in Appendix A's Figure 13. The output splines are given as {Yn}. Note that for consistency, we enforce that Yk = Uk for all keyframes k; we would like to produce good estimates for the remaining output splines. All the keyframe splines must have the same number of control points (M). This is straight forward to maintain; if the artist would to add a new control point on any keyframe, the appropriate Bezier ´ curve is subdivided at the same parametric location in all other keyframes. The manifold may then be recomputed with the increased number of control pointsIt is important to note that we need a minimum of two keyframes (ideally three) before we can begin to construct a shape manifold. For this reason, the artists are asked to produce keyframes for the first and last frame before progressing. This allows the shape model to initialize; before these keyframes are present the system operates without the shape.

4.2 Overview

Our solver may be broken down into three stages. We first estimate a rotation, translation and scale for each key frame spline. We then remove this rigid body transformation to a produce a set of normalized key frame splines which are aligned with one another. With our shape manifold, we wish to capture the changes in deforming shape, which are now the only variations between the normalized splines.

Next, we fit a generative manifold model to the geometry of the normalized key frame splines. This model embeds the high dimension spline data (the collection of all the Bezier control points) into ´ a very low dimensional space such that every location in the low dimensional space interpolates and extrapolates from the key frames to a distinct shape. Furthermore, this low dimensional space is smooth such that smooth changes in the manifold space represent smooth changes in the high dimensional splines. Finally, once we have learned the manifold model we are able to run our solver. Here, we may use the form manifold and tracking data from any source to create a reliable output with smoothly varying representative shapes for the roto-shapes., even when the tracking fails. This allows the user to insert key frames and produce their desired shape quickly, even when parts of the spline are unable to track edges or other image features.

4.3 Keyframe Alignment

We align the keyframe splines by estimating a rotation θk, translation tk, and scale sk for every keyframe. We do this to a high degree of accuracy by using an energy model to estimate the transformation and a reference shape at the same time as a generalized Procrustes analysis problem. We denote the mean reference spline as R. For each keyframe k, our alignment energy is

$$E_{align}^{(k)}(U_k, \theta_k, t_k, s_k, R) = \sum_{m=1}^{M} \left[\begin{bmatrix} U_{xm}^{(k)} \\ U_{ym}^{(k)} \end{bmatrix} - \left(s_k Q_k \begin{bmatrix} R_{x,m} \\ R_{y,m} \end{bmatrix} + t_k \right) \right]$$

where Q_k is the 2D rotation matrix,

$$Q_k = \begin{bmatrix} \cos(\theta_k) & -\sin(\theta_k) \\ \sin(\theta_k) & \cos(\theta_k) \end{bmatrix}$$

over all keyframes to find the optimal $\{\theta_k, t_k, s_k\}$. We initialize with the mean shape and linear estimates for the transformation variables before applying the non-linear least squares Ceres solver directly to Equation 4 using the Gauss-Newton L-BFGS method.

V. ALGORITHM AND EXPLANATION

5.1 Canny Edge Detection Algorithm

Canny edge detection is a multi-stage algorithm employed in image processing to identify and accentuate the edges present within an image. The algorithm begins with a preprocessing step where the image is subjected to Gaussian smoothing, using a convolution operation with a Gaussian kernel. This step helps reduce noise and unwanted details in the image.

Subsequently, the gradient of the smoothed image is computed using convolution with derivative kernels. The gradient magnitude and direction for each pixel are then determined. The gradient magnitude represents the rate of change of intensity, highlighting areas with significant transitions. Non-maximum suppression is a critical step following gradient computation. In this stage, the algorithm examines the gradient magnitudes and retains only local maxima along the edges while suppressing non-maximum values. This effectively thins the edges, preserving only the most pronounced ones.

To trace and connect the edges accurately, hysteresis thresholding is employed. Two thresholds, a high threshold (strong edge) and a low threshold (weak edge), are utilized. Strong edges are those pixels whose gradient magnitudes are higher than the high threshold. Pixels with magnitudes between the high and low thresholds are considered weak edges. To ensure connectivity in the final edge map, weak edges are included if they are adjacent to strong edges; otherwise, they are suppressed.

Canny edge detection is highly regarded for its ability to produce well-defined, continuous edges while suppressing noise and spurious details. Its adaptability to various imaging conditions, ability to handle complex scenes, and capacity to produce thin, accurate edges make it a cornerstone in computer vision and image analysis applications, including edge-aware filtering, feature extraction, and object recognition.

```
 # defining the canny detector algorithm
def Canny_detector(img, weak_th = None, strong_th = None)
```

5.2 Fully Convolutional Networks(FCN)

Fully Convolutional Networks (FCNs) are a class of deep learning models mostly employed in computer vision applications involving semantic segmentation. Unlike traditional convolutional neural networks (CNNs) that are designed for classification, FCNs are tailored for pixel-level prediction tasks, where each pixel in an input image is classified into different categories.

FCNs leverage convolutional layers to process the entire input image and generate output feature maps, preserving spatial information. The key innovation of FCNs lies in their ability to produce dense predictions by employing transposed convolutions or upsampling techniques to recover spatial resolution. This allows FCNs to produce output masks that match the input image dimensions, enabling pixel-wise classification.

The architecture of FCNs typically consists of an encoder-decoder structure. The encoder comprises multiple convolutional and pooling layers, which progressively reduce the spatial dimensions of the input image while capturing hierarchical features. This encoder extracts abstract representations of the input image, preserving spatial information through feature maps.

The decoder component of FCNs consists of upsampling layers, sometimes combined with skip connections, to recover the spatial resolution of the feature maps produced by the encoder. Upsampling techniques like transposed convolutions or interpolation are employed to upscale the feature maps to the original input image 3+size. Skip connections, borrowed from architectures like U-Net, connect corresponding encoder and decoder layers to preserve fine-grained details during upsampling.

During training, FCNs are optimized using loss functions tailored for segmentation tasks, such as cross-entropy loss or variants like Dice loss. These loss functions compare the predicted segmentation masks with ground truth annotations, encouraging the model to produce accurate pixel-level predictions.

FCNs have been widely applied in various computer vision tasks, including image segmentation, object detection, and image-to-image translation. Their ability to generate dense predictions makes them particularly effective for tasks where precise spatial localization is crucial, such as medical image analysis, autonomous driving, and scene understanding.

One of the notable applications of FCNs is semantic segmentation, where the goal is to classify each pixel in an image into predefined categories, such as object classes or semantic regions. FCNs have also been extended to handle instance segmentation, where the task involves not only categorizing pixels but also distinguishing between different instances of the same category.

Fully Convolutional Networks have become a powerful and widely used tool for various computer vision tasks that require dense prediction at the pixel level. Their ability to preserve spatial information and produce output maps with the same spatial dimensions as the input image makes them well-suited for tasks such as semantic segmentation, image-to-image translation, and dense object detection.

In summary, Fully Convolutional Networks (FCNs) are deep learning models designed for pixel-level prediction tasks, particularly semantic segmentation. By leveraging convolutional layers and upsampling techniques, FCNs can produce dense predictions that preserve spatial information, making them suitable for various computer vision applications requiring precise localization and understanding of image content.

VI. EXPERIMENTS

6.1 Real World Rotoscoping Dataset

To evaluate our method we produced an extensive rotoscoping dataset specifically designed to be truly representative of real-world commercial rotoscoping in the post-production industry. The dataset consists of a five minute short movie, that has been professionally rotoscoped by a services company that works on preparation and compositing for high-end VFX and stereoscopic conversion across film, television and commercials. This dataset will be made available under the Creative Commons license agreement (by-nc-sa)1 for non-commercial use. The short movie depicts a story unfolding around a night club and contains shots typical to live-action movies. This illustrates the spectrum of rotoscoping intricacy that one could anticipate in the post-production of a live-action movie. The footage covers the space of rotoshapes that would be required, from simple rigid objects, to isolated articulation motion, more complex articulations with occlusions and intersections, and thin structures for close-ups of hair. The scene content and camera effects cover locked-off and hand-held camera moves, shifts in focus and motion blur, close-up and wide-angle shots, and bright and dark environments. They also include isolated and interacting characters, plus complex with water.

Errors All of the shots are recorded in HD and a typical error of a few pixels is within the motion blur of the majority of scenes at this resolution. For professional artists to achieve sub-pixel accuracy for this footage would require feathering to be used which we do not currently have data for. We discuss this as future work.

We can be sure that the roto-splines will be parameterized in the same way because we utilize the same input keyframes (from the dataset's ground truth) for every technique. This allows us to evaluate the error by considering the RMS pixel error between the control points of the estimated curves and the ground truth splines.

Complexity	Typical Shot Description	Rating
Easy	· Single isolated characters · Trackable objects	1
	· Simple manual keyframing	2
Medium	· Limited motion blur · Limited articulation · Several characters	3
Hard	· Lengthy camera shots · High-speed shots with motion blur	4
	· Many characters with detailed articulation · Detailed shapes for hair / fur	5

Table 1: Complexity examples for different rotoscoping shots. Ratings and descriptions provided by professional artists.

Complexity Rating	Number in Dataset	Rotoscoping effort
2	38 shots	28 frames / day
3	94 shots	14 frames / day
4	14 shots	11 frames / day
5	12 shots	6 frames / day

Table 2 : Breakdown of the complexity and rotoscoping effort required in the rotoscoping dataset. The total effort for the entire dataset was 734 person days. Ratings and effort provided by professional artists.

6.2 Expert Study

To evaluate the Intelligent Drag Tool and our Roto++ tool as a whole we conducted an Expert Study. The aim was to investigate performance with respect to two baseline workflows in a real-world situation with experienced rotoscoping artists and shots from a commercial movie. Figure 8 provides an overview of the expert study that we will now discuss in more detail. First we will describe the protocol used and then we will analyze the quantitative results.

Evaluation Protocol We invited seven professional roto-artists, from movie post-production houses, to take part in the study; the artists all had between two and nine years of rotoscoping experience. Each team had a similar distribution of experience to allow for fair comparisons between the teams. Our Roto++ tool was run in three different modes, Mode 1 presents our method; Mode 2 denotes the Blender planar tracker; and Mode 3 is the linear interpolation. The artists were unaware of the technical details of any of the differences between any of the modes. In addition to our solver, Mode 1 also made the Intelligent Drag Tool available. The roto artists were instructed on how to use it but they were free to use it or not during the study. Similarly for the next keyframe suggestions.

Shape Interaction Using our instrumentation we were able to measure various mouse (pen and tablet) operations. We observe that our solver and intelligent drag tool require fewer mouse operations and achieve a greater accuracy. This is due to the value of the intelligent drag tool moving the control points to the correct location in far fewer moves than editing individual control points.

Rotoscoping Time depicts the point error over time for all baselines on the three different shots. Across all three tests (differing complexities) our solver improves over the baselines. We also note that an earlier version of our solver was used for the expert study that occasionally produced a lag in response (this was commented on by the experts when they were debriefed). We have since improved the solver speed by an order of magnitude and no have no lag in subsequent tests. The dashed line on the plots represents a replay of the log with the solver times set to the new speed; this leads to an additional improvement in performance, really demonstrating the value of our tool.

Scale-Aware Loss Function Adding point and part specific scaling improves performance by 12.3% over ignoring the scale in the loss function (see Supplementary Figure 10 and Table 4). The majority of the benefit is absorbed by the seams and eyebrows (which result in an average MAE of 2.95 and 2.67 when the model is not trained with scaling parameters, respectively, versus 2.45 and 2.15 respectively when the model is trained with scaling parameters), while the eyes do not appear to benefit. This could be due to greater invariance of the eye shape to pose and facial expression. The seams also benefit from point scaling because we can overweight their impact on the loss function.

AI rotoscoping, an emerging field at the intersection of artificial intelligence and visual effects, has garnered significant attention from researchers and practitioners alike. Numerous studies have explored various aspects of AI-driven rotoscoping, ranging from algorithm development to applications in industry settings. These studies aim to advance the state-of-the-art in rotoscoping techniques, improve efficiency, and enhance the quality of visual effects and animation production.

One prominent area of research in AI rotoscoping focuses on algorithm development and optimization. Researchers have proposed novel deep learning architectures tailored specifically for rotoscoping tasks, aiming to improve segmentation accuracy and efficiency. For example, studies have investigated the use of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for semantic segmentation of video footage, demonstrating promising results in automating the rotoscoping process. These algorithms often leverage advanced techniques such as transfer learning and data augmentation to improve generalization performance and robustness across different types of footage.

VII. CONCLUSION AND FUTURE WORK

An AI based Rotoscoping System is developed which uses training models to provide a rotoscoped video. UI is developed for easier use and experience. Converts a labor-intensive work into an automated one which reduces the workload of artists Output video is generated along with matte frames.AI rotoscoping represents a transformative advancement in the field of visual effects and animation, offering unprecedented capabilities for automating the process of matte image creation. Through the integration of artificial intelligence and deep learning techniques, AI-driven rotoscoping algorithms have demonstrated remarkable efficiency, accuracy, and versatility, revolutionizing traditional workflows and enabling new possibilities for creative expression. From algorithm development to software integration and industry applications, research in AI rotoscoping has made significant strides in advancing the state-of-the-art and addressing practical challenges.future work in AI rotoscoping holds immense promise for further innovation and advancement. One avenue for future research is the development of more robust and adaptive algorithms capable of handling complex scenes with greater ease and accuracy. This may involve exploring advanced deep learning architectures, incorporating additional sources of information such as temporal context or semantic cues, and leveraging techniques from computer.

VIII. REFERENCES

[1] L. Bermudez, N. Dabby, Y. A. Lin, S. Hilmarsdottir, N. Sundararajan and S. Kar, "A Learning-Based Approach to Parametric Rotoscoping of Multi-Shape Systems," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 776-785, doi: 10.1109/WACV48630.2021.00082.

[2] Wenbin Li, Fabio Viola, Jonathan Starck, Gabriel J. Brostow, and Neill D. F. Campbell. 2016. Roto++: accelerating professional rotoscoping using shape manifolds. ACM Trans. Graph. 35, 4, Article 62 (July 2016), 15 pages. https://doi.org/10.1145/2897824.2925973

[3] J. Canny, "A Computational Approach to Edge Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, Nov. 1986, doi: 10.1109/TPAMI.1986.4767851

[4] X. Xiong and F. De la Torre Supervised descent method and its applications to face alignment. In 2013 IEEE, pages 532–539.

[5] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. ACM Trans. July 2018 Graph., 37(4).

[6] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. 2019 CoRR, abs/1904.04514.

[7] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[8] H. Zhang, Q. Li, Z. Sun, and Y. Liu. Combining data-driven and model-driven methods for robust facial landmark detection. IEEE Transactions on Information Forensics and Security, Oct 2018 13(10):2409–2422

[9] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, and et al. Lookingood: Enhancing performance capture with real-time neural re-rendering. Dec. 2018 ACM Trans. Graph., 37(6).

[10] Kai Ruhl, Martin Eisemann, and Marcus Magnor. Cost volume-based interactive depth editing in stereo post processing. In Proceedings of the 10th European Conference on Visual Media Production, CVMP ’13, New York, NY, USA, 2013. Association for Computing Machinery.