Machine Learning Video Mattes for Visual Effects

Samuel James Hodge
Adelaide University
Adelaide, South Australia, Australia
ORCID: 0009-0006-6597-5783
Email: samuel.hodge@adelaide.edu.au

***Abstract*—Participants from the visual effects industry were surveyed about the problems associated with matte creation using Rotoscoping in the production of video layers to create visual effects, without the use of chroma screens. Automation of video mattes is available in many forms, including many machine learning approaches, both commercial and open source. The human-centric approach of briefing, feedback, and human-editable output is required to reach an acceptable quality to use in high-end cinematic productions. Automatic approaches allow departments requiring a video matte to work with placeholder data, which will be later replaced by higher-quality human-curated output. As predictions from algorithms are very rarely acceptable for final use, the benefit of using an automatic temporary placeholder can release time in the schedule while handcrafted mattes are being prepared, but this can be negated by scheduling. In contrast, storytelling benefits from using placeholders in post-visualization and editing.**

***Index Terms*—Rotoscoping, Matte extraction, Image segmentation, Machine learning for VFX, Production workflows, Temporal consistency, Compositing, Visual effects pipelines**

I. Introduction

Manual delineation regions in moving images originated in cel animation with the rotoscope device patented by Max Fleischer in 1915 Fleischer (1915). In contemporary visual effects production, rotoscoping Bratt (2018) remains a foundational task in nearly every compositing workflow. Visual effects depend on accurate mattes to isolate actors, props, and environments. Using video mattes enables seamless integration of computer-generated and captured elements. Rotoscoping is conceptually straightforward; it is one of the largest, recurring, time-intensive tasks in post-production. Complex visual effects shots require tens to hundreds of artist-hours to achieve the level of precision demanded to meet quality control standards.

A high-quality matte must faithfully capture object boundaries, partial transparencies, motion blur, and defocus characteristics across tens to thousands of frames. These requirements demand sub-pixel precision and finely tuned artistic judgement, often requiring communication with the review quality control team, which adds to the expense. Spline-based, resolution-independent representations allow artists to retain full control and to iterate on refinements. As production volumes and turnaround

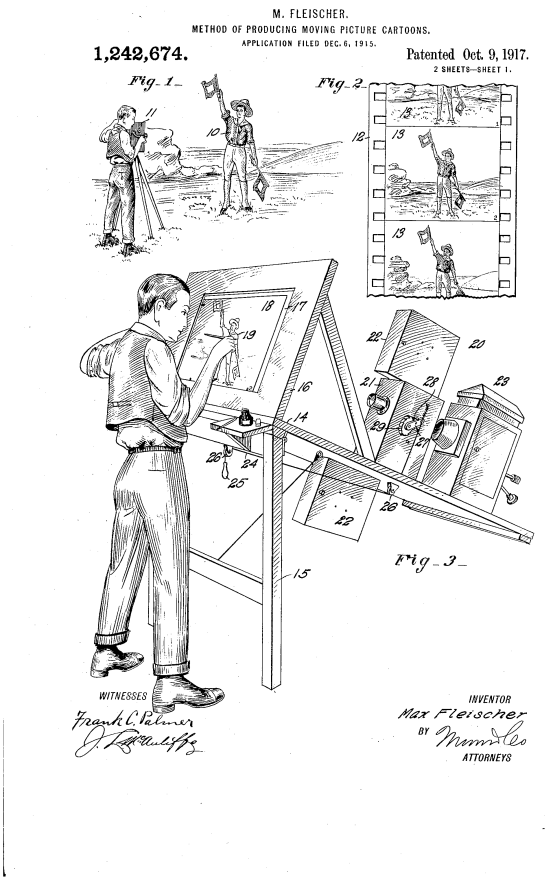

Figure 1: Historical Patent for the Physical Rotoscope Apparatus

expectations continue to grow, studios have sought solutions that reduce manual workload without compromising quality through machine learning automation.

Recent advances in computer vision—particularly deep learning approaches to image segmentation, matte extraction, and temporal propagation—have introduced new possibilities for automation. This automation can create a data footprint of video mattes of lower quality with poor temporal stability and edge fidelity. The ML systems have

limited avenues for human correction. As a result, manual methods are more reliable for the final output.

The degree of adoption of machine learning automation varies widely among participants and is influenced by economic benefit and technology limitations. Academic research has made significant progress in segmentation and video matting; we lack a survey about how these techniques perform in real-world production scenarios. The rationale for the lack of adoption is quality control, drivers, and barriers to adoption. Along with requirements for a system to be useful within visual effects workflows.

This paper presents a production-oriented survey addressing this gap. Drawing on 92 practitioner interviews across a diverse range of organisations, we analyse the role of machine learning in video matting workflows; identify the technical, artistic, and organisational challenges that shape machine learning adoption; and characterise the failure to meet industrial standards. Our goal is to bridge the disconnect between research-driven progress in automated matte extraction and the practical constraints of professional VFX pipelines, providing guidance for future tools that better align with industry needs.

II. Background and Related Work

A. Historical background

Rotoscoping as a technique for frame-by-frame tracing originated in early cel animation; Max Fleischer's rotoscope patent describes the fundamental idea of projecting and tracing live-action frames for subsequent animation workflows Fleischer (1915). Over the ensuing decades, the basic concept evolved into high-precision, artist-driven matte creation used in photographic compositing and film post-production.

B. Classical matting and spline-based mattes

Classical image matting techniques formulate the problem as per-pixel alpha estimation and include influential analytic approaches such as Levin et al.'s closed-form matting, which provides a principled regularisation for natural-image matting and remains a common baseline in the literature Levin et al. (2008). In production, artists commonly author mattes using temporally animated, resolution-independent spline primitives ("roto splines") that are rendered to per-frame grayscale alpha channels; these representations provide full controllability and iterative refinability at the cost of substantial manual effort.

C. Semi-automatic and production tools

Production workflows use a mix of spline-based authoring and semi-automatic tools that reduce manual labour through tracking and user-guided propagation. Planar trackers and sub-planar mesh trackers (for example, Mocha Pro) and dedicated roto/paint systems (for example, Silhouette and the Roto node in Foundry's Nuke) are ubiquitous in VFX pipelines because they pair artist intent with robust tracking primitives and exportable matte caches Boris FX (2024); Inc. (2022–2024). Semi-automatic video cutout systems such as VideoSnapCut provide examples from the academic literature that inspired later production features Bai and Wang (2009).

D. Machine learning approaches: image matting and video matting

Deep learning has driven a step-change in matting research. Early data-driven matting systems moved beyond colour-sampling priors to learn alpha from large paired datasets; examples include Deep Image Matting, which uses encoder–decoder architectures and explicit trimap-conditioning to predict high-quality alpha mattes Xu et al. (2017). For video, recent work focuses on temporal stability and spatio-temporal aggregation: several approaches explicitly model temporal alignment and propagation to produce temporally coherent mattes and to reduce flicker and frame-wise artefacts Lin et al. (2021b); Sun et al. (2021). Robust real-time video matting architectures (e.g., RVM and related recurrent/temporal-memory models) demonstrate feasible production performance for certain human-centric shots Lin et al. (2021a,b).

E. Production realities and tool ecosystems

Despite strong academic progress, production adoption is governed by pragmatic constraints: temporal stability, artist control and editability, pipeline compatibility (file formats, cache management, versioning), and predictable failure modes on challenging footage (motion blur, fine hair, partial transparency). Consequently, studios run hybrid workflows that combine ML-assisted initialisations with artist-driven spline refinement or selective manual cleanup. Commercial tools such as DaVinci Resolve's Magic Mask illustrate this hybrid model and the kinds of trade-offs production teams experience when replacing—or augmenting—traditional rotoscoping workflows with learned systems Blackmagic (2025).

F. Gaps in the literature

While surveys exist for image and video matting methods, there is relatively little formal documentation of how these methods perform in production contexts, how failure modes impact downstream compositing, and what organisational factors drive or inhibit adoption. This motivates our production-oriented survey and taxonomy presented in later sections.

III. Survey Methodology

Survey participants were selected from the author's professional network; this constraint creates a geographic and industry-network bias (figure 2). Respondents covered key VFX roles directly involved with video mattes, including plate preparation, compositing, production management, and production technology. While lighting, animation, and layout also use mattes, these departments are typically

not responsible for matte quality assessment and are less represented (figure 4).

All but one participating organisation specialised in TV episodic, feature film, or related narrative and marketing content. One organisation focused solely on short-form informational and promotional media.

Most interviewees reported more than seven years of professional experience, typically evidenced by twenty or more IMDb credits 6. Participants were industry practitioners responsible for delivering commercially successful narrative content rather than academic researchers (figures 4, 13).

Organisations were classified as major, mid-sized, or small VFX studios based on their global footprint and staffing levels. Major studios operate across multiple countries with hundreds to thousands of employees; mid-sized studios typically maintain operations in fewer than 4 cities with up to several hundred staff; small studios are proportionally smaller by a factor of 2 to 10 (figure 3).

The survey followed a structured discovery call format. Participants were first asked to confirm or refute a baseline assumption regarding current matte-production workflows:

"Video mattes are produced by a dedicated rotoscoping department using Boris Silhouette, and delivered both as pixel-based mattes and as human-editable spline data used in Foundry Nuke."

Subsequent discussion explored the perceived strengths and limitations of this workflow and the impact of machine learning–based approaches on existing production practices.

All participants agreed that the stated workflow accurately reflects common industry practice. Opinions differed about the comparative effectiveness of alternatives. Only a small minority reported that machine-learning matte predictions were consistently reliable enough for final delivery without human refinement.

There is no dispute that survey participants are representative of the creators of visual effects for media and entertainment. Between the participants, they have worked on *1294 titles listed in IMDb* with most credits concerned with the production of visual effects 6.

We can see from the global box office takings (WWB) of IMDb titles (figures 8,9,10,11,12) to which the interview participants have contributed from 1995-2025, there are few films that gross large takings near 2 billion USD and many films that gross 0 to 200 million dollars world wide over the last thirty years, showing the diversity of titles in global economic impact.

Figure 13 shows no one genre that interview participants contributed towards that dominated all other genres, audiences tend to notice *spectacular* visual effects yet ignore *invisible* visual effects.

IV. Taxonomy of Current Solutions

Each of the following solutions will produce and accurate video matte which will be judged via human quality control to see its suitability in context of the final composite.

A. Fully Manual Rotoscoping

Fully manual rotoscoping refers to the frame-by-frame construction of matte boundaries using resolution-independent spline representations such as Bézier or NURBS curves. This practice originates from the classical rotoscope technique introduced by Fleischer Fleischer (1915) and later transitioned into digital form with the emergence of commercial compositing tools in the 1990s Bratt (2018); Brinkmann (1999).

Contemporary rotoscoping systems—including Adobe After Effects, Foundry Nuke, Boris FX Mocha Pro, and Silhouette—provide vector-based spline primitives that are rasterized into pixel-accurate opacity mattes Birn (2012); Bratt (2018). Although these applications differ in interface design and feature sets, they share a common operational paradigm: artists define control points and tangents, animate these shapes through keyframes, and rely on curve interpolation to maintain temporal coherence.

The primary value of this approach is its direct, artist-driven precision, enabling high-fidelity control over occlusion boundaries, sub-pixel detail, and semi-transparent regions. However, this precision is offset by substantial labor overhead and ergonomic constraints. Throughput is fundamentally limited by human–computer interaction—stylus strokes, control-point manipulation, path re-topology—and the cognitive demands of interpreting motion-blurred or visually ambiguous regions Houses (2002). A skilled roto artist may complete several hundred frames for a single object in one hour, whereas complex structures such as hair, cloth, foliage, or articulated hands may require hundreds of overlapping spline shapes.

Quality is largely deterministic when sufficient time and labor are allocated; fully manual pipelines therefore scale primarily via parallelisation. Production workflows often divide a shot across multiple artists, either by temporal segmentation or by partitioning complex subjects into spatial components Birn (2012). While this strategy reduces turnaround time, it does not alter the intrinsic reliance on manual spline animation. As a result, fully manual rotoscoping remains the most reliable—but also the most resource-intensive—method for generating final-quality mattes in contemporary VFX production.

B. Semi-Automatic Tools

Semi-automatic rotoscoping tools occupy a middle ground between fully manual spline animation and fully automated, machine-learning-based matte extraction. These systems augment traditional spline workflows

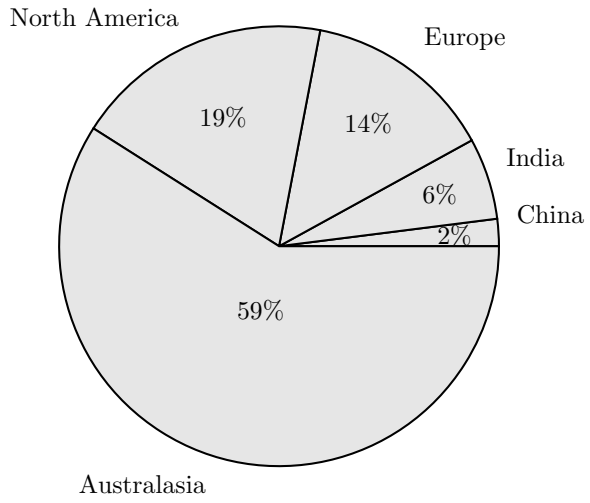

Figure 2: Location

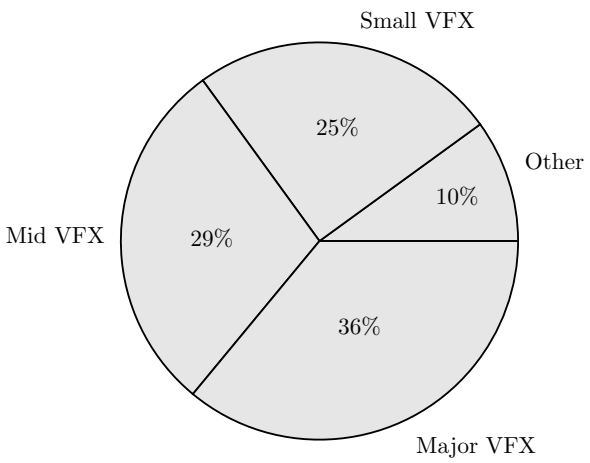

Figure 3: Org. Type

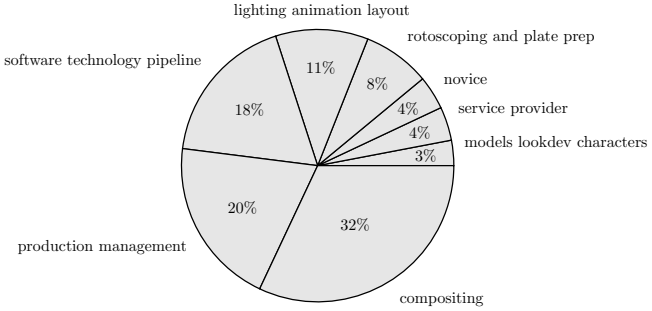

Figure 4: Role of Interview Participants

with tracking, motion analysis, and user-guided propagation. Their purpose is to reduce per-frame manipulation without relinquishing the editability and determinism valued in production environments.

Commercial tools such as Boris FX Mocha Pro and Silhouette provide planar tracking, sub-planar mesh warping, and feature-aligned shape propagation that significantly lessen the need for frame-by-frame keying SilhouetteFX (2020). Similarly, the rotoscoping node in Foundry Nuke integrates point and spline tracking that allows shapes to follow local motion with moderate user correction Foundry (2025b). These systems rasterize vector-based shapes into mattes in the same manner as fully manual workflows, but they automate the interpolation of shape trajectories and contour adjustments across time.

Compared with purely manual rotoscoping Bratt (2018); Brinkmann (1999), semi-automatic methods increase throughput by enabling artists to focus on corrective work rather than primary shape animation. Planar and mesh trackers maintain temporal coherence across sequences containing rigid or quasi-rigid motion, reducing the number of required key-frames and minimising tedious micro-adjustments. In practice, these tools offer substantial efficiency gains for assets such as props, rigid body elements, vehicles, or architectural forms where motion is largely planar. However, complex non-rigid subjects—including hair, cloth, soft tissue, and foliage—often exceed the representational capacity of planar tracking, requiring artists to fall back to manual spline refinement.

The effectiveness of semi-automatic approaches depends heavily on footage characteristics. High-frequency deformation, motion blur, occlusions, and transparency introduce tracking drift or shape distortion, requiring additional corrective labour from artists Houses (2002). Thus, while semi-automatic roto can reduce total key-frame density by an order of magnitude for suitable shots, it rarely eliminates human intervention altogether.

From a production perspective, semi-automatic systems remain more predictable and controllable than machine-learning-driven matte extraction. They produce fully editable spline representations, integrate cleanly with established cache formats, and maintain deterministic behavior across versions and handoffs. As a result, they are broadly adopted in visual effects pipelines as a practical compromise between efficiency and control, offering measurable speed improvements while preserving the editorial flexibility required for final-quality matte production.

C. ML-Assisted Matte Extraction

Machine-learning (ML) approaches to matte extraction aim to reduce manual labour by producing initial alpha mattes or segmentation masks that artists can refine. Survey participants reported a heterogeneous toolset in production use, including commercial and research systems such as **Rotobot** Kognat (2019), Nuke-integrated model nodes (e.g., `SAM2`, `ViTMatte`), and lightweight matting networks such as `MODNet` Ke et al. (2022). SaaS offerings (for example, Runway and Slapshot) were also discussed;

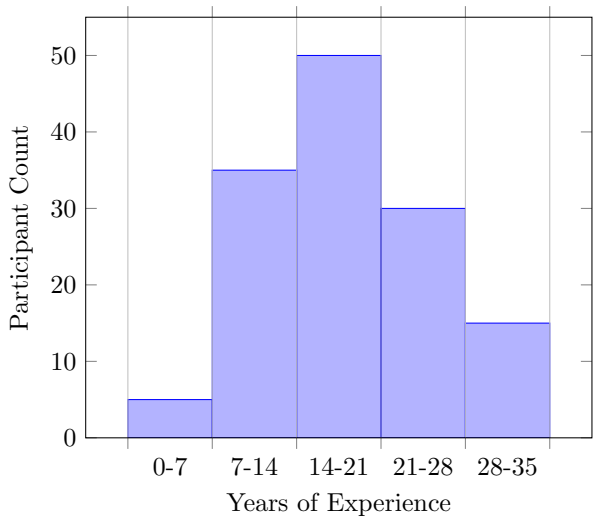

Figure 5: Experience

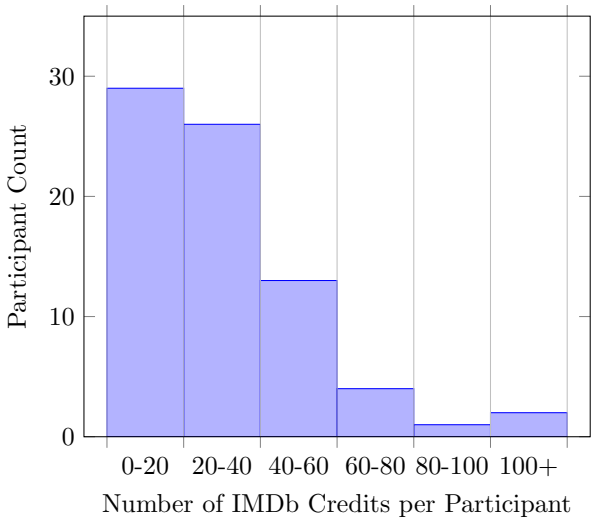

Figure 6: Credit Count

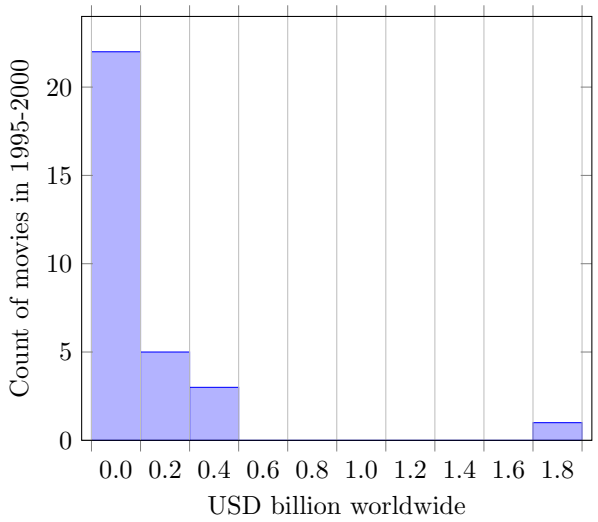

Figure 7: WWB 1995-2000

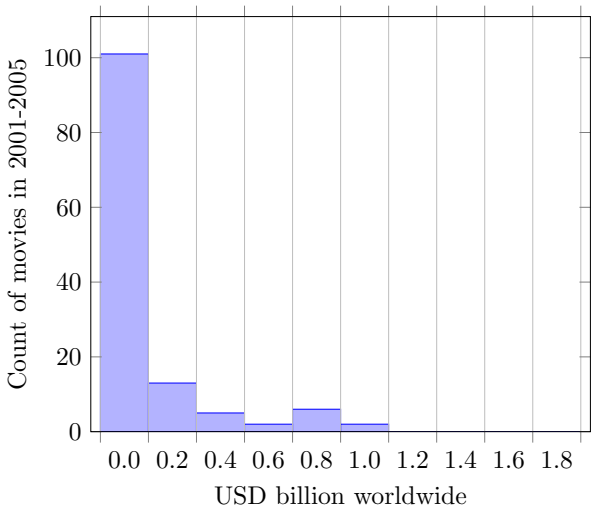

Figure 8: WWB 2001-05

adoption of cloud services is frequently constrained by contractual and legal concerns (data residency, IP and client confidentiality) even when technical performance is competitive ML (2023); Slapshot (2024).

Practitioners emphasized two recurring technical limitations of ML-assisted mattes: (1) *temporal instability* (frame-to-frame flicker and inconsistency) and (2) *boundary inaccuracy* in challenging conditions (motion blur, fine hair, partial transparency). These failure modes reduce trust in fully automated outputs and motivate hybrid workflows in which ML outputs seed artist-driven refinement Lin et al. (2021a); Sun et al. (2021); Xu et al. (2017).

A recent line of research exemplifies the direction toward more robust production-capable systems. The Automated Video Segmentation Machine (AVSM) Pipeline integrates language-guided object selection with matting refinement and temporal aggregation, demonstrating improved semantic control and temporal coherence in benchmark settings Merz and Fostier (2025). While promising, such hybrid architectures still require careful validation on production footage and attention to legal and pipeline-compatibility issues before broad studio adoption.

D. Hybrid Pipelines

Hybrid rotoscoping pipelines combine machine-generated mattes with human refinement, typically using tracking tools, spline editors, and paint-based cleanup. Although attractive in principle, survey participants expressed strong reservations about the practicality of refining machine-learning mattes within production environments. The core issue is representational: ML

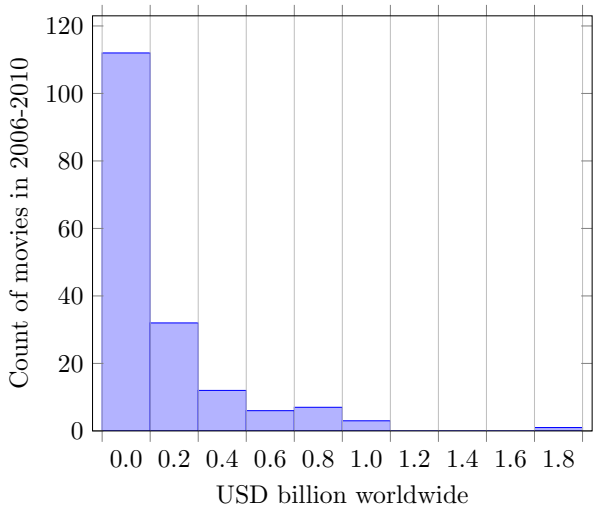

Figure 9: WWB 2006-10

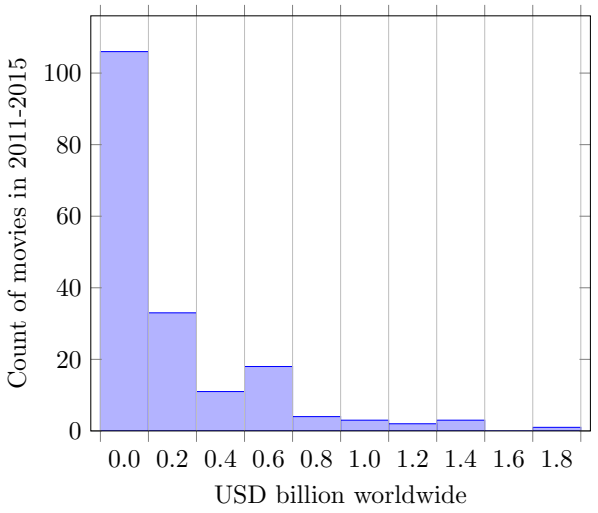

Figure 10: WWB 2011-15

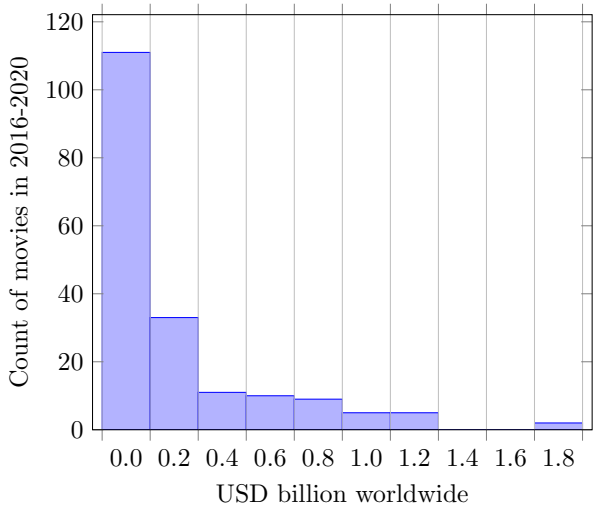

Figure 11: WWB 2016-20

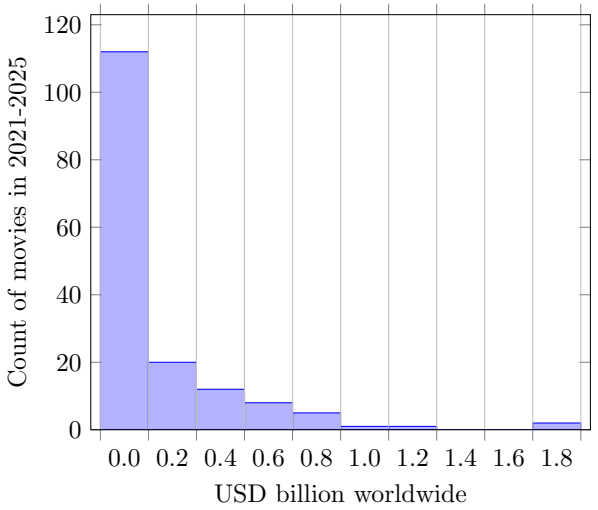

Figure 12: WWB 2021-25

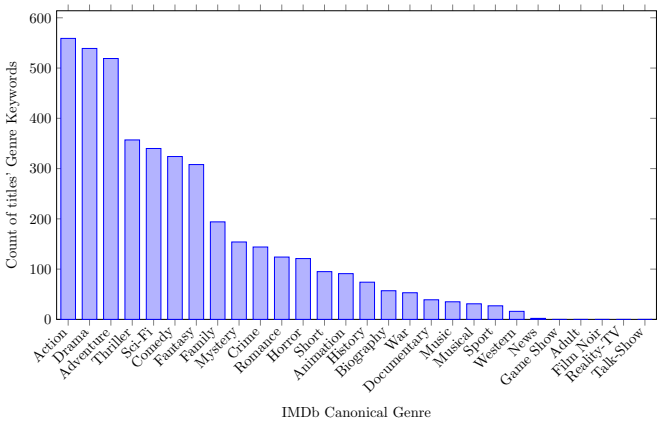

Figure 13: Participants' IMDb Title Count by Genre

Figure 14: Manual Spline Based Rotoscoping

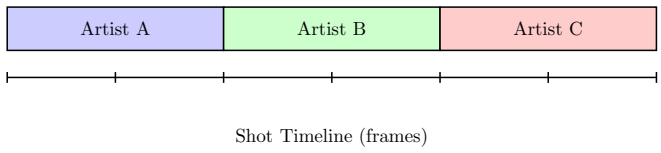

Figure 15: Optimise Task Using Multiple Technical Artists

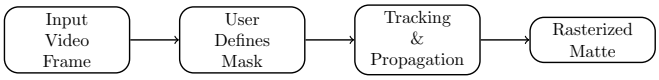

Figure 16: Semi-Automatic: Tracking and Propagation

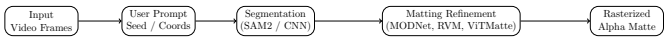

Figure 17: ML Automation: Segmentation to Matting

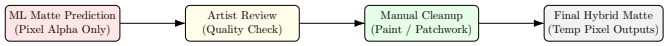

Figure 18: Patch ML Prediction for Early Stages

predictions are delivered as raster alpha mattes—arrays of grayscale pixel values—rather than the spline-based, resolution-independent shapes used in professional rotoscoping workflows Ke et al. (2022); Lin et al. (2021a,b). Because these predictions lack editable control points, artists cannot refine them directly; all improvements must be performed through paint-like operations, which reintroduce labour-intensive manual work and break the non-destructive editing model that spline systems enable.

Participants frequently described a "spent cost" problem: once a team attempts to repair an ML-generated matte, the accumulated cleanup effort often exceeds the cost of having produced a clean, human-editable spline representation from the outset. Pixel-based mattes also accumulate artefacts—noise, instability, and local inconsistencies—that cannot be cleanly corrected without discarding and restarting the matte, further discouraging hybrid repair.

Despite these limitations, hybrid use of ML mattes was considered viable when boundary precision was not critical. In early stages of the post-production pipeline—such as post-visualisation, layout blocking, editorial previews, or rapid internal review—participants reported that ML-assisted masks could accelerate turnarounds, provided that the required level of visual fidelity was moderate and temporal artefacts were acceptable Merz and Fostier (2025); Sun et al. (2021). In such contexts, raster predictions can function as disposable approximations, freeing artist hours without compromising final-shot quality. However, for hero shots or any context requiring production-level precision, participants consistently preferred spline-based workflows due to their editability, stability, and long-term maintainability.

V. Findings From 92 Conversations with Participants in Visual Effects

As participants' views of the video matting process are determined by their role, there will be varied domain expertise in costing, workflow management, quality control, and automation technology.

We find that compositors determine the failure conditions for automation based on the quality standards of the video matte used in the final composite, and technology experts understand the technical barriers to adopting ML algorithms within digital content creation applications such as Foundry's Nuke. Producers have a better view of costs and schedules. Supervisors have a holistic approach to workflow optimisation.

A. Assumptions for Visual Effects Video Mattes

From our interviews, we found that 95 percent of people need the output from video mattes within the compositing package Nuke. 98 percent of interview participants working in compositing and plate preparation teams agreed that the source files for the video mattes were required as animated splines. Outside those teams, the percentage of people who mentioned editable spline data as a requirement was 76 percent.

VFX Role	**Percentage**
Animation	100%
Shots	100%
Assets	100%
Compositing	93.10%
Plate Prep and Rotoscoping	100%
Layout	100%
Software	81.25%
Service Provider	50%
Other	66.67%
Production	94.12%
Founder	81.82%

Table I: Confirm Rotoscoping is use for Video Mattes

The reason given for human-editable output for the video matte was to prevent asynchronous communication to an external team; editing the spline data in the final compositing context was possible, as a last resort.

When time allows, an external team can edit the splines; this iterative process is expedient.

Editing spline data in Foundry's Nuke compositing application degrades the human-computer interface; users must cache the output to disk as soon as edits are complete to restore the interface responsiveness.

We found a small number of users, 1.3 percent, mentioned that machine learning predictions as pixel masks were acceptable for final results.

B. Barriers for Video Matte Automation

Barrier	**Percentage**
Temporal Instability	84%
Edge Consistency	85%
Partial Transparency	72%

Table II: Barriers to Adoption

The barriers to adoption are *temporal instability*, *edge prediction artefacts*, and *semi-transparent edges* (relating to the motion blur, defocus, and semi-transparent objects such as hair and foliage)

Cost of briefing, feedback, and quality control is a theme reported in the union of compositors, plate prep, and production staff, 56 percentage by these members in these roles, but as 12 percentage by members outside of these roles. 65 percent of the compositors recalled communication errors between rotoscoping providers and consumers. The solution to these missed communications

is to acknowledge the use of the video matte in the final composite. 52 percent of creative roles mentioned creative process can adjust the initial requirements for video mattes.

C. Drivers for Adoption for Video Matte Automation

Most interviewees acknowledged the drivers for adoption of video matte automation were: speed gains, cost savings, and impact on the schedule with regard to client deadlines encompassing final acceptable quality control.

Driver	Percentage
Cost Minimisation	100%
Task Duration Minimisation	100%
Schedule Optimisation	97%

Table III: Motivation for Adoption of Automation

D. Paint Points for Video Matte Creation

The pain points of video matte creation and usage relate to a few themes: production budget in terms of time and resources and acceptable quality.

This burden is shared between the production staff who estimate the resources required to deliver completed video mattes suited to the final context within the composite and the team producing the video mattes using manual rotoscoping methods.

Judging the quality of the video masks, accurate briefing, and feedback communication can each be pain points until a systematic approach is developed.

Poor temporal stability and foreground region edge estimation from automated machine learning prediction cause pain for compositors using the video masks.

Video footage can be dimly lit or have poor colour and texture separation between foreground and background items being matted. Manual results rely on the expert eye of a rotoscoping technical artist.

Objects such as strands of hair, foliage and other fine objects can require detailed video masks. Machine learning segmentation or matting solutions with semi-transparency often fail in these use cases.

E. Requirements for Video Matte Automation

85 percent of producers, compositors and plate preparation staff conveyed that machine learning predictions enable downstream tasks to kick off while manual rotoscoping is in progress.

The automated video masks act as a disposable substitute for crafted rotoscoping for internal feedback.

The terms *temps*, *bash comps*, and *post-visualisation* were used to describe these disposable low-quality video mattes. Only to be replaced by curated high-quality rotoscoping when available.

The placeholders often cannot shorten the schedule because the period when the disposable masks are in use is shorter than the time required to complete the final quality rotoscoping mattes.

80 percentage of compositors mentioned updating the brief due to creative adjustments, as the narrative of a visual effects shot invalidates the original context of the video matte.

We can narrow the scope of the manual rotoscoping mattes by spatial and temporal aspects of how the video matte is used. By using disposable mattes to inform creative decisions in editorial and downstream departments such as animation, matte painting, layout, camera matchmove, lighting, and effects.

F. Video Matte Automation Adopted by Visual Effects Practitioners

Only some participants are aware of the details of automated solutions. Based on these participants alone, we present the frequency of automation technology solutions in use for visual effects use cases.

ML Tool	Percentage
Segment Anything Meta 2	73.8%
Kognat Rotobot OpenFX Plugin	64.8%
MODNet Cattery	10.8%
ViT Matte Cattery	40.5%
Slapshot SaaS	5.4%
AVSM	10.8%
DaVinci Resolve Magic Mask	35.1%

Table IV: Particpants that Mentioned ML Tools by name

The human computer interface for machine learning predictions will depend on the model. Whereas human-centric communication can be illustrated and annotated. If a model is semantic and can only recognise a fixed number of classes, this lack of flexibility can limit the effectiveness of the predictions.

By the same token, if interactivity is slow, defining the matte region requires the operator to add and subtract regions very often per iteration, which increases the cost of the *semi automation*.

Using natural language processing (NLP) as an interface, as mentioned by 18 percentage of people who had been exploring machine learning solutions.

Full automation was mentioned by 6 per cent of users who had employed machine learning. A template was used for common shots that require a matte. It was seen as a surprise that the end result could be used in the final matte.

56 per cent of compositors think placeholder machine learning automatic mattes are beneficial, while 26 percent prefer to use human-curated video layers alone.

VI. Comparative Analysis

Automation of Video Mattes is only successful for internal use to inform the creative process and create a data footprint that will later be replaced by manual mattes created via rotoscoping in Silhouette (Silhouette) and imported into Nuke Foundry (2024).

Machine learning automation tools mentioned were: Segment Anything Meta 2 in Nuke Rafael Silva (2024a), Ravi et al. (2024), Yang et al. (2024)

SAM2 Rafael Silva (2024a) was used with ViTMatte Rafael Silva (2024b), to increase the quality of edges

Participants used commercial solutions, Magic Mask feature in DaVinci Resolve Blackmagic (2025), Kognat Rotobot OpenFX Plugin Kognat (2019), Runway ML (2023), and Slapshot Slapshot (2024) are software as a service products.

The consensus was that segmentation gives binary decisions between foreground and background pixels, which cannot represent the anti-aliasing required for high-quality masks.

While matting tools like ViTMatte and RVM allow for antialiasing, they can fail when there is self-similarity between foreground and background pixels' colour and texture.

The all in one tool Ke et al. (2022), which can be used in Nuke Foundry (2025a), was seen as a tool that can reach final quality with fine tuning.

A participant mentioned that implementing a number of models available from Nuke Cattery Foundry (2025a) and testing all of the output to determine which model would be most successful. No one model that is most suitable for all tasks, and it becomes a task to explore a variety of models to determine which is most suited to the requirements of the use case.

Participants mentioned that fine-tuning models would lead to improved results using both Nuke Cattery Foundry (2025a) and Rotobot Butler from Kognat Kognat (2021)

Comparative metrics are possible with machine learning models, human-based quality control by compositing supervisors in shows that while the accuracy of these models may improve on a leader-board they are rarely acceptable for production quality, and we are reliant on human inference and human quality control rather than automation for video mattes and the assessment of quality.

VII. Discussion

Models need to predict results as a spatially and temporally sparse data structure that is human-editable with near real-time assessment in the context of its final use. If the output is not available in the final context, can we know it is fit for use? Real-time feedback and editing become more important than automation.

The best candidate for this data structure is time-varying NURBS or Bézier splines. Although there is research published about time varying topology Dalstein et al. (2015), otherwise consistent topology between splines over time, particularly because this is supported for editing in Nuke Foundry (2024).

If we can construct a machine learning model with an end-to-end differentiable equations for this sparse data structure with the intended use of a video matte, we can potentially create a continuous learning pipeline where the edits made by hand can be used to train the model, creating the predictions that generalise the model for future inferences.

Human communication to brief a video-matting task is inexact, which means that the human-computer interface instructing an automated model will suffer the same problem. We will need failsafe approaches to instruct which spatial and temporal regions need to be isolated as a video matte.

Similarly, the feedback on how to improve the video matte through human edits requires a complete understanding of how the matte is to be used. When this understanding is within a single user, there is no need to abstract this as a transferable message; the user can simply edit the data structure to their requirements for the final composite.

VIII. Conclusion

We need human-editable video mattes, temporally varying splines of open and closed shapes are well-suited.

Splitting the labour between roto-scoping and compositing departments created a communication problem. The reason for this split in tasks is economic; the cost of human resources for the Rotoscoping task is significantly less than for a more specialist compositing task.

The problem with automated video mattes is that they are rarely perfect, and editing the automated predictions costs more than manual rotoscoping.

As a result, until automated video mattes are human-editable, they will not be widely adopted in industrial applications.

There is a view that the need for image intermediates is decreasing as we can generate all the components of a novel image using stable video diffusion as a single layer, negating the need for video mattes to adjust elements of the generative video.

References

Xue Bai and Jue Wang. Towards temporally-coherent video matting. In *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, pages 1510–1517, Kyoto, Japan, 2009. IEEE. doi: 10.1109/ICCV.2009.5459258.

Jeremy Birn. *Digital Lighting & Rendering*. New Riders, Berkeley, CA, 3 edition, 2012. ISBN 978-0321928429.

Blackmagic. Begginers guide to davinci resolve, 2025. URL https://documents.blackmagicdesign.com/UserManuals/DaVinci-Resolve-17-Beginners-Guide.pdf.

Boris FX. Mocha pro user guide. https://borisfx.com/products/mocha/, 2024. Accessed 2025.

Benjamin Bratt. *Rotoscoping: Techniques and Tools for the Aspiring Artist*. Routledge, 2018. ISBN 978-1138474253.

Ron Brinkmann. *The Art and Science of Digital Compositing*. Morgan Kaufmann, 340 Pine Street, Sixth Floor; San Francisco; CA; United States, 1999. ISBN 9780121339609.

Boris Dalstein, Rémi Ronfard, and Michiel van de Panne. Vector graphics animation with time-varying topology.

ACM Trans. Graph., 34(4), July 2015. ISSN 0730-0301. doi: 10.1145/2766913. URL https://doi.org/10.1145/2766913.

Max Fleischer. Method of producing moving-picture cartoons, 1915. URL https://patents.google.com/patent/US1242674A/en.

Foundry. Nuke roto node (documentation), 2024. URL https://learn.foundry.com/nuke/content/reference_guide/draw_nodes/roto.html.

Foundry. Cattery: Community open source models in nuke, 2025a. URL https://community.foundry.com/cattery.

Foundry. Nuke rotoscoping and spline tools documentation. https://learn.foundry.com/nuke, 2025b. Accessed 2025.

Philip Houses. Rotoscoping in the digital domain: Tools, techniques, and production practices, 2002. Course #12: Digital Compositing.

Adobe Inc. Roto brush and refine matte (after effects documentation), 2022–2024. URL https://helpx.adobe.com/after-effects/using/roto-brush-refine-matte.html.

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:1140–1147, 06 2022. doi: 10.1609/aaai.v36i1.19999.

Kognat. Rotobot openfx plugin documentation, 2019. URL https://rotobot-docs.readthedocs.io/en/latest/.

Kognat. Rotobot butler: Read the docs, 2021. URL https://rotobot-butler-docs.readthedocs.io/en/latest/.

Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30 (2):228–242, 2008. doi: 10.1109/TPAMI.2007.1177.

Shanchuan Lin, Jiaqi Sun, Xiaohong Wu, Bilal Kayhan, and Stefano Soatto. Robust video matting (rvm). https://github.com/PeterL1n/RobustVideoMatting, 2021a. GitHub repository; access 2025.

Yen-Chun Lin, Zhaoyang Liu, Ning Wang, Ming Ding, and Jing Liao. Rovomatte: Robust human video matting via temporal decomposition. In *SIGGRAPH Asia 2021 Technical Communications*, New York, NY, USA, 2021b. Association for Computing Machinery. doi: 10.1145/3478512.3488628.

Johannes Merz and Lucien Fostier. Automated video segmentation machine learning pipeline. In *Proceedings of the Digital Production Symposium*, DigiPro '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720086. doi: 10.1145/3744199.3744635. URL https://doi.org/10.1145/3744199.3744635.

Runway ML. Runway: Cloud-based creative ml tools. https://runwayml.com, 2023. Commercial SaaS; Accessed 2025.

Rafael Silva. Sam2 node for nuke (segment-anything model integration). https://github.com/rafaelperez/Segment-Anything-for-Nuke, 2024a. Accessed 2025.

Rafael Silva. Vitmatte: Vision-transformer based matting node. https://github.com/rafaelperez/ViTMatte-for-Nuke, 2024b. Repository / Nuke integration; Accessed 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. URL https://arxiv.org/abs/2408.00714.

Boris FX (Silhouette). Silhouette — industry standard for rotoscoping and paint, 2024. URL https://borisfx.com/products/silhouette/.

SilhouetteFX. Silhouettefx user manual. https://borisfx.com, 2020. Accessed 2025.

Slapshot. Slapshot saas: Video segmentation services. https://slapshot.ai, 2024. Commercial SaaS; Accessed 2025.

Ke Sun, Jiaolong Yang, Dongxu Liu, Zhe Zhang, Hailin Bao, and Jia Li. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1406–1416, Virtual Conference, 2021. IEEE. doi: 10.1109/CVPR46437.2021.00690.

Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2970–2979, Honolulu, HI, USA, 2017. IEEE. doi: 10.1109/CVPR.2017.317.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory, 2024. URL https://arxiv.org/abs/2411.11922.