JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 1, pp. 61-70, Jan. 2026

Check for updates

AI 기반 비디오 모션캡처 도구 성능 비교 및 평가 프레임워크

갈 평 건[1] · 정 진 헌[2*]
[1]동국대학교 영상대학원 멀티미디어학과 박사과정
[2]동국대학교 영상대학원 멀티미디어학과 교수

Comparative Evaluation Framework for AI-Based Video Motion Capture Tools

Ping-Jian Jie[1] · Jean-Hun Chung[2*]

[1]Doctoral's Course, Department of Multimedia, Graduate School of Digital Image & Contents, Dongguk University, Seoul 04620, Korea
[2]Professor, Department of Multimedia, Graduate School of Digital Image & Contents, Dongguk University, Seoul 04620, Korea

[요 약]

AI 기반 비디오 모션캡처 도구의 활용이 확산되면서, 다양한 도구의 성능 차이를 이해하는 것이 중요해지고 있다. 본 연구는 단일 카메라 환경에서 회전, 걷기, 쪼그리기, 점프, 손동작 등 다섯 가지 모션을 이용하여 Rokoko Vision, DeepMotion, Meshcapade의 성능을 비교하는 실용적 평가 프레임워크를 제시한다. 각 도구는 시각적 추적 품질, 제스처 인식, 모션 안정성, 사용 편의성을 기준으로 평가되었으며, DeepMotion은 상지·손동작 표현이 우수하고 Meshcapade는 전신 움직임의 해부학적 일관성이 높았다. Rokoko Vision은 접근성과 기본 추적 안정성이 강점으로 나타났다. 본 연구는 프리비주얼라이제이션, 소규모 스튜디오 제작, 애니메이션 교육 환경에서 적합한 도구 선택에 유용한 지침을 제공한다.

[Abstract]

With the rapid adoption of AI-based video motion capture tools in digital animation, understanding their comparative performance has become essential. This study presents a comparative performance analysis and practical evaluation framework for AI-based video motion capture tools. Rokoko Vision, DeepMotion, and Meshcapade were assessed for their perceived motion-tracking quality, gesture recognition, motion stability, and usability under standardized conditions using a single-camera setup and five representative motions, namely rotation, walking, squatting, jumping, and hand gestures. DeepMotion exhibited stronger upper-body and gesture fidelity, Meshcapade produced more anatomically consistent full-body motion, and Rokoko Vision offered stable baseline tracking with high accessibility. The proposed framework provides workflow-specific insights, particularly for previsualization, small-studio production, and animation education, thereby helping creators to select tools that align with their production goals.

색인어 : 모션 캡처, 로코코 비전, 딥모션, 메쉬카페이드, 애니메이션
Keyword : Motion Capture, Rokoko Vision, DeepMotion, Meshcapade, Animation

http://dx.doi.org/10.9728/dcs.2026.27.1.61

Received 26 October 2025; **Revised** 24 November 2025
Accepted 02 January 2026

***Corresponding Author; Jean-Hun Chung**
Tel: ※ 개인정보 표시제한
E-mail: evengates@gmail.com

Ⅰ. Introduction

The rapid integration of artificial intelligence (AI) into digital design and animation has revolutionized motion capture (MoCap) workflows. Traditionally, optical systems required expensive multi-camera setups and controlled studio lighting, while inertial systems depended on wearable sensors prone to drift and synchronization issues. In contrast, AI-based video solutions can now extract motion data directly from standard RGB footage, significantly reducing cost and complexity. These advancements democratize access to high-quality animation tools, empowering small studios and independent creators to adopt professional-grade pipelines.

Despite these developments, academic research comparing AI-based MoCap tools remains limited. Most prior studies emphasize algorithmic accuracy or hardware calibration rather than their performance in real production environments. Consequently, creators often rely on anecdotal evidence when choosing between tools, leading to uncertainty regarding accuracy, usability, and workflow efficiency.

To address this gap, this study systematically compares three widely used AI-driven motion capture tools—Rokoko Vision, DeepMotion, and Meshcapade—under standardized test conditions. The experiments analyze each tool's ability to capture various human motions, including rotation, walking, jumping, squatting, and gestures.

Therefore, this study aims to answer the following research questions (RQs):

RQ1: How do AI-based video motion capture tools perform in capturing diverse human motions using standardized video inputs, as evaluated through multi-rater visual assessments?

RQ2: What are the performance and usability differences among Rokoko Vision, DeepMotion, and Meshcapade?

RQ3: How can these findings guide efficient AI-integrated workflows for small studios and independent creators?

This research contributes both practical insights and academic significance by proposing a replicable comparative framework and connecting empirical analysis to the broader discourse on automation and creative efficiency in digital design.

Ⅱ. Theoretical Background of AI-Based Motion Capture

Motion capture (MoCap) refers to the digital recording and reconstruction of physical movements, enabling the translation of real-world performance into virtual environments. Traditional MoCap systems are generally categorized as optical or inertial. Optical systems provide sub-millimeter precision using infrared cameras and reflective markers but require costly studio setups and are sensitive to occlusion. Inertial systems utilize wearable sensors to track orientation and acceleration, offering portability but suffering from cumulative drift over time.

The emergence of AI-driven MoCap represents a paradigm shift. Deep learning-based pose estimation models can infer 2D and 3D skeletal structures directly from monocular RGB video without markers or suits. Foundational studies such as OpenPose[1] and VNect[2] demonstrated that convolutional neural networks (CNNs) can reconstruct full-body motion from single-camera input. More recent transformer-based architectures have further improved robustness to lighting variation, occlusion, and camera movement[3].

While AI MoCap increases accessibility, recent research shows that performance remains affected by environmental conditions (lighting, resolution, background) and performer characteristics (clothing contrast, body shape)[4]. Therefore, comparative evaluation under controlled conditions is essential for assessing practical usability in real-world animation pipelines.

In the contemporary animation industry, Rokoko Vision, DeepMotion, and Meshcapade have emerged as leading commercial solutions due to their affordability, accessibility, and integration with Unreal Engine, Unity, and MetaHuman. Market analyses by Zion Market Research[5] and Mordor Intelligence[6] indicate that AI-based motion capture technologies are among the fastest-growing segments in the animation sector, with significant adoption in independent and educational production environments.

These theoretical and market foundations provide the rationale for selecting these three tools as representative case studies for the present comparative analysis.

Ⅲ. Related Work

Previous research on motion capture (MoCap) technologies has evolved from traditional optical and inertial systems to data-driven and AI-based methods. Early studies such as Moeslund et al.[7] and Menache [8] laid the foundation for understanding motion tracking principles, calibration requirements[9], and limitations in conventional systems. These studies emphasized the trade-off between precision, setup complexity, and scalability, which remain relevant to the development of modern AI-based MoCap solutions.

With the advancement of deep learning, researchers began applying convolutional neural networks (CNNs) and transformer architectures to estimate human pose directly from images or videos. Mehta et al.[2] proposed VNect, a real-time 3D pose estimation framework from monocular video, marking a pivotal step toward markerless motion capture. Subsequent frameworks such as OpenPose[1] and AlphaPose expanded this approach to multi-person tracking, improving detection speed and joint coherence.

More recent surveys highlight how deep learning-based MoCap integrates with creative and industrial applications. Alemi and Pasquier[10] reviewed data-driven movement generation for animation and gaming, emphasizing how neural networks can automate motion style transfer and synthesis. Similarly, Mourot et al.[11] provided a comprehensive overview of skeleton-based animation using AI, identifying both potential and current limitations—such as spatial drift and temporal instability in AI-generated motion data.

Although these studies contribute to the theoretical understanding of AI motion modeling, few have conducted comparative evaluations of commercial AI MoCap tools under standardized creative conditions. Existing analyses typically focus on single-tool performance or algorithmic accuracy (e.g., OpenPose vs. BlazePose) rather than usability and production efficiency relevant to small-scale animation workflows.

In the field of design and digital media production, recent applied studies demonstrated the growing need for accessible MoCap tools to enhance creative efficiency in 3D design education and animation content creation. This aligns with industrial market reports[4],[12] indicating increasing adoption of AI-based MoCap tools like Rokoko Vision, DeepMotion, and Meshcapade for educational and small-studio production environments.

Therefore, this research expands on prior theoretical frameworks by providing a systematic, practice-oriented comparison of three commercial AI-driven MoCap platforms using consistent motion datasets, thereby addressing the methodological and empirical gaps observed in earlier literature.

Ⅳ. Research Methodology

This study conducts a comparative evaluation of three AI-driven video motion capture tools—Rokoko Vision, DeepMotion, and Meshcapade—to assess their performance in capturing human motions with varying complexity.

The research framework consists of four stages: experimental setup, data acquisition, evaluation criteria, and analysis.

4-1 Experimental Overview

The experiment was designed to analyze how each system performs when given identical motion input data. All three tools process video-based motion capture through single-camera monocular input, and the evaluation focused on motion accuracy, gesture recognition, path consistency, visual artifacts, and operational limitations.

A standardized test video dataset was produced for this purpose, containing five types of motion commonly used in animation workflows:

Rotation – horizontal turning motion for upper and lower body.

Circular walking – locomotion involving consistent foot planting and direction change.

Squatting – testing lower-body flexibility and joint tracking.

Jumping – vertical motion with body extension.

Hand gestures – fine movements assessing upper-limb tracking and recognition.

These motions were chosen based on the guidelines from previous motion capture research as representative categories to test both gross and fine motor performance[8].

For each motion type, a single representative clip was recorded under identical lighting, background, and camera settings, forming a compact but fully

standardized dataset that allows consistent cross-tool comparison and reproducibility.

4-2 Recording Setup

All recordings were conducted in a controlled indoor environment with stable ambient lighting and a neutral background to minimize visual noise and shadow interference.

The performer—a single female model (height: 165 cm)—was selected to provide a consistent baseline for skeletal tracking and to ensure clear visualization of limb articulation across all tools. Her body proportions are close to the average used in character animation studies, which facilitates consistent joint detection and minimizes variability caused by unusual body shapes. The performer wore a dark top and gray shorts to create sufficient contrast with the background, improving keypoint detection during motion capture. It should be noted that AI MoCap performance can vary significantly across different body types, genders, attire, and environmental conditions. Therefore, the results are limited to this specific setup and should not be generalized to all scenarios.

Video data were captured using a single 1080p RGB camera (60 fps) positioned approximately two meters in front of the performer at shoulder height, ensuring a balanced field of view that captures both upper and lower body movements without perspective distortion.

To ensure methodological consistency, the camera, spatial setup, and performer attire followed the official recommendations provided in Rokoko Vision's Capture Space Setup Guide, as well as technical guidance available from DeepMotion and Meshcapade documentation.

Each motion sequence was recorded under the same environmental conditions and subsequently uploaded to the official online platforms of the three tools.

No additional post-processing or external 3D software (e.g., Maya or Unreal Engine) was used; all analyses were conducted using each tool's built-in animation preview interface to maintain native performance evaluation and minimize external influence. Details of camera position and performer attire are further discussed in Section 4-5.

4-3 Data Processing and Evaluation Framework

After recording, the captured videos were uploaded separately to the official online platforms of Rokoko Vision, DeepMotion, and Meshcapade, each operating through browser-based interfaces.

All analysis was conducted using the native animation preview functions provided by each tool, rather than exporting data to external 3D software. This approach ensured that tool performance was evaluated in its intended environment, maintaining the integrity of the built-in motion estimation algorithms and real-world usability.

To evaluate the captured results, this study applied five qualitative and functional criteria, defined based on both prior research[10],[11] and practical animation production needs:

Joint Tracking Accuracy – Evaluates how precisely each system reconstructs the skeletal structure and maintains correct joint positioning throughout motion sequences.

Gesture Recognition – Measures the tool's ability to capture and reproduce expressive hand and arm gestures such as rotations, pointing, and "OK" signs.

Path Consistency – Assesses the spatial stability and continuity of locomotion (e.g., circular walking) to detect drift or directional error.

Visual Artifacts – Observes anomalies such as jitter, floating, or unnatural limb deformation during playback.

Operational Constraints – Examines the usability aspects of each system, including upload limits, processing time, and platform accessibility.

Each tool was tested under identical environmental conditions and camera parameters, ensuring methodological consistency.

All motion outputs were reviewed frame-by-frame within each tool's built-in 3D preview window at standard playback speed, allowing direct visual comparison across systems.

To ensure fairness across systems, all three tools were accessed exclusively through their web-based interfaces rather than desktop or mobile applications. Rokoko Vision provides both "Single Camera – Record" and "Single Camera – Upload" modes; however, to maintain methodological consistency, this study used the identical pre-recorded video uploaded to each platform.

Table 1. Evaluation criteria and scoring framework

Evaluation Criteria	Evaluation Focus	Definition	Scoring Standard (1–5)
Joint Tracking Quality	Skeletal alignment precision	Measures how well the system visually maintains skeletal alignment (head, spine, limbs) relative to the source motion	1 = Frequent misalignment → 5 = Highly precise and stable
Gesture Recognition	Hand and upper-body detail	Evaluates accuracy of expressive or symbolic motions (e.g., pointing, waving, "OK" gesture)	1 = Unrecognized / distorted → 5 = Fully preserved gesture fidelity
Path Consistency	Spatial trajectory reliability	Assesses stability and realism of motion paths (e.g., walking circle or jump arc)	1 = Severe drift → 5 = Smooth and realistic continuity
Visual Artifacts	Rendering and stability issues	Examines frame discontinuities, jittering, or floating artifacts in animation output	1 = Frequent artifacts → 5 = No visible instability
Operational Constraints	Tool usability and limitations	Considers processing speed, upload limits, hardware dependency, and ease of operation	1 = Highly restrictive → 5 = Efficient and user-friendly

Additionally, Rokoko Vision allows unlimited free recordings of up to 15 seconds in its single-camera mode, whereas DeepMotion and Meshcapade offer only limited free trial usage. As a result, Rokoko Vision's higher usability and efficiency scores partially reflect its more accessible trial conditions, while all functional comparisons were conducted under the same upload-based workflow to ensure evaluation fairness.

The evaluation used a five-point qualitative scale (1 = Poor to 5 = Excellent) to rank performance per criterion. Each standardized motion sequence lasted approximately 4–8 seconds and was recorded once per motion type under identical environmental conditions. During evaluation, the outputs from the performer and all three motion capture tools were displayed side by side within a single interface: first played once at normal speed, and then replayed at 0.25× slow motion. This procedure allowed evaluators to closely inspect joint behavior, gesture articulation, and trajectory continuity across tools. The playback order of motion types was randomized to reduce ordering effects, and all evaluators followed a shared scoring guideline describing the criteria for each level (1–5).

For each tool and criterion, the final score is the average of three evaluators' ratings, summarized in Table 2. These scores represent subjective visual assessments of motion quality rather than objective error metrics based on ground-truth data.

4-4 Evaluation Process

Each recorded motion sequence was evaluated based on the five qualitative criteria defined in Table 1.

The assessment focused on observable motion realism and tool responsiveness within the animation preview interfaces of each application, rather than external 3D engines. All tools were tested under identical video inputs to ensure comparability.

The evaluation procedure followed a step-by-step observation of body articulation, gesture detection, and trajectory reproduction. Each motion type (rotation, circular walking, squatting, jumping, and hand gestures) was visually analyzed to identify strengths and weaknesses in the generated animation. The scores were calculated as the average of three evaluators' ratings, each based on prior experience in digital animation. This averaging process enhanced inter-subjective reliability while preserving the qualitative nature of the assessment.

4-5 Camera and Clothing Configuration

All motion videos were captured using a single fixed RGB camera positioned approximately 2.5 meters from the performer at chest height, providing a balanced field of view to capture both upper and lower body movements without perspective distortion. The filming space followed the official setup recommendations provided by Rokoko Vision, DeepMotion, and Meshcapade to ensure methodological consistency.

The performer wore a dark top and gray shorts, creating sufficient contrast with the background to improve keypoint detection and minimize occlusion errors.

The chosen performer (single female model, height: 165 cm) was selected for her average body proportions, which ensure clear visualization of limb

articulation and consistent skeletal detection across all tools, facilitating reproducible evaluation under standardized conditions.

Lighting conditions were evenly distributed, and the background was plain and unobstructed to avoid interference with body segmentation. All captured data were processed within the tools' native animation preview systems, without further retargeting or post-processing in external software such as Maya or Unreal Engine.

Table 2. Evaluation protocol summary

Item	Specification
Camera setup	Single RGB camera, fixed position
Motion types	Rotation, walking, squatting, jumping, hand gestures
Motion duration	Approximately 4–8 seconds per motion
Takes	Single take per motion
Input method	Identical pre-recorded video uploaded to each tool
Playback	1× normal speed followed by 0.25× slow motion
View mode	Side-by-side comparison within the preview interface
Scoring	Five-point Likert scale, averaged across three evaluators

This protocol summary is provided to clarify the technical conditions under which the proposed qualitative comparison framework can be reproduced by other researchers.

V. Results and Analysis

This section presents the comparative evaluation of three AI-based, video-driven motion capture tools—Rokoko Vision, DeepMotion, and Meshcapade—based on the standardized motion set introduced in Section 4.3 and the five evaluation criteria: joint tracking accuracy, gesture recognition, path consistency, visual artifacts, and operational constraints.

Each result represents the average of three evaluators' visual assessments using a five-point Likert scale (1–5), based on qualitative observation of the output animations in each tool's built-in 3D preview interface. These scores reflect subjective perception of motion fidelity rather than objective computational accuracy.

All evaluators had prior experience in digital animation production and followed a shared scoring guideline during the evaluation process. Although no formal inter-rater reliability statistics were computed, the evaluators exhibited generally consistent rating tendencies across tools and motion types, with limited disagreement observed in most criteria. Accordingly, the reported scores should be interpreted as indicative comparative tendencies rather than precise or absolute measurements.

5-1 Overall Performance

As summarized in Table 3, each tool demonstrates distinct strengths depending on motion complexity and purpose of use.

DeepMotion achieved the highest overall performance, particularly excelling in gesture fidelity and spatial tracking.

Meshcapade delivered the most realistic body movement and spatial smoothness through its biomechanical modeling, while Rokoko Vision provided an accessible, cost-effective solution for broad body movement, despite lower accuracy in detailed gestures.

While mean scores are reported for clarity, some degree of score dispersion across evaluators is inherent to qualitative visual assessment.

5-2 Motion-Type Analysis

1) Rotation Test — Fig. 1

DeepMotion achieved the most stable hip and shoulder alignment with minimal drift (avg. 4.3), while Meshcapade produced smoother rotational balance through biomechanical correction (avg. 4.5).

Table 3. Average evaluation scores (1 – 5) across motion categories and criteria

Tool / Criterion	Joint Tracking Quality	Gesture Recognition	Path Consistency	Visual Artifacts	Operational Constraints	Average Score
Rokoko Vision	3.4	2.1	3.2	3.5	4.7	3.4
DeepMotion	4.2	4.8	4.6	4.1	3.2	4.2
Meshcapade	4.5	3.9	4.4	4.3	3.6	4.1

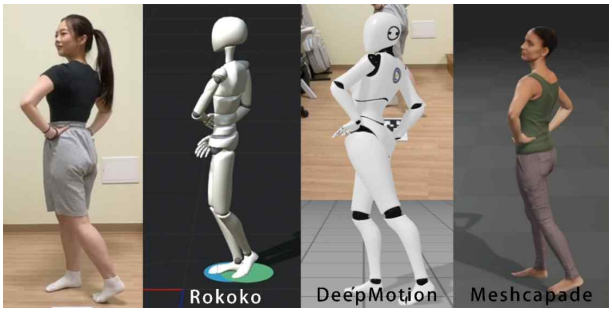

Fig. 1. Hands-on-hips rotation

Rokoko Vision maintained acceptable stability (avg. 3.2) but exhibited frame jitter in mid-turn frames.

2) Circular Walking — Fig. 2

DeepMotion showed precise foot locking and consistent trajectory tracking (avg. 4.7).

Fig. 2. Circular walking motion

Meshcapade achieved realistic arm swing and pelvic motion (avg. 4.5).

Rokoko Vision followed the circular path accurately (avg. 3.4) but produced slight shoulder offset near turning points.

3) Squatting — Fig. 3

All tools struggled with occlusion and depth ambiguity.

DeepMotion preserved temporal rhythm (avg. 4.0) but displayed ankle misalignment at the lowest position.

Meshcapade maintained body proportion (avg. 4.2) though occasionally produced floating feet, while Rokoko Vision lost lower-body tracking (avg. 2.8) when legs overlapped the torso.

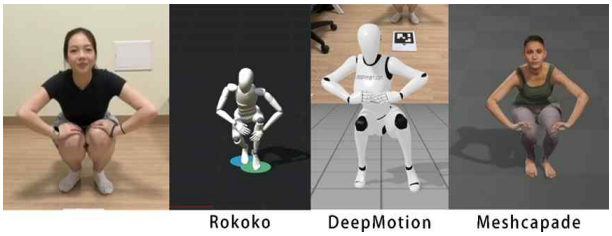

Fig. 3. Squatting

4) Jumping — Fig. 4

Rokoko Vision captured vertical motion and posture stably (avg. 4.0), suitable for dynamic, full-body actions.

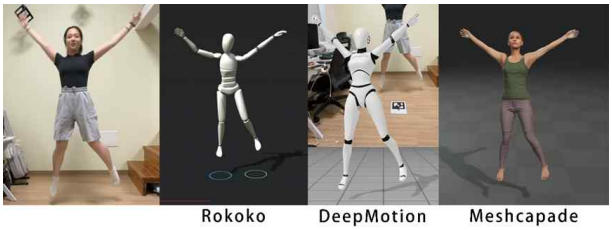

Fig. 4. Jumping

DeepMotion recorded takeoff-landing continuity with slight landing jitter (avg. 4.3).

Meshcapade generated the smoothest transition curve (avg. 4.4) but slightly exaggerated elevation due to smoothing bias.

5) Hand Gesture — Fig. 5

DeepMotion achieved the most accurate finger tracking (avg. 4.8), successfully recognizing small-scale gestures such as the "OK" sign.

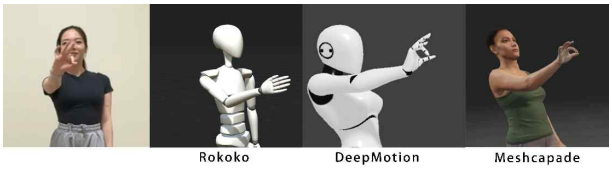

Fig. 5. OK gesture

Meshcapade kept proportional structure (avg. 3.9) but lacked fine articulation.

Rokoko Vision failed to capture individual fingers (avg. 2.1), focusing primarily on upper-body alignment.

Figs. 1–5. Visual Comparison of Output Animations for Five Standardized Motions

Each figure displays raw motion output from Rokoko Vision, DeepMotion, and Meshcapade for identical input sequences.

5-3 Comparative Findings

Overall, DeepMotion produced the highest combined visual accuracy and gesture fidelity as perceived by the evaluators, making it the most suitable tool for character animation requiring expressive detail.

Meshcapade, leveraging its biomechanical modeling, delivered visually natural results ideal for cinematic production and high-end visualization.

Rokoko Vision, though less detailed, provided stable large-body motion capture with minimal setup, making it advantageous for small studios or previsualization.

The performance differences among the three tools can be interpreted in light of their underlying technical architectures. DeepMotion, which utilizes a cloud-based deep learning pipeline with model ensembles, demonstrates stronger gesture recognition due to its higher-capacity inference models. Meshcapade, built upon the SMPL biomechanical body model, delivers more anatomically natural full-body motion and smoother joint transitions, reflecting its parametric model constraints. Rokoko Vision relies primarily on 2D keypoint detection followed by monocular pose lifting, which provides stable but less detailed reconstruction, particularly in fine hand motions. These architectural distinctions help explain the comparative strengths and weaknesses observed in the evaluation.

Despite these strengths, all three tools share common weaknesses in occlusion-heavy or self-contact motions (e.g., squats, crossed arms), indicating the persistent limitation of monocular AI-based systems in resolving depth ambiguity and multi-limb overlap.

These findings reaffirm the need for future studies exploring multi-view capture, improved pose inference algorithms, and hybrid quantitative-qualitative evaluation frameworks to advance AI-driven MoCap reliability.

Ⅵ. Discussion and Conclusion

6-1 Discussion

The evaluation framework draws upon theoretical foundations from prior MoCap assessment research and animation production studies. Joint tracking quality and gesture recognition align with established criteria used in computer vision–based motion analysis, which emphasize skeletal alignment, joint stability, and fine-grained articulation. Path consistency and visual artifacts reflect animation workflow concerns, where spatial drift, jitter, and deformation strongly impact retargeting and editing. Operational constraints and usability follow principles from HCI and digital content creation research, recognizing that accessibility, processing time, and workflow efficiency directly influence tool adoption in production environments. Accordingly, the five-criteria framework integrates both technical MoCap standards and practice-oriented evaluation dimensions.

The purpose of this study was to evaluate and compare three AI-based, video-driven motion capture tools—Rokoko Vision, DeepMotion, and Meshcapade—under standardized motion conditions to determine their motion-tracking performance, usability, and creative applicability in digital animation workflows.

A key methodological limitation of this study is the absence of objective ground-truth motion data. Because the test motions were not captured simultaneously with a marker-based MoCap system, and because Rokoko Vision, DeepMotion, and Meshcapade do not provide unified access to raw 3D joint coordinates, established quantitative metrics such as MPJPE or OKS could not be computed. These constraints arise from the structural design of commercial AI-based video MoCap tools rather than from experimental omission.

Despite this limitation, a visually oriented comparative evaluation holds practical relevance, as animation workflows—particularly in previsualization, education, and small-studio environments—depend primarily on perceived motion quality rather than marker-based numerical accuracy. Therefore, the qualitative framework adopted in this study reflects the conditions under which these tools are most commonly used in real production contexts.

The results demonstrate that while all three systems effectively replicate general body motion, their precision and consistency vary according to motion complexity and application context.

DeepMotion achieved the highest overall performance, particularly excelling in gesture recognition and path tracking. Its results confirm prior research that deep learning–based systems can reliably infer detailed limb and hand movement when trained on large video datasets.

Meshcapade, leveraging biomechanical modeling, demonstrated superior anatomical realism and

continuity, aligning with findings from motion reconstruction studies that emphasize the role of body-parameterized models such as SMPL.

In contrast, Rokoko Vision achieved practical efficiency and accessibility but showed limitations in fine-grained tracking accuracy, especially for occluded or self-contact motions.

These findings highlight the current trade-off in AI motion capture between precision and accessibility—while cloud-based and browser-based tools democratize motion capture, they still face fundamental constraints in spatial depth inference and occlusion handling.

From a design perspective, this suggests that tool selection should depend on production goals: DeepMotion offers expressive accuracy for gesture-intensive animation; Meshcapade provides smooth articulation for cinematic realism; and Rokoko Vision delivers usability with minimal setup for rapid previsualization.

These tool-specific recommendations should also be interpreted within the limitations of the single-performer, single-camera setup used in this study.

The introduction of a multi-rater evaluation protocol strengthened reliability by minimizing subjective bias, establishing a foundation for future quantitative benchmarking.

However, the study also reveals that current evaluation frameworks in AI motion capture research are largely qualitative.

While quantitative metrics such as MPJPE or OKS are commonly used in traditional marker-based MoCap research, they could not be applied in this study due to the lack of unified 3D joint coordinate outputs and the absence of simultaneous marker-based ground-truth capture. Future research may incorporate such metrics only if commercial AI-based systems begin providing standardized skeletal data formats or if dual capture setups are used. In such cases, combining quantitative measurements with perceptual quality assessments could more robustly connect computational accuracy with creative usability.

This mixed-methods approach would enhance the theoretical rigor of AI motion analysis and support standardized evaluation practices within digital design research.

6-2 Conclusion

This study provides an empirical framework for comparing AI-based motion capture tools in the context of animation and digital media production.

By applying standardized motion tests and a five-criteria evaluation model, it identifies the comparative strengths and weaknesses of three widely used systems—Rokoko Vision, DeepMotion, and Meshcapade.

Under the specific experimental conditions of this study—a single performer, a fixed indoor environment, and monocular RGB input—the findings indicate that DeepMotion excels in gesture fidelity, Meshcapade achieves natural motion continuity, and Rokoko Vision ensures ease of use and affordability. These conclusions should therefore be understood as context-dependent, rather than universally applicable across all user profiles and production scenarios.

Beyond its practical contribution, this research offers theoretical implications for AI-integrated design methodology, emphasizing how machine learning tools can reshape traditional workflows in animation and content creation.

The results also underline the necessity of developing hybrid evaluation systems that combine computational accuracy with aesthetic and design-based criteria, aligning the field of motion capture with broader digital design scholarship.

Nevertheless, limitations remain: the study was conducted with a single performer, a fixed indoor setting, and monocular video input.

Future research should expand to multi-camera environments, varied body types, and dynamic lighting conditions to improve generalizability.

Additionally, integrating real-time motion capture data with engines such as Unreal Engine 5 and MetaHuman will allow further exploration of practical animation pipelines.

Ultimately, this study contributes to the ongoing dialogue between AI engineering and design research, offering both empirical data and methodological insights that can inform future developments in AI-assisted character animation and digital design.

As AI-driven performance capture continues to evolve, the convergence of computer vision and design methodologies will define the next paradigm of digital character creation.

Based on these comparative findings, scenario-specific recommendations can be made for practical animation and educational contexts. Rokoko Vision is suitable for entry-level training, classroom demonstrations, and rapid previsualization due to its accessibility and minimal setup requirements. DeepMotion is recommended for projects that demand expressive upper-body or hand-driven performances, such as character acting or stylized animation. Meshcapade is preferable for productions that prioritize anatomically realistic motion and stable full-body trajectories, making it well suited for realistic character work and research-oriented environments. These guidelines may assist educators, independent creators, and small studios in selecting tools that align with their production goals and resource constraints.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172-186, 2021. https://doi.org/10.1109/TPAMI.2019.2929257

[2] D. Mehta, S. Sridhar, O. Sotnychenko, H. Phodin, M. Shafiei, H.-S. Seidel, ... and C. Theobalt, "VNect: Real-Time 3D Human Pose Estimation with a Single RGB Camera," *ACM Transactions on Graphics (TOG)*, Vol. 36, No. 4, 44, 2017. https://doi.org/10.1145/3072959.3073596

[3] N. Kolotouros, G. Pavlakos, M. Black, and K. Daniilidis, "Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, pp. 2252-2261, 2019. https://doi.org/10.1109/ICCV.2019.00234

[4] A. Mathis, S. Schneider, J. Lauer, and M. W. Mathis, "A Primer on Motion Capture with Deep Learning: Principles, Pitfalls, and Perspectives," *Neuron*, Vol. 108, No. 1, pp. 44-65, 2020. https://doi.org/10.1016/j.neuron.2020.09.017

[5] Zion Market Research, AI Tools for Animation Market Growth, Size, Share, Trends, Zion Market Research, New York, Technical Report ZMR-AITA-2024, 2024.

[6] Mordor Intelligence, 3D Motion Capture Market Analysis: USD 281.85 Million in 2025, Projected 13.5% CAGR, Mordor Intelligence Research Center, Hyderabad, Technical Report MI-3DMC-2025, 2025.

[7] T. B. Moeslund, A. Hilton, and V. Krüger, "A Survey of Advances in Vision-Based Human Motion Capture and Analysis," *Computer Vision and Image Understanding*, Vol. 104, No. 2-3, pp. 90-126, 2006. https://doi.org/10.1016/j.cviu.2006.08.002

[8] A. Menache, *Understanding Motion Capture for Computer Animation*, San Francisco, CA: Morgan Kaufmann, 2010.

[9] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, Long Beach: CA, pp. 7745-7754. https://doi.org/10.1109/CVPR.2019.00794

[10] O. Alemi and P. Pasquier, "Machine Learning for Data-Driven Movement Generation: A Review of the State of the Art," arXiv:1903.08356, 2019. https://doi.org/10.48550/arXiv.1903.08356

[11] L. Mourot, L. Hoyet, F. Le Clerc, F. Schnitzler, and P. Hellier, "A Survey on Deep Learning for Skeleton-Based Human Animation," *Computer Graphics Forum*, Vol. 41, No. 1, pp. 122-157, 2022. https://doi.org/10.1111/cgf.14426

[12] DataInsightsMarket, AI 3D Model Generators Future Pathways: Strategic Insights to 2033, DataInsights Institute, Singapore, Technical Report DIM-505601, 2025.

갈평건(Ping-Jian Jie)

2022년 : 동국대학교 공학대학 멀티미디어공학 (학사)

2024년 : 동국대학교 영상대학원 멀티미디어학과 콘텐츠디자인 (석사)

2024년～현 재: 동국대학교 영상대학원 멀티미디어학과 박사과정

※관심분야 : AI 아트, 애니메이션, XR, VFX, 메타버스 등

정진헌(Jean-Hun Chung)

1992년 : 홍익대학교 미술대학 시각디자인 BFA

1999년 : 미국 Academy of Art University, Computer Arts MFA

2001년～현 재: 동국대학교 영상대학원 멀티미디어학과 교수

※관심분야 : AI 아트, 컴퓨터 아트, 뉴미디어디자인, XR, VFX, 웹툰, 메타버스 등