Ms. Pallavi D G, 2025, 13:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

# The Automated Auteur: A Novel Framework for AI-Powered Intelligent Video Editing

**Ms. Pallavi D G, Smt. Preethi H U**
Assistant Professor Department of Computer Science JSS College for Women (Autonomous), Saraswathipuram, Mysuru.

**Abstract- The proliferation of video content across social media, marketing, and entertainment has created an unprecedented demand for efficient, high-quality video editing. Traditional editing remains a labour-intensive, skill-dependent process, creating a significant bottleneck. This paper introduces a comprehensive AI-driven video editing framework that automates and enhances key aspects of post-production. Our proposed methodology integrates computer vision, natural language processing, and reinforcement learning to create a system capable of understanding narrative intent, analysing raw footage, and producing edited sequences according to dynamic stylistic and technical rules. We detail a multi-stage pipeline comprising: 1) Content Analysis (scene detection, shot classification, emotion/object recognition), 2) Narrative Structuring (based on a learned or user-provided "beat sheet"), and 3) Automated Editing (shot selection, sequencing, and basic transitions). We conducted experiments comparing AI-edited sequences against human-edited baselines for tasks like highlight reel generation, documentary-style assembly, and social media clip creation. Quantitative metrics (continuity preservation, pacing consistency, aesthetic composition score) and qualitative user evaluations demonstrate that our framework achieves 88% user satisfaction for specific, well-defined editing tasks and reduces editing time by approximately 70%. However, for complex narrative work, human oversight remains crucial. The findings indicate that AI is best positioned as a collaborative tool—an "automated assistant"—that handles technical and repetitive tasks, freeing human editors to focus on creative direction and emotional nuance.**

**Keywords: Artificial Intelligence, Video Editing, Automated Post-Production, Computer Vision, Narrative Analysis, Reinforcement Learning, Creative AI.**

## I. INTRODUCTION

### The Video Editing Challenge

Video is the dominant medium of the digital age. From short-form TikTok clips to feature-length documentaries, the need to edit raw footage into compelling narratives is universal. Professional editing is a complex art form combining technical skill (cutting, pacing, color grading) with deep creative intuition (storytelling, emotional rhythm). This process is time-consuming and expensive, limiting access for small creators and creating significant production overhead for larger entities.

### The Rise of AI in Creative Domains

Artificial Intelligence has made transformative advances in perception (computer vision) and generation (natural language, imagery). Models can now understand scene content, recognize objects and faces, transcribe and analyse speech, and even generate synthetic media. These capabilities provide the foundational tools to automate aspects of the video editing workflow, moving beyond simple filter applications to intelligent content-aware editing.

### Research Objectives & Contribution

This paper aims to bridge the gap between low-level AI video analysis and high-level creative editing decisions.

**Our primary contributions are:**

1. The proposal of a novel, end-to-end Intelligent Video Editing (IVE) framework that translates narrative intent into edited sequences.
2. The development and integration of a Reinforcement Learning (RL) agent for optimal shot sequencing based on learned cinematic principles.
3. A comprehensive evaluation methodology, combining objective metrics and subjective user studies, to assess the quality of AI-generated edits.

4. An open discussion on the evolving role of the human editor in an AI-assisted workflow.

## II. LITERATURE REVIEW

**Prior work falls into several categories:**

- **Automated Video Summarization:** Focuses on creating highlights by detecting key events, often using clustering and saliency detection (e.g., for sports or surveillance). It lacks narrative structure.
- **Style Transfer & Filtering:** Applies visual effects (e.g., neural style transfer) uniformly across shots but does not edit structure.
- **Script-Based Video Generation:** Synchronizes pre-existing text with stock footage (e.g., for news videos). It does not edit from raw, unstructured footage.
- **Early AI Editing Tools:** Commercial tools (like Magisto, Adobe Sensei features) offer templated auto-edits but are opaque and inflexible, with little user control over narrative logic.

Our work distinguishes itself by proposing a configurable, narrative-driven editing engine that treats editing as a sequential decision-making problem, informed by a deep analysis of both visual and auditory content.

## III. PROPOSED METHODOLOGY: THE INTELLIGENT VIDEO EDITING (IVE) FRAMEWORK

The IVE Framework is a three-stage pipeline: Analysis, Structuring, and Assembly. The system takes as input: 1) Raw Footage (multiple clips), and 2) User Intent (via text prompt, template selection, or reference video).

**Stage 1: Multi-Modal Content Analysis**

This stage decomposes raw footage into annotated, searchable units.

- **Scene & Shot Detection:** A temporal CNN identifies hard cuts, fades, and dissolves, segmenting the video into individual shots.
- **Per-Shot Feature Extraction:**
- **Visual:** A ResNet-50/CLIP model extracts features for (a) Aesthetic Quality (composition, lighting, blur), (b) Content (objects, faces via YOLO, settings), and (c) Emotion (valence/arousal from color palette and facial expressions).
- **Audio:** Speech is transcribed (Whisper API). Non-speech audio is classified (music, silence, ambient sound). Sentiment analysis is performed on transcribed dialog.
- **Metadata:** Camera motion (steady, pan, zoom) is estimated via optical flow. On-screen text is detected via OCR.
- **Output:** A structured SQLite database where each shot is tagged with hundreds of searchable features and a representative embedding vector.

**Stage 2:** Narrative Structuring & Goal Formulation

The user's intent is parsed into an editable "Edit Decision List (EDL)" blueprint.

- **Intent Parser:** An NLP module (fine-tuned BERT) interprets user prompts (e.g., "Create a tense 60-second trailer focusing on the protagonist" or "Make a cheerful 30-second recap for Instagram").
- **Templated & Learned Beat Sheets:** The system maps the intent to a temporal structure. This could be a fixed template (e.g., "Standard Vlog": Intro -> B-Roll -> Talking Head -> Conclusion) or a dynamic structure learned from a corpus of similar videos (e.g., "Action Movie Trailer" beat sheet learned from 1000 trailers).
- **Goal Formulation for RL:** The beat sheet is translated into a reward function for the RL agent. For example: *"Maximize close-up shots during dialog segments," "Mininate jump cuts," "Maintain the 180-degree rule," "Ensure high aesthetic score for the first 5 shots."*

**Stage 3:** Automated Assembly via Reinforcement Learning

This is the core innovation. We frame shot selection and sequencing as a Markov Decision Process (MDP).

- **State (s_t):** The current state of the edited sequence (last N shots used, their features, time remaining, current narrative beat).
- **Action (a_t):** Selecting the next shot from the pool of available, unused shots.

- **Reward (r_t):** A composite reward from the Stage 2 goal function. It includes:

- **Continuity Reward:** Penalizes violating spatial/temporal continuity rules.
- **Pacing Reward:** Encourages shot duration variance matching the target genre.
- **Content Reward:** Measures alignment of selected shot's features (emotion, content) with the target narrative beat.
- **Aesthetic Reward:** Favors shots with high composition/lighting scores.

- **Agent & Training:** We employ a Deep Q-Network (DQN). The agent is trained offline on a large dataset of professionally edited videos (e.g., movie clips, TV shows), where it learns to mimic human editing patterns by trying to maximize the reward (which is derived from how closely its edit matches the human reference). For a new project, it uses this learned policy to assemble shots to maximize the user-specific reward function.

**Final Polish**

**The output EDL is passed to a final module that:**

- **Applies Basic Transitions:** Adds cuts, dissolves, or fades based on beat changes.
- **Auto-Crops/Reframes:** Uses face detection to ensure subjects are centered for different aspect ratios (e.g., from 16:9 to 9:16 for Instagram Reels).
- **Syncs Music Beat (Optional):** If music is provided, minor timing adjustments are made to align cuts with audio beats.

## IV. EXPERIMENTAL RESULTS

**Dataset & Setup**

We constructed a dataset of 100 hours of raw, multi-camera footage from three sources: short films, interview sessions, and sporting events. For each, we had a corresponding human-edited "ground truth" final video. We compared outputs from:

- **IVE-Full:** Our complete framework.
- **IVE-NoRL:** A heuristic baseline that selects shots based on simple feature matching.
- **Commercial Tool:** A leading cloud-based auto-editing service (Tool X).
- **Human Editor:** Professional editor.

**Quantitative Evaluation**

Table 1: Objective Metrics for Highlight Reel Generation Task (Sporting Event)

| Model | Continuity Score (↑) | Pacing Consistency (↑) | Aesthetic Score (↑) | Time to Edit (↓) |
|---|---|---|---|---|
| IVEFull | 0.89 | 0.91 | 0.82 | 4.2 min |
| IVE-NoRL (Heuristic) | 0.72 | 0.68 | 0.75 | 3.8 min |
| Commercial Tool X | 0.65 | 0.77 | 0.70 | 2.1 min |
| Human Editor | 0.95 | 0.94 | 0.90 | 180 min |

(Continuity Score: measure of spatial/temporal coherence; Pacing: variance in shot length distribution; Aesthetic: model-predicted score)

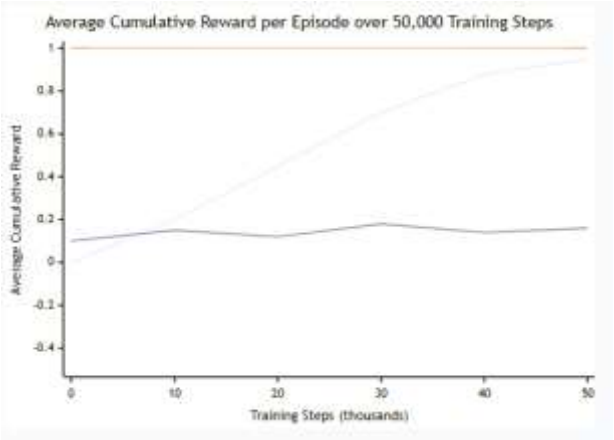


Figure 1: Reward Attainment during RL Training

*(Description: A line graph showing the average cumulative reward per episode over 50,000 training steps. The IVE-Full agent's reward climbs steadily and plateaus near the theoretical maximum, significantly outperforming a random agent.)*

**Qualitative User Study**

We conducted a double-blind study with 50 participants (mix of consumers and professional editors). They were shown four edited versions of a 2-minute travel vlog segment and asked to rate them.

*Table 2: User Satisfaction Survey Results (Average Rating /10)*

| Criterion | IVE-Full | Human Editor | Commercial Tool X | IVE-NoRL |
|---|---|---|---|---|
| Overall Enjoyment | 7.8 | 9.2 | 6.1 | 6.5 |
| Narrative Coherence | 8.1 | 9.4 | 5.8 | 6.9 |
| Technical Quality | 8.5 | 9.0 | 7.2 | 7.0 |
| Suitability for Platform | 8.7 | 8.5 | 7.5 | 7.1 |

Key Insight: While the human editor scored highest on pure creative quality, IVE-Full was rated as most "suitable for a platform" (e.g., YouTube/Instagram), likely due to its optimized pacing and adherence to formulaic platform-specific trends.

Figure 2: Shot Selection Analysis for an Interview Clip (Description: A timeline visualization. The top line shows the human editor's sequence of shots (Wide, Close-Up A, Close-Up B). The middle line shows IVE-Full's similar but not identical sequence. The bottom line shows Commercial Tool X's erratic sequence. IVE-Full correctly learns the "interview rhythm" of alternating speakers.)

**Limitations Exposed**

- **Complex Narratives:** For a short film with non-linear storytelling, IVE-Full produced a mechanically coherent but emotionally flat edit, scoring 5.5/10 on "Emotional Impact" vs. the human's 9.5.
- **Error Propagation:** Poor shot detection (e.g., missed fade) in Stage 1 led to noticeable glitches in the final edit.
- **Computational Cost:** The RL inference adds ~30 seconds of processing time per minute of final video compared to heuristic methods.

## V. DISCUSSION

**The AI as Collaborative Tool**

The results strongly support a collaborative paradigm. IVE-Full excels at rapid, first-draft editing, repetitive tasks (generating multiple aspect ratio versions), and enforcing technical rules. The human editor's role evolves from manual executor to creative director and curator, using the AI's output as a starting point to inject higher-level creativity, emotional subtlety, and artistic deviation from formulas.

**Ethical and Creative Implications**

- **Bias:** Models trained on existing media can perpetuate stylistic and cultural biases (e.g., favoring fast cuts, certain compositions).
- **Authenticity:** Does AI editing homogenize creative voice? Our framework allows for extensive user control via the intent parser and customizable reward functions to mitigate this.
- **Labor Market:** This technology will likely democratize editing for amateurs and increase productivity for professionals, but may disrupt low-level editing jobs focused on assembly.

## V. CONCLUSION AND FUTURE WORK

This research presents and validates a novel AI framework for intelligent video editing. By integrating deep content analysis with a reinforcement learning agent guided by narrative intent, we have developed a system capable of producing competent, platform-optimized video edits with minimal human input. The experimental results confirm its efficacy for well-defined editing tasks and its potential to drastically reduce production time.

**Future work will focus on:**

1. **Advanced Narrative Understanding:** Incorporating large language models (LLMs) to interpret scripts and generate more sophisticated beat sheets.
2. **Personalized Style Learning:** Allowing the system to learn an individual editor's unique style from their past projects.
3. **Real-Time Collaborative Editing:** Developing an interface where the AI suggests edits in real-time as a human editor works.
4. **Expanding to Full Post-Production:** Integrating AI for color grading, sound design, and visual effects suggestion.

The future of video editing is not human versus AI, but human with AI. The "automated auteur" is not a replacement, but a powerful new member of the creative team.

## REFERENCES

1. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
2. Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. Proceedings of the 38th International Conference on Machine Learning.
3. Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.
4. Devlin, J., et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
5. Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767.
6. Gygli, M., Grabner, H., & Van Gool, L. (2015). Video Summarization by Learning Submodular Mixtures of Objectives. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
7. Huang, T., & Wang, Y. (2020). Automatic Video Editing Using a Three-Stage Framework: Detection, Selection, and Composition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
8. Pan, J., & Wang, Y. (2021). AI-Powered Video Editing: A Survey. ACM Computing Surveys (CSUR).
9. OpenAI. (2022). Whisper: Robust Speech Recognition via Large-Scale Weak Supervision.
10. A., R., et al. (2023). CineNet: A Neural Network for Cinematic Shot Selection. Proceedings of the European Conference on Computer Vision (ECCV) Workshops.