Liu *et al. Art Int Surg* 2024;4:92-108
DOI: 10.20517/ais.2024.19

Artificial Intelligence Surgery

Original Article

Open Access

Check for updates

A human mesh-centered approach to action recognition in the operating room

Benjamin Liu[1], Gilles Soenens[2], Joshua Villarreal[3], Jeffrey Jopling[4], Isabelle Van Herzeele[2], Anita Rau[5,#], Serena Yeung-Levy[5,#]

[1]Department of Computer Science, Stanford University, Stanford, CA 94305, USA.
[2]Department of Thoracic and Vascular Surgery, Ghent University Hospital, Gent 9000, Belgium.
[3]Department of Surgery, Stanford University, Stanford, CA 94305, USA.
[4]Department of Surgery, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.
[5]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA.
[#]Authors contributed equally.

Correspondence to: Benjamin Liu, Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA. E-mail: bencliu@cs.stanford.edu

How to cite this article: Liu B, Soenens G, Villarreal J, Jopling J, Van Herzeele I, Rau A, Yeung-Levy S. A human mesh-centered approach to action recognition in the operating room. *Art Int Surg* 2024;4:92-108. https://dx.doi.org/10.20517/ais.2024.19

Received: 14 Mar 2024 **First Decision:** 15 May 2024 **Revised:** 25 May 2024 **Accepted:** 11 Jun 2024 **Published:** 30 Jun 2024

Academic Editor: Andrew A. Gumbs **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

Abstract

Aim: Video review programs in hospitals play a crucial role in optimizing operating room workflows. In scenarios where split-seconds can change the outcome of a surgery, the potential of such programs to improve safety and efficiency is profound. However, leveraging this potential requires a systematic and automated analysis of human actions. Existing methods predominantly employ manual methods, which are labor-intensive, inconsistent, and difficult to scale. Here, we present an AI-based approach to systematically analyze the behavior and actions of individuals from operating rooms (OR) videos.

Methods: We designed a novel framework for human mesh recovery from long-duration surgical videos by integrating existing human detection, tracking, and mesh recovery models. We then trained an action recognition model to predict surgical actions from the predicted temporal mesh sequences. To train and evaluate our approach, we annotated an in-house dataset of 864 five-second clips from simulated surgical videos with their corresponding actions.

Results: Our best model achieves an F1 score and the area under the precision-recall curve (AUPRC) of 0.81 and 0.85, respectively, demonstrating that human mesh sequences can be successfully used to recover surgical actions

from operating room videos. Model ablation studies suggest that action recognition performance is enhanced by composing human mesh representations with lower arm, pelvic, and cranial joints.

Conclusion: Our work presents promising opportunities for OR video review programs to study human behavior in a systematic, scalable manner.

Keywords: Action recognition, human mesh recovery, operating room, surgery, artificial intelligence, computer vision, deep learning

INTRODUCTION

In recent years, video review programs in hospitals have grown rapidly, particularly within critical settings such as intensive care units, trauma bays, and operating rooms (OR)[1]. These programs use video recordings, which serve as a veritable source of truth and offer insights into case challenges, systematic flaws in operating workflows, and opportunities for improvement. Healthcare providers can leverage these insights to design safer and more efficient interventions. In environments where mere seconds can significantly alter a patient's life course, the integration of video review programs holds enormous potential to improve patient outcomes.

Realizing this potential in a scalable, efficient way requires a systematic approach to video review that enables the granular analysis of movements, spatial dynamics, and actions made by human subjects. However, this vision is untenable with the manual methods that dominate modern programs. Manual review of OR videos is labor-intensive and difficult to perform systematically. Previous studies focusing exclusively on the analysis of OR movements required several mobility experts to review videos individually, discuss observations, and consolidate findings[2-5]. Extrapolating objective insights on team performance presents even more substantial challenges, as communications and team dynamics can be subtle despite their overwhelming importance to a successful operation[6]. While human analysis falls short, artificial intelligence (AI) algorithms are equipped to identify such subtle human motions in an efficient, scalable manner. Conventional AI-centered approaches process videos or images holistically[7]. However, visual cues differ between ORs across various institutions and specialties, potentially leading to model overfitting in low data regimes. Instead of processing videos directly, we thus leverage human meshes in sequence to analyze movements and actions in the OR.

Human mesh recovery (HMR) is a rapidly emerging technique for estimating detailed 3D human body meshes from 2D images. HMR harnesses deep learning architectures and parametric human body models to capture the shape and pose of a person. Recent increases in available high-resolution 3D motion capture data, alongside significant advances in HMR methods[8-11], present a compelling opportunity for the systematic analysis of OR videos. Resulting shape and pose estimates can be used to derive high-fidelity human meshes, providing a basis for studying underlying human behaviors based on the change of an individual's mannerisms and associated poses throughout time. For example, a common prelude to a human greeting may involve the extension of one's hand, the quick turn of one's neck, or the opening of one's upper arms for an embrace. All these actions can be interpreted clearly with the progression of arm and neck joints from estimated human meshes. Previous studies have applied HMR to diverse simulated and real-world settings, such as the analysis of striking techniques in sports, the reconstruction of clothed human bodies, and the modeling of avatars in virtual reality environments[12].

Despite their potential for analyzing behavior-rich scenes, HMR-based methods have yet to be explored for analyzing human behavior across long time frames (i.e., more than one minute). Several studies have developed temporal-based approaches to HMR; however, these methods are limited to short videos spanning less than one minute due to computational constraints[11]. Similar studies have investigated frame-based approaches, but focus on either frames with a single human subject or singular frames with multiple people[12]. Few studies have leveraged HMR techniques to analyze group dynamics, individual behavior, and global movements. Furthermore, to our knowledge, no previous studies have investigated the development of HMR-based methods to analyze human behavior in OR videos.

We propose an HMR-based computer vision framework for detecting, recovering, and tracking human meshes in surgical simulation videos. Our framework integrates a dual human head-body detector[13], a Kalman filter-based tracker[14], and a frame-based HMR model[15] trained on accessible, large-scale human mesh and human detection datasets[16-19]. Our framework presents a unified approach to studying human behavior in surgical scenes by deriving metrics on human attention, human movement, and hand-tool interactions from a small dataset of simulated surgical videos. To evaluate the potential of leveraging our estimated human mesh sequences for downstream surgical prediction tasks, we trained and evaluated a customized multi-layer perceptron (MLP) Mixer model on a self-curated dataset of human mesh sequences annotated with common, short-duration surgical actions. We show that sequences of mesh embeddings can be leveraged successfully to discriminate between actions with similar physical behaviors yet striking differences in surgical significance. Overall, our work advances efforts in systematizing OR video review for the study of human behavior with HMR.

METHODS

We designed an integrated, scalable method to identify individual actions and analyze individual behavior from OR videos [Figure 1]. In the following sections, we describe each successive component of our method in detail.

Human mesh recovery framework

The analysis of each individual's behavior in the OR requires three key steps: robustly detecting human subjects in each frame, tracking human subjects from frame to frame, and recovering human mesh parameters from each detected individual [Figure 1A].

Human and head detection

We performed human detection to identify regions in the image that our HMR model should focus on to recover human mesh features (Figure 2, middle column). We obtained a YoloV5 model[20] pretrained on the COCO 2017 dataset[17] and finetuned it separately on the detection of whole human subjects and human heads using the CrowdHuman dataset[13]. To improve the precision of our detections, we cross-referenced each human subject prediction with a prediction of an associated human head attained from a second object detection model.

Subject tracking

We leveraged these predictions in subsequent tracking with a simple Kalman Box-based tracker[14]; tracking results were used to associate human meshes in each frame with meshes in prior frames, allowing us to construct a sequential view of each individual's changes in movement and pose throughout time. To improve the fidelity of our tracking procedure, we introduced constraints on the nature in which a tracklet can be abandoned or created based on its most recent estimated position. We defined a tracklet as a temporal sequence of consecutive human mesh observations, associated with a single individual in the scene.

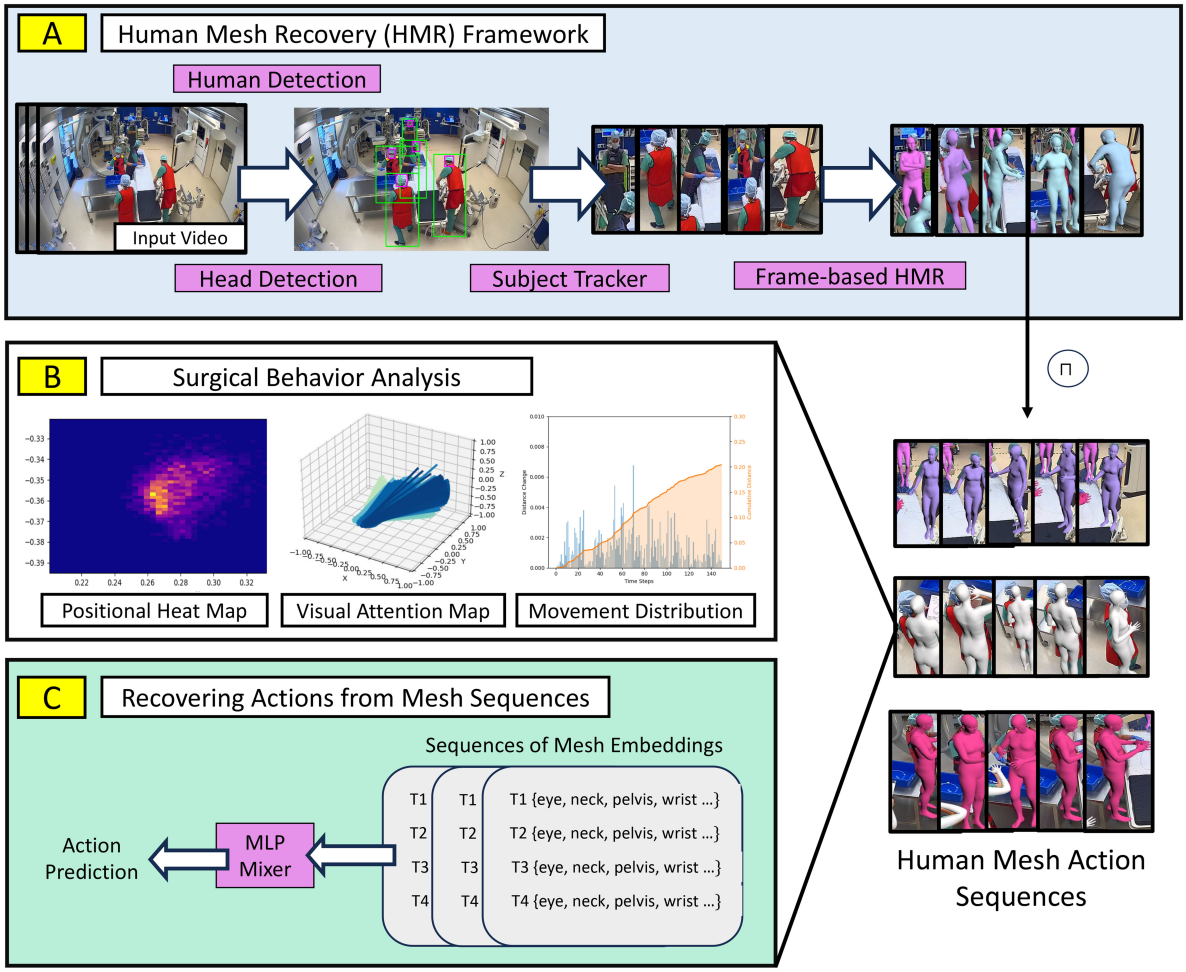

Figure 1. Overview of our framework.

Frame-based human mesh recovery

To recover human meshes from video frames (Figure 2, right column), we adopted the architecture and training procedure proposed by Li *et al.*, with one important deviation[15]. We swapped the Skinned Multi-Person Linear (SMPL) parametric model with SMPL eXpressive (SMPL-X) to extend the mesh recovery process to include more granular representations of hand and face joints[21]. We achieved comparable evaluation scores in Mean Per Joint Position Error (MPJPE), Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE), and Per Vertex Error (PVE) when evaluating on the benchmarks outlined in the study, which verified our trained model.

Surgical behavior analysis

Collating the results of our tracking and HMR procedures, we next performed a comprehensive set of qualitative analyses to interpret the movements and behaviors of individuals in the simulated scenes throughout time [Figure 1B]. In our analysis, we focused on metrics that were both significant to understanding the effectiveness of an OR procedure and clearly discernible from physical pose alone (i.e., sharp turn of neck to indicate attention switch). Specifically, we focused on:

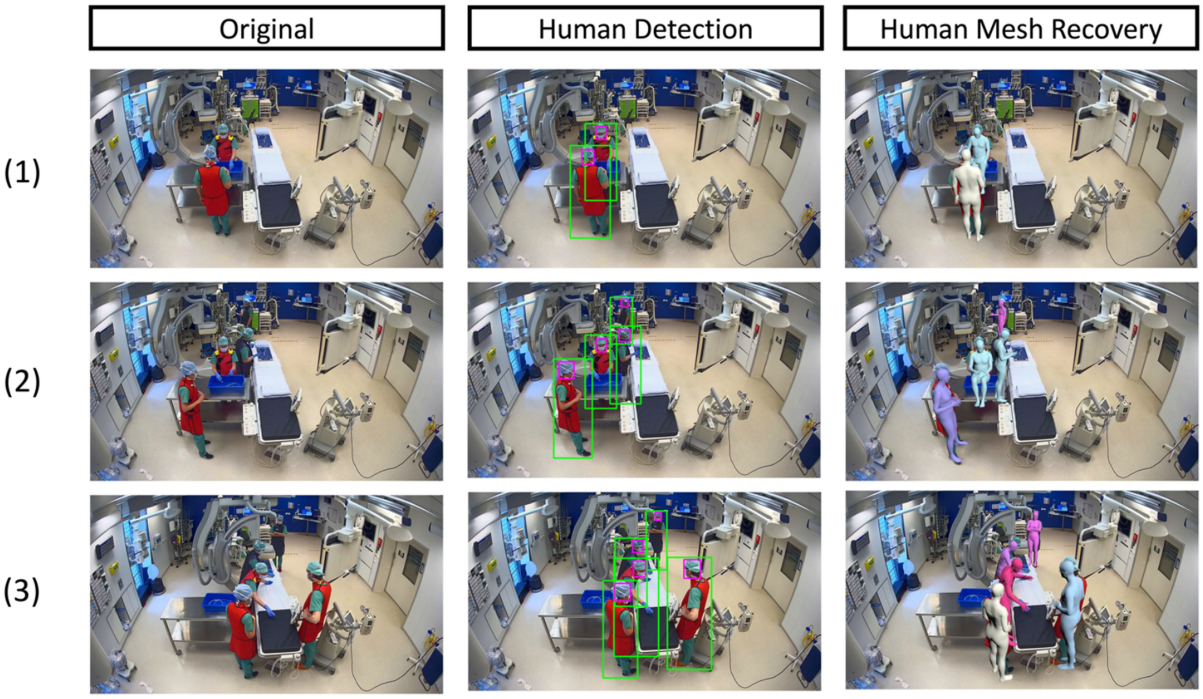

Figure 2. Overlay of original images (left) with corresponding outputs of human detection (middle) and HMR models (right) across different stages of the simulated surgery. HMR: Human mesh recovery.

(1) Movement patterns and distance traversal, in accordance with the association between movements and surgical stage transitions;
(2) Changes in positioning relative to the room due to the collective emphasis on optimal OR layout and minimization of collision points to facilitate smooth flow patterns[5];
(3) Visual attention switches over time due to the importance of task focus in the OR[22].

Movements and positioning

To analyze movement patterns and changes in positioning, we approximated the position of each mesh in each frame by the predicted pelvis joint of the mesh. Using these estimated positions, we constructed OR flow maps and associated heat maps to visualize tracklet trajectories and compute statistics on movement patterns, such as cumulative distance traversed.

Visual attention

To compute the visual attention field of an individual i at timestep t, we firstly calculated the midpoint $\mu \in \mathbb{R}^3$ between the left eye joint $L \in \mathbb{R}^3$ and right eye joint $R \in \mathbb{R}^3$. With neck joint $N \in \mathbb{R}^3$, we then constructed a plane $P \subset \mathbb{R}^3$ that was perpendicular to L, R, and N while simultaneously anchored by μ. Finally, we normalized across the resultant plane to obtain our view direction $v_i^t \in \mathbb{R}^3$. To detect a potential attention switch at timestep $t2$, $AS_i^{t2} \in \{0, 1\}$, we measured cosine similarity between viewing directions v_i^{t1} and v_i^{t2} where $t1$ and $t2$ represent timesteps with a difference of one-third of a second. We defined a switch in attention as a direction change of more than 45 degrees.

Recovering actions from mesh sequences

To demonstrate the utility of our HMR framework to downstream surgical prediction tasks that rely on a physical understanding of the scene, we trained and evaluated a deep learning model to perform a multi-

class classification task. Specifically, our model predicts the action associated with a mesh sequence [Figure 1C].

Architecture and experimental design

We leveraged a customized MLP Mixer model for our action recognition task. An MLP-Mixer model leverages MLPs to process channel-wise and token-wise information, allowing it to capture complex interactions among channels and patches for the modeling of image-based inputs[23]. Follow-up studies have applied these architectural principles to successfully model dependencies in non-image modalities and sequential data[24,25], demonstrating the suitability of the architecture for our action recognition task; we aimed to separately capture relationships among (1) different joints of a single subject in a given frame and (2) joints in sequential frames. In our experiments, we adapted the original MLP-Mixer architecture to accept an input sequence of human mesh-based embeddings by discarding the image patch layer and performing token mixing across the temporal and embedding dimensions of the human mesh input sequence. We defined mesh-based embedding as a vector representation of a human in a frame by any combination or subset of HMR-derived parameters, including estimated 3D joint poses and positions. Each mesh-based embedding effectively captures information on how the individual can be physically modeled at a specific point in time. The temporal dimension brings together mesh-based representations in sequence, which, we argue, can collectively represent a specific action, gesture, or behavior. In each training step, the MLP Mixer model takes in an input sequence of human mesh-based embeddings, representing a subject's physical motion across a 5-second clip, and outputs a predicted action class from the options of (1) hand-tool interaction, (2) walking movement, and (3) visual observation of peers.

Throughout initial experiments, we explored training with different numbers of mixer layers, learning rates, optimization algorithms, and mesh representation strategies. For all experiments, we used an unweighted cross-entropy loss function,

$$L(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (1)$$

Where y_c is the real class label for the mesh sequence, and $\hat{y}_c$ is the predicted confidence score for the designated class.

To understand the representative power of the HMR-derived parameters in distinguishing short-duration, common surgical actions, we experimented with different formulations of mesh-based embeddings. Specifically, we performed an ablation study where mesh-based embeddings were constructed solely from the 3D positions of joints from major joint categories, such as the pelvic, thorax, and cranial joints for the same task. Each joint set $J = \{j_1, ..., j_n\}$ was composed of n joints for which each joint $j_i \in \mathbb{R}^3$ represents the 3D position of the joint in the global scene, and the corresponding mesh embedding is a concatenation of all $j_i \in J$. We performed a similar ablation study, where mesh-based embeddings were represented strictly with predicted 3D joint poses rather than 3D joint positions. Specifically, we collected pose parameters in accordance with the joint categories defined previously. Each pose set $P = \{p_1, ..., p_n\}$ was composed of a concatenation of n flattened pose vectors, for which each $p_i \in \mathbb{R}^{3\times3}$ is defined by a rotation matrix that represents the pose of joint j_i. We performed follow-up studies looking into the performance effects of ablating specific joints that are crucial for performing hand-tool interactions and computing visual attention. Furthermore, we also studied the dependency of our approach on the rate of video frame sampling to provide insight into the scalability of our method to videos with longer durations.

We trained our MLP-mixer model for 200 epochs with a learning rate of 1e-4, batch size of 16, and an Adam Optimizer with default parameters[26]. During training, we selected best-performing models based on the reported F1 score during validation at the end of each training epoch. We ceased training if this metric did not improve over 50 epochs. Training was performed on a single GeForce RTX 2080 and completed in approximately one hour.

Evaluation metrics

We determined the predicted action class for each mesh sequence by selecting the class with the highest corresponding predicted probability. Our most important performance metrics included (1) precision, which quantifies the ratio of predicted images that correctly conform to a specific action class; (2) recall, which quantifies the ratio of mesh sequences correctly designated to a specific action class; and (3) F1, which combines precision and recall using a harmonic mean. We defined precision, P_c, and recall, R_c, for class c as:

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{2}$$

and

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{3}$$

where TP_c, FP_c, and FN_c denote true positives, false positives, and false negatives corresponding to a given action class. To provide more comprehensive measures of model performance, we calculated the area under the precision-recall curve (AUPRC). For all metrics, we computed a weighted average based on class prevalence.

RESULTS

This section describes the qualitative and quantitative insights into our framework's ability to analyze surgical behavior and recover short-duration actions from human mesh sequences of OR videos.

Datasets

To train our HMR model, we use a broad set of commonly used, open-access HMR datasets. As no surgical HMR datasets exist, to the best of our knowledge, we employed diverse datasets from general settings. We followed the widely referenced schema outlined by Kolotouros *et al.* for querying examples[16] from the Common Objects in Context (COCO) dataset and the Max Planck Institute for Informatics (MPII) Human Pose dataset along with associated 2D keypoints[17,27]. We also added examples and 3D ground truth from the 3D Poses in the Wild (3DPW) and Human 3.6M (H36M) datasets. We conducted our evaluation on the official train/test data splits of 3DPW, an in-the-wild dataset capturing humans in diverse poses and camera angles, and H36M, which captures human activities in controlled environments[18,19].

For human detection, we train our model on CrowdHuman, a large, richly annotated dataset of human subjects in crowded, natural scenes to mimic the crowded nature of OR scenes[13].

We curated an in-house dataset based on simulated surgical videos for experiments on surgical behavior analysis. These videos replicated actions in the OR by real clinical personnel but did not employ actual patients or procedures. Accordingly, our data do not include Protected Health Information (PHI) and do

not require de-identification, such as blurring of faces. All visible subjects provided written formal consent to be recorded and agreed to the usage of the data for this research. We analyzed eight simulated surgical videos with a total runtime of approximately 40 min with our integrated HMR framework. Videos were gathered from multiple perspectives from a single hybrid OR and depicted team members, including a surgeon, scrub nurse, circulating nurse, and anesthesia nurse, entering the room, preparing the OR table along with associated technical instruments, and engaging in attentive hand-tool movements to mimic a real endovascular procedure. To demonstrate the utility of the derived HMR features in modeling human behaviors, we curated a dataset of 5-second clips with discernible, common actions exhibited in endovascular surgery. Specifically, we derived tracklet sequences from our simulated videos for each human subject, which we further separated into 864 5-second clips. We manually annotated each clip with common surgical actions, including (1) hand-tool interaction, (2) walking movement, and (3) visual observation of peers, ensuring that actions were mutually exclusive for each clip in our dataset. Our curated action dataset included 313 examples of "hand-tool interaction", 91 examples of "walking movement", and 460 examples of "visual observation of peers" [Table 1].

Surgical behavior analysis

We performed a qualitative analysis of surgical scenes with mesh-derived visual attention, positioning, and movement metrics to enhance our understanding of how human behavior emerges from human mesh-based representations.

Movements and positioning

In comparing the positional heat maps displayed between different individuals, we found that individuals engaging in hand-tool interactions (normally near the operating table) have distinctly concentrated positional heatmap signatures compared to individuals engaging in walking movements [Figure 3]. While the positional heatmaps signatures can vary in concentration for individuals directly observing peer activities, they are not always clearly discernable from those of individuals engaging in hand-tool interactions. We observe a similar trend relative to our graphical comparisons of movement patterns, noting a substantial increase in movement patterns in walking movement clips compared to all other clips [Figure 4].

Visual attention

Analysis of the visual attention profiles of individuals revealed that the distribution of visual attention shows significant differences when a given subject is engaging in hand-tool interactions, walking movements, and observation of peers [Figure 5]. Unlike the positional heatmaps and movement pattern graphs, the visual attention maps displayed clear qualitative differences in the dispersion of attention between subjects engaging in hand-tool interactions and observations of peers.

Recovering actions from mesh sequences

Motivated by our qualitative observations of action-specific differences in visual attention, positional metrics, and movement metrics, we leveraged sequences of mesh-based embeddings for the classification of common surgical actions from 5-second tracklet clips.

Experiments on the choice of mesh embedding representation showed that composing mesh embeddings from 3D joint positions improved model performance in the F1 score, precision, and recall by 0.03, 0.04, and 0.04, respectively, compared to representing mesh embeddings as 3D joint poses (Table 2, bolded entries). In both representation strategies, we observed notable performance improvements with the new inclusion of joints from the "cranial" and "arm" categories, with minor performance differences seen in the further inclusion of joints in the "thorax", "spine", and "leg" categories. In our experiments with 3D joint

Table 1. Breakdown of our dataset containing short-duration action clips for downstream surgical task evaluation leveraging recovered human meshes

Action type	Train	Validation	Test	All splits
Hand-tool interaction	219	46	48	313
Walking movement	64	13	14	91
Visual observation of peer(s)	322	68	70	460
All types	605	127	132	864

We separately used subsets of four, two, and two simulated surgical videos to create our train, validation, and test splits, respectively. We learned MLP mixer model parameters using our training set, tuned hyperparameters with our validation set, and evaluated our model on our held-out test set. MLP: Multi-layer perceptron.

Table 2. Performance of our multi-class classification model under ablations that form the mesh embeddings separately from 3D joint positions (top) and 3D joint poses (bottom)

Pelvic	Arm	Cranial	Thorax	Spine	Leg	Recall↑	Precision↑	F1↑	AUPRC↑
Mesh embeddings as 3D joint positions									
√	-	-	-	-	-	0.62	0.38	0.47	0.57
√	√	-	-	-	-	0.75	0.72	0.73	0.74
√	√	√	-	-	-	**0.83**	**0.82**	**0.81**	**0.85**
√	√	√	√	-	-	0.82	0.80	0.81	0.81
√	√	√	√	√	-	0.78	0.78	0.77	0.74
√	√	√	√	√	√	0.73	0.73	0.72	0.71
Mesh embeddings as 3D joint poses									
√	-	-	-	-	-	0.75	0.75	0.75	0.72
√	√	-	-	-	-	0.78	0.77	0.77	0.77
√	√	√	-	-	-	0.78	**0.78**	**0.78**	0.83
√	√	√	√	-	-	0.78	0.77	0.77	0.81
√	√	√	√	√	-	0.78	0.77	0.77	**0.85**
√	√	√	√	√	√	**0.79**	0.77	0.78	0.81

Both ablations rely on the same major categories of joints, and check marks indicate that parameters from the joints in the referenced category are used to form the mesh embedding. For example, in the second row of the top table, 3D positions of the joints categorized under the pelvic and arm regions [Supplementary Material] are concatenated together to form the mesh embedding in each frame. Mesh embeddings from sampled frames in the 5-second action clip are collated, forming one dataset example, together with its associated action class label. Bolding indicates a top score.

poses, we observed less variance in model performance among joint categories, with 0.04, 0.03, and 0.03 as the maximal differences between the lowest and highest performing experimental settings in the metrics of recall, precision, and F1, respectively (Table 2, bottom).

Based on the results observed in Table 2, we performed further experiments to understand the contributions of specific joints to modeling the action recognition task, using 3D joint positions to construct mesh embeddings. Specifically, we performed ablation of joints in the "cranial" and "arm" joint categories, such as the wrists, elbows, eyes, ears, and head joints, while including joints from the "pelvic" joint category as a positional anchor. We chose to ablate joints from these specific categories, because these categories were previously observed to introduce substantial gains in model performance [Table 2]. We observed that optimal performance was achieved only after all individual cranial joints were included (Table 3, row 5). An ablation of pelvic joints from mesh embeddings that included all arm and cranial joints saw a considerable decrease in performance from its non-ablated baseline (Table 3, row 6).

Table 3. Performance of our multi-class classification model observed when we ablate the inclusion of key individual joints that are central to modeling lower-arm orientations and computing visual attention

Pelvic joints	Arm joints		Cranial joints						
All	**Wrists**	**Elbows**	**Eyes**	**Head**	**Ears**	**Recall↑**	**Precision↑**	**F1↑**	**AUPRC↑**
√	√	-	-	-	-	0.77	0.74	0.75	0.78
√	√	√	-	-	-	0.73	0.72	0.72	0.78
√	√	√	√	-	-	0.80	0.79	0.79	0.80
√	√	√	√	√	-	0.79	0.79	0.79	0.82
√	√	√	√	√	√	**0.83**	**0.82**	**0.81**	**0.85**
-	√	√	√	√	√	0.77	0.75	0.76	0.82

Embeddings are modeled as 3D joint positions. Bolding indicates a top score. AUPRC: The area under the precision-recall curve.

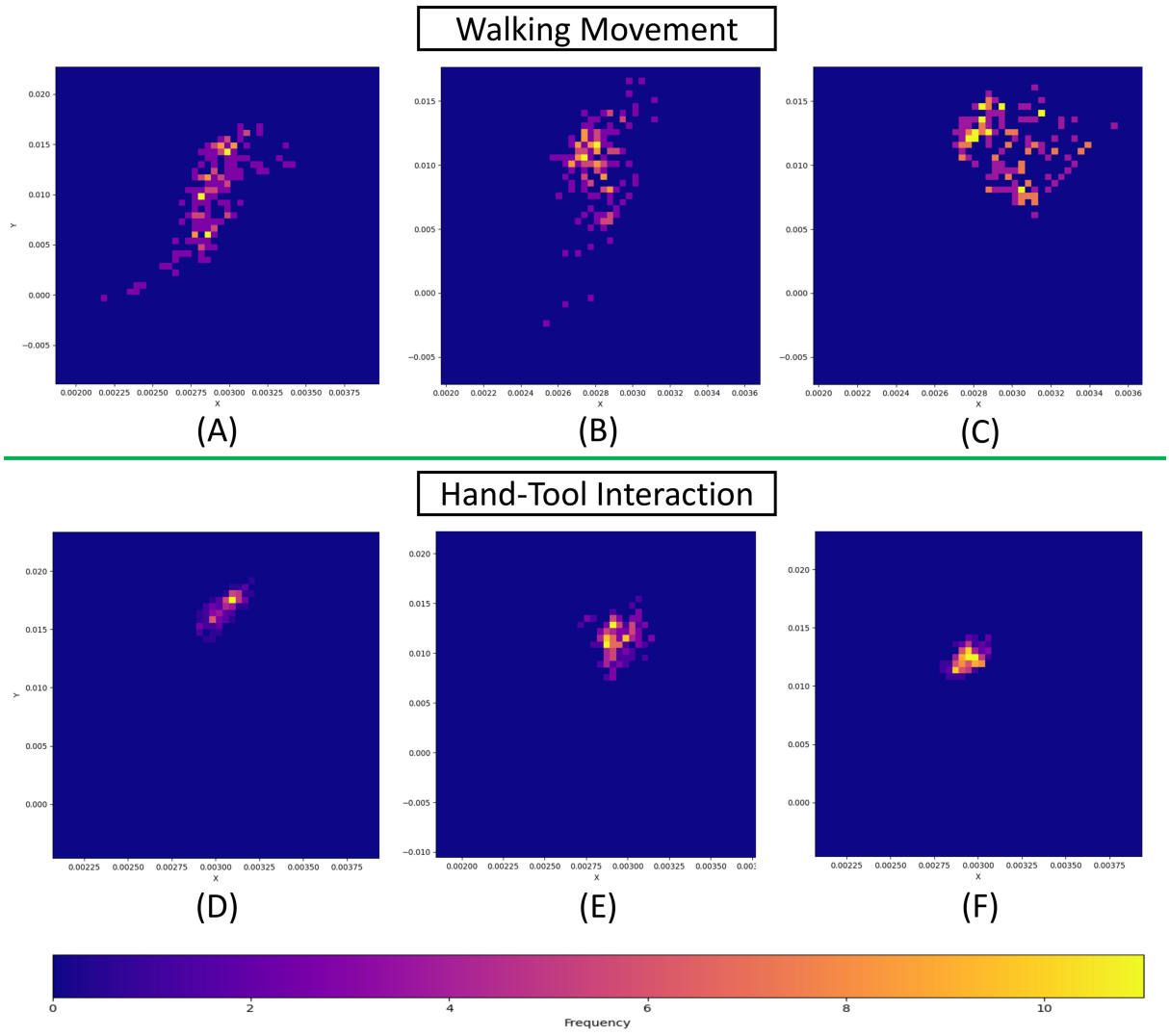

Figure 3. Comparisons of positional heat maps among tracklets engaging in walking movements (A-C) and tracklets engaging in hand-tool interactions (D-F). Tracklets engaging in walking movements (A-C) are more positionally dispersed, represented by the wide spread of their positional heat signature, while tracklets engaging in hand-tool interactions (D-F) are more visibly concentrated in a position close to the operating table.

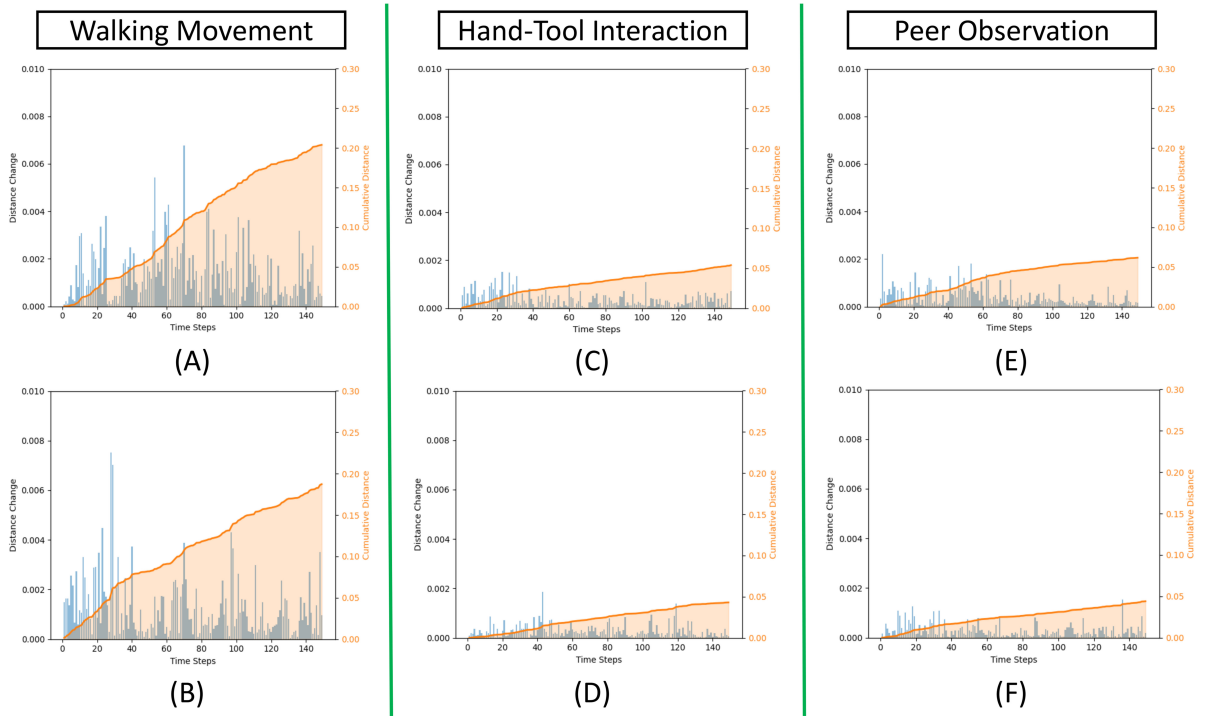

Figure 4. Graphical comparisons of distance traversal patterns among individual tracklets engaging in unique actions captured by 5-second clips. Subjects engaging in walking movements (A and B, left column) exhibit substantially higher levels of cumulative and local distance changes compared to individuals engaging in hand-tool interactions (C and D, middle column) and in observation of peer activities (E and F, right column).

Lastly, we performed an experiment to analyze the effect of different frame sampling rates on model performance. We found that sampling 10 frames for each second in the clip and constructing the mesh sequence from the corresponding frames is optimal for performance. We observed performance drops across all metrics at frame sampling settings that were lower and higher than this optimal setting [Table 4].

DISCUSSION

This section describes overarching interpretations of the results surrounding our HMR framework, surgical behavior analysis, and action recognition model. We also discussed the limitations and practical implications of our study.

Experimental interpretation

Our results provided evidence that an automated, human mesh-centered approach to OR video understanding can produce meaningful insights into surgical behaviors and short-duration OR actions. Notably, we showed that in addition to information on subject behavior that can be analyzed from meshes in single frames, such as visual attention, subject positioning, and joint pose, we can make more nuanced inferences on actions that persist across short durations. These capabilities are significant in the OR, as subtle actions and behaviors can be important predictors of team performance, operation trajectory, and patient outcomes.

Our experiments on surgical behavior analysis underscored the utility of granular mesh embeddings in representing individual behavior. Comparisons of positional heatmaps provided an interpretable way to understand how subjects are positionally distributed in the OR. Substantial differences in the positional and

Table 4. Performance of our multi-class classification model under different mesh sequence lengths used to model each 5-second action clip

FPS	Recall↑	Precision↑	F1↑	AUPRC↑
30	0.77	0.76	0.77	0.76
25	0.71	0.71	0.71	0.77
20	0.79	0.76	0.76	0.77
15	0.78	0.78	0.78	0.76
10	**0.80**	**0.79**	**0.79**	**0.82**
5	0.77	0.77	0.77	0.81

Videos are initially captured at 30 FPS. Embeddings are modeled as 3D joint positions. Bolding indicates a top score. AUPRC: The area under the precision-recall curve; FPS: frames-per-second.

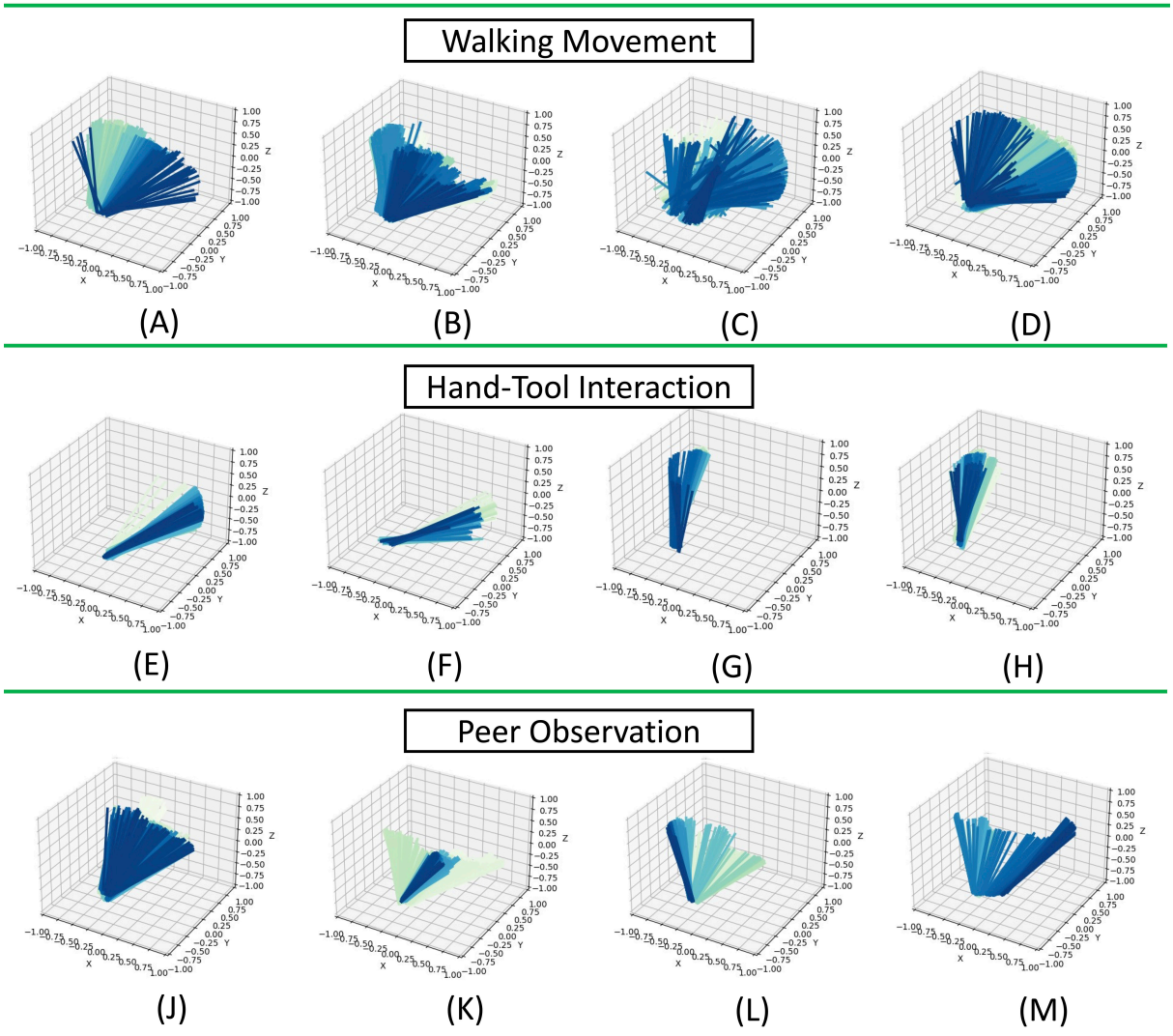

Figure 5. Comparisons of visual field-of-view composites from tracklets engaging in walking movements (A-D), hand-tool interactions (E-H), and observation of peers (J-M), where actions are mutually exclusive for each respective tracklet. The visual attention profiles for tracklets engaging in walking movements (A-D) are substantially more dispersed relative to tracklets engaging in hand-tool interactions (E-H) and moderately interspersed relative to tracklets engaging in observation of peer activities (J-M).

movement patterns between subjects in transition and focusing on a stationary task can be applied to identify transition periods in an OR procedure and instances of supply retrieval. Furthermore, visual attention profiles can provide a broad assessment of one's focus on a stationary task and be used to disseminate between surgical tasks that require different levels of visual attention.

These qualitative behavioral differences served as inspiration for comprehensive ablation studies on recovering surgical actions from mesh sequences. Overall, these studies provided critical insights for applying mesh-level features to downstream surgical prediction tasks. In comparing model performance when training with different mesh embedding compositions, we found that constructing mesh embeddings from 3D joint positions resulted in improved performance over 3D joint pose compositions. One possible explanation for this is that unlike joint poses, joint positions implicitly capture poses while carrying information about a subject's position in the overall scene. The observed performance difference suggests that scene positioning is important for telling apart surgical actions and that poses can be learned by our action recognition model from joint positions.

Leveraging this finding, we studied the impact of various joint categories on model performance and observed that the inclusion of joints from the "pelvic", "arm", and "cranial" joint categories was optimal. This quantitative result was consistent with previous qualitative observations, underscoring the differences in attention, movement, and positional patterns between human subjects performing different surgical actions. Interestingly, the further inclusion of joints in the "thorax", "spine", and "leg" categories resulted in successive performance drops. One possible explanation for this trend is that estimations of joints from these categories may be more imprecise due to higher tendencies for occlusion by adjacent equipment, specifically in joints of the "spine" and "leg" categories. We observed this phenomenon in a recovered human mesh in Figure 2, row 2, which erroneously modeled a standing subject in a sitting position. Similar to the effects of occlusion, each subject in our videos displayed a homogenous appearance due to their surgical attire, which may have affected the precision of pelvic and spine joint estimations. These challenges have been observed less frequently in previous HMR studies dealing with natural imagery due to common distinctive features between the upper and lower body attire of human subjects in natural settings[17,19]. Future work should explore methods to mitigate these errors and assess the uncertainty of joint predictions in surgical scenes.

Follow-up investigations into individual joints in the "arm" and "cranial" joint categories provided empirical evidence on the importance of individual joints that are closely tied to arm movements and visual attention for disseminating surgical actions. Specifically, we observed considerable, isolated improvements to model performance over ablated baselines when testing the separate inclusion of (1) pelvic joints, (2) anchoring joints for visual field computation (Section "Surgical behavior analysis"), such as the head and ear joints, and (3) arm joints, such as the wrist and elbow joints. Due to the importance of modeling intricate hand movements to analyze surgical performance, we hope to perform future studies that recover finger joints to discriminate between different hand movements. While a granular understanding of hand geometry was not central to our study of basic actions, our findings lay the groundwork for future studies on hand movements by providing evidence that mesh sequences can effectively encode physical actions. Furthermore, previous HMR studies have demonstrated the recovery of finger joints from in-the-wild scenes, supporting the feasibility of this research direction[28,29].

Altogether, our findings on recovering actions from mesh sequences demonstrated that we can consistently recover actions from sequences of human mesh features alone. Conventional approaches that directly use video frames to make predictions can be prone to overfitting, and our approach may circumvent this due to

its reliance on basic mesh-level features rather than image features that may vary widely among OR room appearances. To ensure that the risk of overfitting our models was accurately assessed, we stratified our training, evaluation, and test sets such that action clips comprising each data subset were derived from separate videos. Furthermore, we experimented with various regularization strategies, such as sequence frame sampling and reducing the number of modeled mesh features, to improve model robustness within the action recognition task. Our models performed comparably on our test set relative to our training set, providing strong evidence of their ability to generalize to new domains.

Limitations

There are important limitations of our work, one of which is that we focused on simulated surgery videos. These videos are similar to those of real endovascular procedures, since the OR room layout is identical, the same procedural steps are simulated, and similar equipment is used. However, the simulated videos featured fewer people and not all standard protective accessories, such as operating gowns, were used. While these simulated videos strongly resembled those of real procedures, we plan to study any challenges that may arise from applying our methods to videos of real procedures in future work.

An additional limitation of our study is that our simulated videos featured relatively similar room layouts and human appearances. To assess the capacity for our approach in generalizing to new surgical settings, we hope to test our approach across a larger video dataset, comprising full-duration endovascular procedures that capture a wide range of OR layouts and human appearances. In these future experiments, we plan to adhere to the ethical and legal guidelines outlined by Doyen *et al.* on performing continuous video recordings of the OR[30]. Ensuring patient privacy and practicing data stewardship are critical considerations in the safe integration of computer vision approaches with surgical video analysis; hence, these criteria are important for future HMR studies analyzing OR videos with real patients.

One final limitation of our study is that we focused on a small subset of short-duration surgical actions performed by individuals. While these actions are common across OR procedures and provided a conceptual foundation for our study, it is important for future efforts to focus on a larger range of surgical actions over longer time frames. Expanding the range of identifiable surgical actions would aid in the automatic reconstruction of procedure timelines and the identification of critical events, which are active areas of surgical research[31-34]. Specifically, sequences of human actions preceding critical milestones could be automatically parsed, and identified surgical events could be viewed collectively to attain a full picture of the procedure timeline. Computer vision provides a natural way to streamline this analysis in a scalable manner[31,35], and our study is a proof-of-concept for how this may be achieved with HMR. In addition to expanding the temporal dimension of our work, we also hope to investigate the extrapolation of subject-level behavior to an understanding of team dynamics and interpersonal communication, which are crucial hallmarks of success in the OR[6].

Clinical relevance

Our work has important implications for surgical video analysis that seeks to improve OR efficiency and patient outcomes. Previous studies have found that environmental distractions (i.e., auditory and visual) and workflow inefficiencies in the OR can have adverse effects on team performance, resulting in unfavorable patient outcomes[5,22,36]. This observation has been the basis of several observational studies that have focused on uncovering the root causes of OR inefficiencies from the lens of human behavior. Lynch *et al.*, for example, performed a manual video review of 28 surgical cases to monitor OR foot traffic and associated infection risk[2]. Hazlehurst *et al.* performed an ethnographic study of audiovisual data from 20 open-heart surgical cases to better understand team interactions in the OR[3]. Harders *et al.* examined perioperative flow patterns within 20 ORs during a three-month period to design interventions for reducing

nonoperative time[4]. These studies collectively highlight the notion that investigations on improving OR efficiency heavily rely on manual observation as the substrate for their analysis. Hence, the most significant barrier to improving surgery along these dimensions is the ability to analyze video data in a timely manner. Our HMR-based approach for surgical activity recognition serves as a foundation for exponentially scaling our ability to understand and improve the performance of both individuals and teams in the operating room.

CONCLUSION

In this paper, we presented a unified approach to systematically analyze the behavior and actions of individuals from OR videos using a human mesh-centered approach. Leveraging a novel, ensemble method for human detection, tracking, and mesh recovery, we demonstrated that substantial quantitative differences between surgical actions can emerge in the form of visual attention, movement patterns, and positional occupancy. We further showed that sequences of mesh embeddings formed from 3D joint positions can be used to train downstream machine learning models for surgical action recognition, paving the way for important downstream surgical tasks that rely on a rich understanding of human behavior. Overall, our work presents opportunities for video review programs to study human behavior in the OR in a systematic and scalable way. To our knowledge, we are the first study to have investigated the development of HMR-based approaches to analyze OR videos.

DECLARATIONS

Author contributions

Conceptualization, investigation, methodology, software, validation, visualization, writing - original draft, writing - review and editing: Liu B
Conceptualization, data curation, writing - review and editing: Soenens G
Conceptualization, writing - review: Villarreal J
Conceptualization, writing - review, supervision: Jopling J, Yeung-Levy S, Rau A
Conceptualization, data curation, writing - review, supervision: Van Herzeele I

Availability of data and materials

Code will be made available upon request.

Financial support and sponsorship

This work was supported by Wellcome Leap SAVE (No. 63447087-287892). Soenens G was supported by a PhD Fellowship (No. 11A5721-3N), and Van Herzeele I was supported by a Senior Clinical Fellowship (No. 802314-24N), both provided by the Fund for Scientific Research - Flanders, Belgium.

Conflicts of interest

All authors declare that there are no conflicts of interest.

Ethical approval and consent to participate

All visible subjects used in the simulated videos provided written informed consent to being filmed and agreed to the data's use in scientific research. As our research does not deal with any patient data or broadly, PHI, we did not obtain an institutional review board (IRB) for this study.

Consent for publication

All visible subjects provided written formal consent for publication.

REFERENCES

1. Dumas RP, Vella MA, Hatchimonji JS, Ma L, Maher Z, Holena DN. Trauma video review utilization: a survey of practice in the United States. *Am J Surg* 2020;219:49-53. DOI PubMed PMC
2. Lynch RJ, Englesbe MJ, Sturm L, et al. Measurement of foot traffic in the operating room: implications for infection control. *Am J Med Qual* 2009;24:45-52. DOI PubMed
3. Hazlehurst B, McMullen CK, Gorman PN. Distributed cognition in the heart room: how situation awareness arises from coordinated communications during cardiac surgery. *J Biomed Inform* 2007;40:539-51. DOI PubMed
4. Harders M, Malangoni MA, Weight S, Sidhu T. Improving operating room efficiency through process redesign. *Surgery* 2006;140:509-14. DOI PubMed
5. Palmer G 2nd, Abernathy JH 3rd, Swinton G, et al. Realizing improved patient care through human-centered operating room design: a human factors methodology for observing flow disruptions in the cardiothoracic operating room. *Anesthesiology* 2013;119:1066-77. DOI PubMed
6. Catchpole K, Mishra A, Handa A, McCulloch P. Teamwork and error in the operating room: analysis of skills and roles. *Ann Surg* 2008;247:699-706. DOI PubMed
7. Mottaghi A, Sharghi A, Yeung S, Mohareri O. Adaptation of surgical activity recognition models across operating rooms. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. Medical image computing and computer assisted intervention - MICCAI 2022. Cham: Springer; 2022. pp. 530-40. DOI
8. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision - ECCV 2016. Cham: Springer; 2016. pp. 561-78. DOI
9. Li H, Zech J, Hong D, Ghamisi P, Schultz M, Zipf A. Leveraging OpenStreetMap and multimodal remote sensing data with joint deep learning for wastewater treatment plants detection. *Int J Appl Earth Obs Geoinf* 2022;110:102804. DOI PubMed PMC
10. Yuan Y, Iqbal U, Molchanov P, Kitani K, Kautz J. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. pp. 11038-49. Available from: https://openaccess.thecvf.com/content/CVPR2022/html/Yuan_GLAMR_Global_Occlusion-Aware_Human_Mesh_Recovery_With_Dynamic_Cameras_CVPR_2022_paper.html. [Last accessed on 21 Jun 2024].
11. Kocabas M, Athanasiou N, Black MJ. Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020. pp. 5253-63. Available from: https://openaccess.thecvf.com/content_CVPR_2020/html/Kocabas_VIBE_Video_Inference_for_Human_Body_Pose_and_Shape_Estimation_CVPR_2020_paper.html. [Last accessed on 21 Jun 2024].
12. Tian Y, Zhang H, Liu Y, Wang L. Recovering 3D human mesh from monocular images: a survey. *IEEE Trans Pattern Anal Mach Intell* 2023;45:15406-25. DOI
13. Shao S, Zhao Z, Li B, et al. CrowdHuman: a benchmark for detecting human in a crowd. arXiv. [Preprint.] Apr 30, 2018 [accessed 2024 Jun 21]. Available from: https://arxiv.org/abs/1805.00123.
14. Weng SK, Kuo CM, Tu SK. Video object tracking using adaptive Kalman filter. *J Vis Commun Image Represent* 2006;17:1190-208. DOI
15. Li Z, Liu J, Zhang Z, Xu S, Yan Y. CLIFF: carrying location information in full frames into human pose and shape estimation. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision - ECCV 2022: 17th European Conference; 2022 Oct 23-27; Tel Aviv, Israel. Cham: Springer; 2022. pp. 590-606. DOI
16. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K. 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). 2019. pp. 2252-61. Available from: https://openaccess.thecvf.com/content_ICCV_2019/html/Kolotouros_Learning_to_Reconstruct_3D_Human_Pose_and_Shape_via_Model-Fitting_ICCV_2019_paper.html. [Last accessed on 21 Jun 2024].
17. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. In: Computer Vision - ECCV 2014. Cham: Springer; 2014. pp. 740-55. DOI
18. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans Pattern Anal Mach Intell* 2014;36:1325-39. DOI PubMed
19. von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-moll G. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision - ECCV 2018. Cham: Springer; 2018. pp. 614-31. DOI
20. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. pp. 779-88. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html. [Last accessed on 21 Jun 2024].
21. Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019. pp. 10975-85. Available from: https://

openaccess.thecvf.com/content_CVPR_2019/html/Pavlakos_Expressive_Body_Capture_3D_Hands_Face_and_Body_From_a_CVPR_2019_paper.html. [Last accessed on 21 Jun 2024].

22. Mentis HM, Chellali A, Manser K, Cao CG, Schwaitzberg SD. A systematic review of the effect of distraction on surgeon performance: directions for operating room policy and surgical training. *Surg Endosc* 2016;30:1713-24. DOI PubMed PMC
23. Tolstikhin IO, Houlsby N, Kolesnikov A, et al. MLP-Mixer: an all-MLP architecture for vision. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Available from: https://proceedings.neurips.cc/paper/2021/hash/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Abstract.html. [Last accessed on 21 Jun 2024].
24. Choe J, Park C, Rameau F, Park J, Kweon IS. PointMixer: MLP-mixer for point cloud understanding. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision - ECCV 2022. Cham: Springer; 2022. pp. 620-40. DOI
25. Ekambaram V, Jati A, Nguyen N, Sinthong P, Kalagnanam J. TSMixer: lightweight MLP-mixer model for multivariate time series forecasting. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM; 2023. pp. 459-469. DOI
26. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv. [Preprint.] Dec 22, 2014 [accessed 2024 Jun 21]. Available from: https://arxiv.org/abs/1412.6980.
27. Mehta D, Rhodin H, Casas D, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision. arXiv. [Preprint.] Nov 29, 2016 [accessed 2024 Jun 21]. Available from: https://arxiv.org/abs/1611.09813.
28. Moon G, Choi H, Lee KM. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. pp. 2308-17. Available from: https://openaccess.thecvf.com/content/CVPR2022W/ABAW/html/Moon_Accurate_3D_Hand_Pose_Estimation_for_Whole-Body_3D_Human_Mesh_CVPRW_2022_paper.html. [Last accessed on 21 Jun 2024].
29. Zhang X, Li Q, Mo H, Zhang W, Zheng W. End-to-end hand mesh recovery from a monocular rgb image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019. pp. 2354-64. Available from: https://openaccess.thecvf.com/content_ICCV_2019/html/Zhang_End-to-End_Hand_Mesh_Recovery_From_a_Monocular_RGB_Image_ICCV_2019_paper.html. [Last accessed on 21 Jun 2024].
30. Doyen B, Gordon L, Soenens G, et al. Introduction of a surgical Black Box system in a hybrid angiosuite: challenges and opportunities. *Phys Med* 2020;76:77-84. DOI PubMed
31. Garrow CR, Kowalewski KF, Li L, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg* 2021;273:684-93. DOI PubMed
32. Hardie JA, Hunn D, Mitchell TE, Brennan PA. Patient, Procedure, People (PPP): recognising and responding to intraoperative critical events. *Ann R Coll Surg Engl* 2022;104:409-13. DOI PubMed PMC
33. Fasting S, Gisvold SE. Serious intraoperative problems - a five-year review of 83,844 anesthetics. *Can J Anaesth* 2002;49:545-53. DOI PubMed
34. Yu X, Xiao H, Wang R, Huang Y. Prediction of massive blood loss in scoliosis surgery from preoperative variables. *Spine* 2013;38:350-5. DOI PubMed
35. Chadebecq F, Vasconcelos F, Mazomenos E, Stoyanov D. Computer vision in the surgical operating room. *Visc Med* 2020;36:456-62. DOI PubMed PMC
36. Nasri BN, Mitchell JD, Jackson C, Nakamoto K, Guglielmi C, Jones DB. Distractions in the operating room: a survey of the healthcare team. *Surg Endosc* 2023;37:2316-25. DOI PubMed PMC