Benchmarking 2D human pose estimators and trackers for workflow analysis in the cardiac catheterization laboratory

Butler, Rick M.; Frassini, Emanuele; Vijfvinkel, Teddy S.; van Riel, Sjors; Bachvarov, Chavdar; Constandse, Jan; van der Elst, Maarten; van den Dobbelsteen, John J.; Hendriks, Benno H.W.

DOI
10.1016/j.medengphy.2025.104289

Publication date
2025

Document Version
Final published version

Published in
Medical Engineering and Physics

Citation (APA)
Butler, R. M., Frassini, E., Vijfvinkel, T. S., van Riel, S., Bachvarov, C., Constandse, J., van der Elst, M., van den Dobbelsteen, J. J., & Hendriks, B. H. W. (2025). Benchmarking 2D human pose estimators and trackers for workflow analysis in the cardiac catheterization laboratory. *Medical Engineering and Physics*, *136*, Article 104289. https://doi.org/10.1016/j.medengphy.2025.104289

Medical Engineering and Physics 136 (2025) 104289

Contents lists available at ScienceDirect

Medical Engineering and Physics

journal homepage: www.elsevier.com/locate/medengphy

Paper

Benchmarking 2D human pose estimators and trackers for workflow analysis in the cardiac catheterization laboratory

Rick M. Butler [a,*], Emanuele Frassini [a], Teddy S. Vijfvinkel [a], Sjors van Riel [c], Chavdar Bachvarov [c], Jan Constandse [b], Maarten van der Elst [a,b], John J. van den Dobbelsteen [a], Benno H.W. Hendriks [a,c]

[a] Delft University of Technology, Delft, the Netherlands
[b] Reinier de Graaf Gasthuis, Delft, the Netherlands
[c] Philips Healthcare, Best, the Netherlands

ARTICLE INFO

Keywords:
Computer vision
Deep learning
2D pose estimation
2D pose tracking
Cardiac catheterization laboratory
Workflow analysis

ABSTRACT

Workflow insights can improve efficiency and safety in the Cardiac Catheterization Laboratory (Cath Lab). As manual analysis is labor-intensive, we aim for automation through camera monitoring. Literature shows that human poses are indicative of activities and therefore workflow. As a first exploration, we evaluate how marker-less multi-human pose estimators perform in the Cath Lab. We annotated poses in 2040 frames from ten multi-view coronary angiogram (CAG) recordings. Pose estimators AlphaPose, OpenPifPaf and OpenPose were run on the footage. Detection and tracking were evaluated separately for the Head, Arms, and Legs with Average Precision (AP), head-guided Percentage of Correct Keypoints (PCKh), Association Accuracy (AA), and Higher-Order Tracking Accuracy (HOTA). We give qualitative examples of results for situations common in the Cath Lab, with reflections in the monitor or occlusion of personnel. AlphaPose performed best on most mean Full-pose metrics with an AP from 0.56 to 0.82, AA from 0.55 to 0.71, and HOTA from 0.58 to 0.73. On PCKh OpenPifPaf scored highest, from 0.53 to 0.64. Arms, Legs, and the Head were detected best in that order, from the views which see the least occlusion. During tracking in the Cath Lab, AlphaPose tended to swap identities and OpenPifPaf merged different individuals. Results suggest that AlphaPose yields the most accurate confidence scores and limbs, and OpenPifPaf more accurate keypoint locations in the Cath Lab. Occlusions and reflection complicate pose tracking. The AP of up to 0.82 suggests that AlphaPose is a suitable pose detector for workflow analysis in the Cath Lab, whereas its HOTA of up to 0.73 here calls for another tracking solution.

1. Introduction

The field of workflow analysis is gaining traction in medical environments [1–4]. During surgery, insight into workflow is necessary in order to optimize procedures. Example use-cases are improved procedure efficiency, safety, and training.

Manual workflow analysis is a laborious task that requires experts to carry out. Automation enables cost-effective, large-scale deployment and additional use-cases like real-time feedback or support [5–8]. Personnel activities, which can be found from human pose tracklets [9–11], are descriptive of workflow.

Multi-object keypoint detection—also called pose estimation—aims to localize predefined objects and their keypoints in an image. Fig. 1 shows keypoints and edges ('limbs') for the 'Human' class as defined in [12]. Pose estimators output a continuous pixel (px) location and confidence score per detected keypoint. Modern works often take one of two approaches:

Top-down: Detect object bounding boxes [13] and estimate a pose in each of them [14–16].
Bottom-up: Detect keypoints and assemble them into objects [17–19].

In this work we refer to human keypoints using the abbreviations and groupings from Fig. 1a, where a leading '*l*' or '*r*' denotes 'left' or 'right'.

The temporal element in videos gives need to multi-object tracking: the assignment of the same identity (ID) to the same object in different

* Corresponding author.
E-mail address: r.m.butler@tudelft.nl (R.M. Butler).

https://doi.org/10.1016/j.medengphy.2025.104289
Received 5 August 2024; Received in revised form 23 December 2024; Accepted 7 January 2025

Available online 10 January 2025

Nomenclature

Algorithms

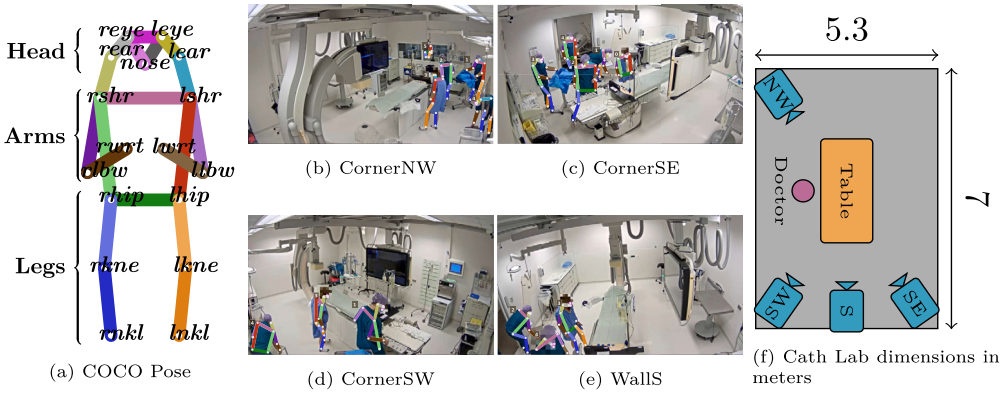
T	Tracking
AlphaP	AlphaPose
OpenPP	OpenPifPaf
OpenP	OpenPose

Metrics

AP	Average Precision
AA	Association Accuracy
DA	Detection Accuracy
FN	False Negative
FP	False Positive
HOTA	Higher-Order Tracking Accuracy
IoU	Intersection over Union
OKS	Object Keypoint Similarity
τ_{OKS}	OKS threshold
PCKh	Head-guided Percentage of Correct Keypoints
TP	True Positive

Units

px	Pixel
pp	Percentage Point



(a) COCO Pose (b) CornerNW (c) CornerSE (d) CornerSW (e) WallS (f) Cath Lab dimensions in meters

Fig. 1. (a) COCO pose [12] facing the reader, (b)-(e) Camera viewpoints with annotated poses, (f) Map of the Cath Lab with measurements in meters, table and cameras not to scale.

video frames. This can be done causally [20], non-causally [21], jointly with detection [18], or separately after detection [14]. We denote algorithms with tracking capabilities with a superscript 'T'.

Annotations are required to train or test keypoint detectors and trackers. Human annotators label 'ground-truth' poses with their video-specific ID, presence, and location. Algorithms learn to mimic the annotation process and are tested against ground-truths.

Medical environments present challenges like significant occlusion between personnel and objects, and appearance similarities due to sterile clothing. General-purpose datasets like [12] are not representative of such settings. Evaluating pose estimator performance requires recordings of real procedures, which are scarce due to privacy regulations [22]. MVOR [23] is a public dataset with recordings from the hybrid operating room (OR). It was recorded in four days during different procedures in a university hospital. To capture workflow information, however, more data and procedure uniformity are needed.

Human pose estimation in ORs was investigated in [24–26]. Reference [26] tests a state-of-the-art 2D pose estimator on a single metric, and focuses on the step to 3D pose estimation. To our knowledge, optimality of the chosen 2D pose estimator in a medical setting has not been verified. References [27,28] investigate exoskeletal control through the tracking of local limb movements. Where their methods rely on measurement through wearables, we investigate measurement through camera monitoring. Reference [29] investigates the scalability of object detection to different Cath Labs, but does not consider pose detection.

The Cardiac Catheterization Laboratory (Cath Lab) is a specialized operating room (OR) where minimally invasive cardiovascular procedures take place. This work evaluates the performance of human pose estimators and trackers as a potential tool for workflow analysis in the Cath Lab. To this end, we record real coronary angiogram (CAG) procedures in a regional hospital from the four camera (Axis M1125) views shown in Fig. 1. The videos capture workflow before, during and after procedures. Poses are annotated in ten procedures, showing five different workflow phases. The Cath Lab presents unique challenges to computer vision like concealing clothing, occlusion, and reflections. To our knowledge, no video dataset of real Cath Lab procedures exists in literature at the time of this study. An estimator to analyze any future recordings can be selected in line with results from this work.

We test several pre-trained state-of-the-art 2D human pose estimators in the Cath Lab. Three algorithms were selected by the criteria that they i) can detect an arbitrary number of poses per image, and ii) provide implementation details in peer-reviewed work: AlphaPose [14] (AlphaP), OpenPifPaf [18] (OpenPP), and OpenPose [17] (OpenP). As AlphaP is a top-down estimator and OpenPP and OpenP are bottom-up, results should give an idea of which approach works best in the Cath Lab. AlphaP and OpenPP also provide causal tracking models AlphaP^T and OpenPP^T. We quantitatively measure detection- and tracking performance and support these metrics with qualitative observations.

The main contributions of this work are:

- We introduce a unique multi-view dataset of real CAG procedures in the Cath Lab with pose annotations.
- We evaluate the performance of several state-of-the-art 2D human pose estimators in the Cath Lab.
- We discuss—from a workflow perspective—pitfalls for pose estimation that are Cath Lab-specific.

Section 2 starts with a description of our dataset, included algorithms, and evaluated metrics. Section 3 lists the results and highlights trends and differences. Then, section 4 discusses and explains observed outcomes. We theorize what the results imply for our setting and identify an algorithm for use in future work. Finally, section 5 gives a summary.

2. Materials and methods

This section describes all components that make up our benchmark. Section 2.1 starts with a description of the dataset and its recording process. Section 2.2 provides a brief explanation of the used pose estimators. Section 2.3 concludes with our used metrics and other evaluations.

2.1. Video recordings

Four cameras (Axis M1125) were hung in the Cath Lab of the Reinier de Graaf Gasthuis, Delft, NL. With approval of a local medical ethics committee and the hospital board, and informed consent from the patients and staff, CAG procedures were recorded from the viewpoints in Fig. 1 and stored with a resolution of 1920 px × 1088 px and framerate of 25 frames per second. A cardiologist, scrub nurse, up to two lab assistants, and the patient were present during each procedure. We record and annotate ten procedures, where we ensure that each shows a different medical team for variability. CAGs follow a strict, consistent workflow with little to no variation. Because of this uniformity, the ten chosen procedures cover the typical cases. Local doctors helped select the procedures to include some rare deviations. For instance, there is a procedure during which the cardiologist had to move the monitor, one where ultrasound was needed to find the radial artery for endovascular access, and one where the staff struggled to reposition the lead shield.

2.1.1. Annotation

In each procedure, poses were annotated in 51 frames sampled uniformly over 30 seconds, from four synchronized viewpoints. This gives a total of 10 (procedures) × 51 (frames) × 4 (viewpoints) = 2040 annotated frames. The 30 seconds per procedure were hand-picked to show one of five unique workflow phases:

- The patient entering and lying down.
- Realization of endovascular access through the wrist.
- Use of ultrasound to detect the radial artery for endovascular access.
- X-Ray imaging.
- Closure of the entrywound.

Each phase was selected twice from different procedures. Poses were annotated in Computer Vision Annotation Tool (CVAT) [30] by two of the authors with a background in engineering, and their quality confirmed by a third who has been a practicing interventional cardiologist for over 13 years. We did not use the CVAT interpolation feature in order to preserve fine positioning, which we expect to be important for workflow analysis in the Cath Lab. One annotated example frame is shown per viewpoint in Fig. 1. Fully occluded individuals and keypoint reflections in e.g. the monitor were not labeled. People in the control room and hallway were included.

We define a person to be 'visible' on a frame if any of their keypoints can be seen directly in that frame without obstruction. To describe the dataset we label each frame by presence of situations that arise in the Cath Lab:

- Occluded fully: A person is inside the camera view but not visible.
- Occluding person: Segmentations of visible persons overlap.
- Occluding object: An object segmentation overlaps a visible person.
- Occluding sheet: The surgical sheet overlaps the visible patient.
- Occluding clothes: Sterile clothes conceal visible elbows, knees or hips.
- Occluding window: The control room window overlaps a visible person.
- Occluding view: A wall or frame boundary overlaps a visible person.
- Horizontal patient: The visible patient shows non-vertically in the view.
- Reflecting monitor: The monitor shows a reflected person.
- Reflecting window: The control room window shows a reflected person.

Some situations are viewpoint-specific, e.g., CornerSE sees no reflective surfaces and the patient is vertical from WallS even when lying down. The situations are labeled per frame, i.e., if a situation occurs multiple times in the same frame it is counted as a single instance. In addition, we record the number of visible people per frame using the same methodology. Finally, we count the total number of annotated keypoints per class where multiple can be counted per frame.

2.2. Pose estimation

AlphaP is implemented as a parallel pipeline which aims for high inference speeds. A fast object detector [31,32] detects Human bounding boxes, in each of which a Convolutional Neural Network (CNN) generates a heatmap per keypoint. At the maximum of this heatmap, the keypoint is placed. This per-bounding box processing makes AlphaP a top-down algorithm. Optionally a second CNN extracts features per bounding box for tracking and trajectory smoothing, where background noise is mitigated by masking with the detected pose. A low object detector confidence threshold avoids false negatives but yields redundant detections. Pose Non-Maximum Suppression removes resulting duplicate poses. A translation-invariant approximation of the loss function gradient is used during optimization. Additionally, heatmaps are normalized such that calculated confidences become invariant of keypoint scale.

OpenP has a CNN encode limb presence and orientation over the entire image into Part Affinity vector Fields (PAFs). A second CNN generates keypoint heatmaps and locations from these PAFs like AlphaP. Poses are assembled in bottom-up fashion with bipartite matching: each candidate limb is scored by integration over its PAF, and a set of limbs is selected to maximize the sum of scores.

OpenPP replaces heatmaps with Composite Intensity Fields which encode keypoint confidence, scale, and location offset. PAFs are replaced with Composite Association Fields (CAFs) which encode i) probability of limb presence and ii) endpoint scales and location offsets. Temporal CAFs model limbs between keypoints of the same class in adjacent frames for tracking. Poses are grown bottom-up by greedy matching from a high-confidence seed keypoint, guided by these intensity and association fields. Keypoint-level Non-Maximum Suppression removes duplicate poses. Redundant limbs are modeled for robustness against occlusion.

2.3. Experimental setup

The following sections describe the used pre-trained models, evaluation metrics, and other validation procedures.

2.3.1. Model settings

Each algorithm offers several pre-trained models, which can be split into i) a backbone which extracts image features and ii) a head which estimates poses and/or IDs. We test the models in Table 1 on our dataset without retraining, using an NVIDIA GeForce RTX 3090 GPU. The used model parameters are available publicly for AlphaP,[1] OpenPP[2] and OpenP.[3] For fair comparison we sample all outputs to the format from Fig. 1a. We define pose confidence as the mean of all its nonzero keypoint confidences.

[1] Available: https://github.com/MVIG-SJTU/AlphaPose/blob/master/docs/MODEL_ZOO.md.
[2] Available: https://openpifpaf.github.io/intro.html.
[3] Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/master/models.

Table 1
Tested pose estimators where a superscript 'T' denotes tracking capabilities.

Algorithm	Backbone	Head	Training dataset
AlphaP	YOLOv3-SPP [31,32]+ResNet152 [33]	FastPose (DUC) [14]	COCO [12]
AlphaP^T	YOLOv3-SPP [31,32]+ResNet152 [33]	FastPose (DUC)+Human-ReID [14]	COCO [12]
OpenPP	shufflenetv2k30 [18,34]	CifCaf [18]	COCO [12]
OpenPP^T	tshufflenetv2k30 [18,34]	TrackingPose [18]	COCO [12]
OpenP	OpenPose [17]	OpenPose [17]	COCO [12]+Human Foot [17]

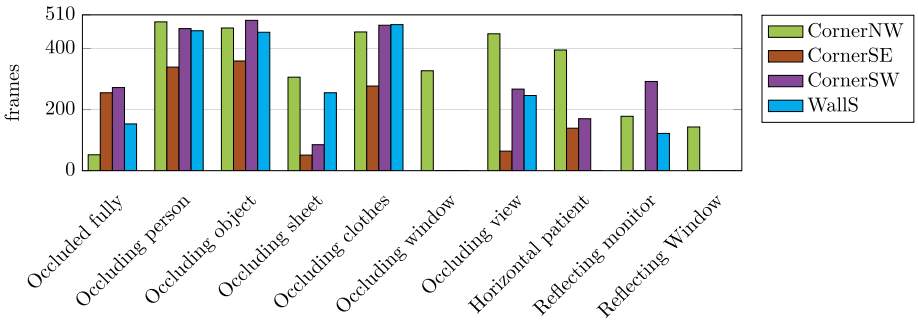


Fig. 2. Number of frames per situation from section 2.1.1 per viewpoint.

2.3.2. Quantitative metrics

Metrics measure detection- and tracking performance per viewpoint. They are calculated with True Positives (TPs), False Positives (FPs), and False Negatives (FNs), using only visible keypoints.

Average Precision (AP) [12,35] evaluates Full-pose detection. As it was designed for bounding boxes, we replace its use of Intersection over Union with Object Keypoint Similarity (OKS) as suggested in [12], where we estimate segmentation area with the tightest-fit pose bounding box. For a more detailed evaluation we calculate AP separately for three subposes: Head, Arms, and Legs, in addition to the Full pose. We calculate $AP^{\tau_{OKS}}$ at OKS thresholds $\tau_{OKS} = 0.5$ (low), $\tau_{OKS} = 0.75$ (high), and averaged from 0.5 to 0.95 with step size 0.05 $\tau_{OKS} = 0.5 : 0.95$ (ranged).

Head-guided Percentage of Correct Keypoints (PCKh) [36] evaluates detection per keypoint. We first use the Hungarian algorithm [37] to match annotated and estimated poses by OKS where—as opposed to AP—we do not threshold confidence. PCKh is evaluated per match. Since the COCO pose has no headbone, we threshold TPs with 0.5 times the longest annotated *shr-ear* distance instead, and only use poses with such an annotated limb. With the obtained per-keypoint TPs, FPs and FNs we calculate

$$\text{PCKh} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{1}$$

We evaluate tracking for each viewpoint and subpose with Association Accuracy (AA) [38], and replace its use of Localization Similarity with OKS as was done for AP. Finally, Higher-Order Tracking Accuracy (HOTA) [38] summarizes detection and tracking performance in a single metric. Although HOTA is an aggregation of AA and Detection Accuracy, we do not evaluate the latter, as its purpose is similar to that of the more commonly used AP.

We show metrics evaluated per individual video, each of which shows one of five workflow phases from different procedures. Error bars show two standard deviations around the mean metric. If one situation yields a better score than others, we say this situation is 'preferred'. Unless explicitly stated otherwise, discussed results are mean Full-pose scores for ranged τ_{OKS}.

2.3.3. Statistical significance

We evaluate the significance of performance differences between each pair of algorithms with a two-sample Hotelling's T-Squared [39]. AP and PCKh are used as dependent variables, as AA and HOTA can not be calculated for every tested algorithm. Specifically, we include AP for each separate subpose with ranged τ_{OKS}, and PCKh per keypoint for a total of 3 (subposes) + 17 (keypoints) = 20 parameters per sample. Each single-view video represents a sample for a total of 40 samples. We consider p-values of 0.05 or below to show statistical significance.

We repeat the same analysis to compare AlphaP^T and OpenPP^T on AA, where again the three ranged-τ_{OKS} subposes are used as separate input variables. Instead of repeating again with HOTA, we test on AA jointly with AP and/or PCKh.

2.3.4. Qualitative analysis

To get insight into problems specific to our setting, results are manually evaluated. Specific example situations are selected by the authors to demonstrate strengths and weaknesses of each algorithm. Results are shown with detected poses, confidence scores and IDs. We show all detections regardless of their confidence.

3. Results

This section shares the results obtained from the experiments described in section 2.3. Section 3.1 begins with an analysis of the dataset. Sections 3.2 to 3.5 report performance on various metrics. Statistical significance of the differences between algorithms is investigated in section 3.6. Finally, section 3.7 shows some qualitative examples.

3.1. Dataset composition

Figs. 2 to 4 show a description of the dataset as described in section 2.1.1. 1749 frames (85.7 % of the dataset) contain occlusion between persons, 1771 (86.8 %) occluding objects, and 1685 (82.6 %) occluding clothes. CornerNW, CornerSE, CornerSW and WallS saw 3257, 1484, 2519 and 2165 frame situations respectively, where counts exceed the dataset size due to single frames showing multiple situations. Most full occlusions and monitor reflections occur from CornerSW. Window occlusions- and reflections occur only from CornerNW. This viewpoint sees five persons on most frames and is the only viewpoint to ever see six. CornerSE usually sees four people and CornerSW and WallS three.

CornerNW, CornerSE, CornerSW and WallS respectively see a total of 20 404, 15 317, 16 012 and 16 518 keypoints. CornerNW sees the highest counts per subpose and class except for the *rear*, *rlbw* and *rhip*, which are seen more often from CornerSW, and the wrists which are seen more from WallS. CornerNW sees the highest mean count per class of 1200.2, paired with the highest standard deviation of 320.7. The lowest total and average counts are seen by CornerSE, although CornerSW sees lower counts per class more often.

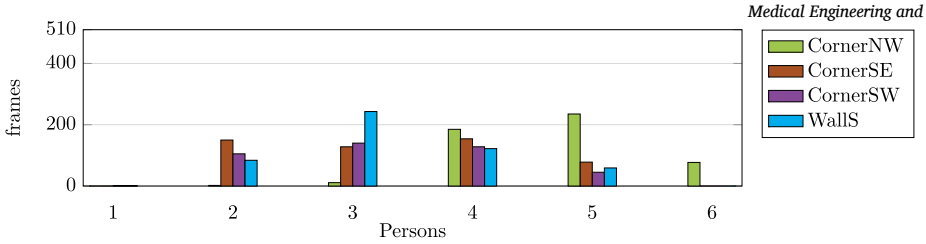


Fig. 3. Number of frames per person count per viewpoint.

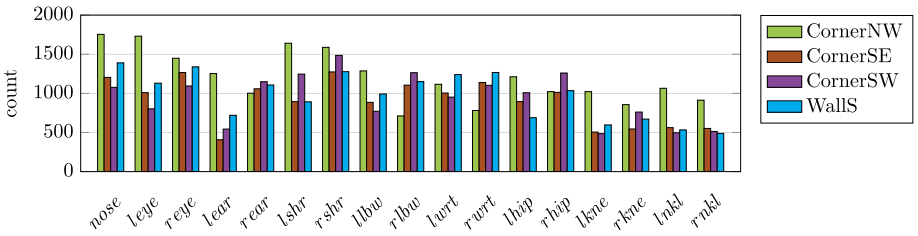


Fig. 4. Number of appearances per keypoint class from Fig. 1a per viewpoint.

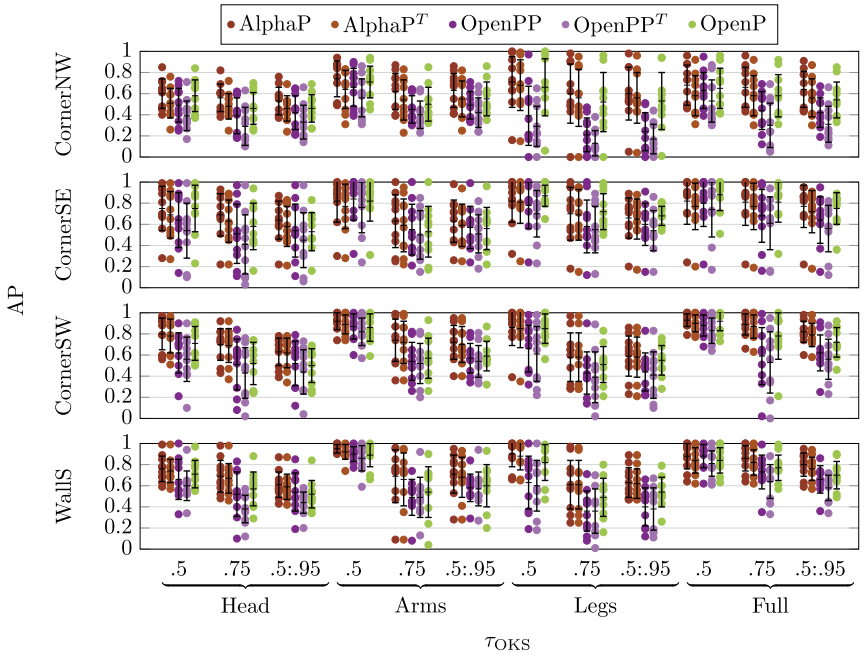


Fig. 5. Average Precision per individual video where error bars show two standard deviations around mean results.

3.2. Average precision

Fig. 5 shows AP. Here, AlphaP yields the highest Full-pose mean scores of up to 0.82. Non-tracking algorithms perform up to 11 percentage points (pp) better than their tracking counterparts. Arms yield the best scores of up to 0.72, and Head the worst of up to 0.64. AlphaP(^{T}) prefers the CornerSW viewpoint, OpenPP(^{T}) WallS, and OpenP CornerSE. CornerSW is most often preferred with up to 3 pp over the second-choice viewpoint per individual algorithm. On the Head and Arms, AlphaP(^{T}) shows scoring drops of up to 10 pp and 26 pp between low and high τ_{OKS}, which is 22 pp and 37 pp for other algorithms. On the Legs, OpenPP(^{T}) shows the lowest scoring drop of up to 27 pp which is 34 pp for others. Results on the Arms show standard deviations of up to 20 pp. For the Head and Legs this is 26 pp and 27 pp respectively. OpenP shows standard deviations of up to 17 pp, AlphaP(^{T}) of 20 pp and OpenPP(^{T}) of 22 pp.

From CornerSE, one outlier procedure performs worse than the rest for all algorithms. Here the cardiologist and patient are mostly occluded by the monitor. Their few visible keypoints were not detected, or merged into a single pose.

3.3. Head-guided percentage of correct keypoints

PCKh in Fig. 6 shows that most keypoints prefer OpenPP or OpenPP^T except the *nose*, which prefers AlphaP instead. All algorithms prefer WallS most often, followed by CornerSE for AlphaP and OpenPP, and CornerSW for AlphaP^T and OpenP. AlphaP, OpenPP^T and OpenP prefer CornerNW least often, which is CornerSE for AlphaP^T and CornerSW for OpenPP. With scores of up to 0.57 and 0.87 the Legs and Head score the lowest and highest respectively. The *hips* are detected worst with a maximum score of 0.23. All subposes prefer WallS. All algorithms had the highest standard deviation on CornerSW. The lowest standard deviations for the Head are achieved on CornerSE and WallS, for the Arms on WallS, and for the Legs on CornerNW.

The outlier procedure from CornerSE at the end of section 3.2 shows the same poor performance on PCKh. From CornerSW, we see another

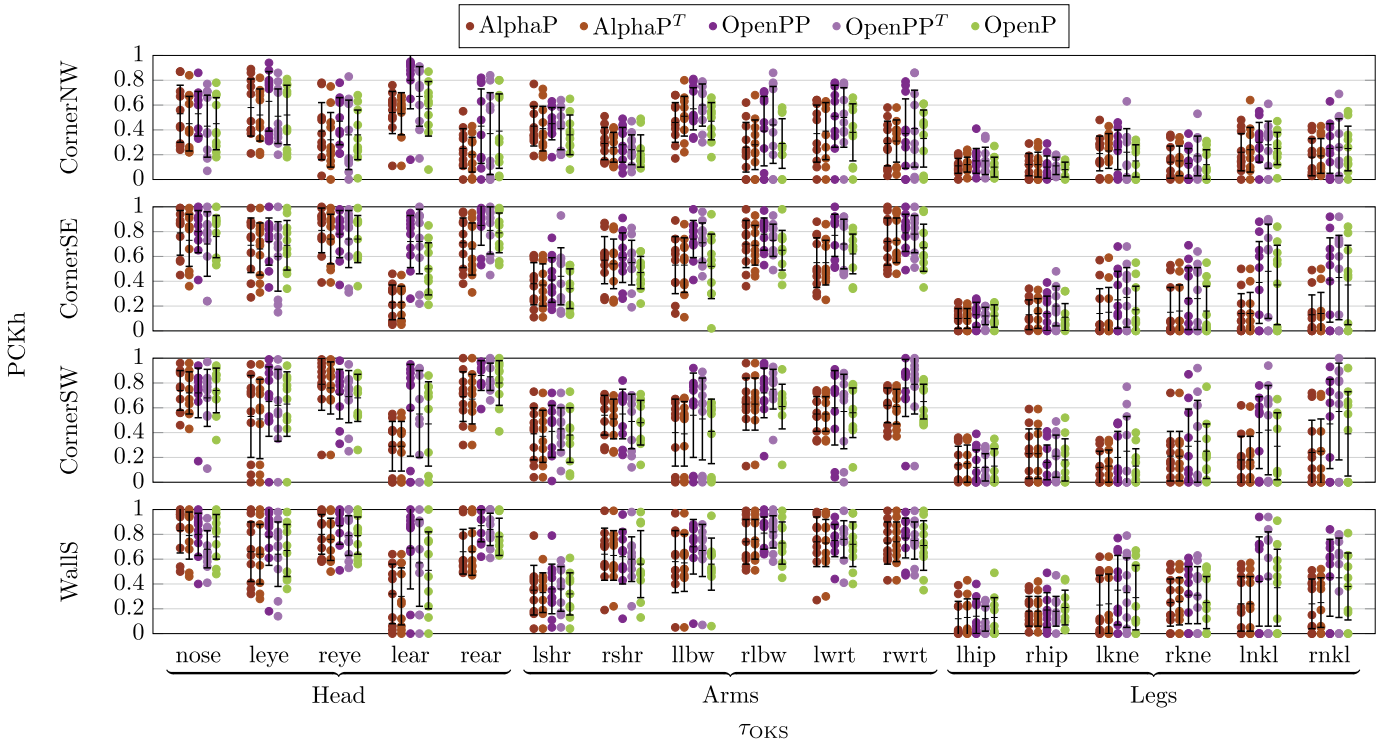


Fig. 6. Head-guided Percentage of Correct Keypoints over the entire dataset where error bars show two standard deviations around mean results.

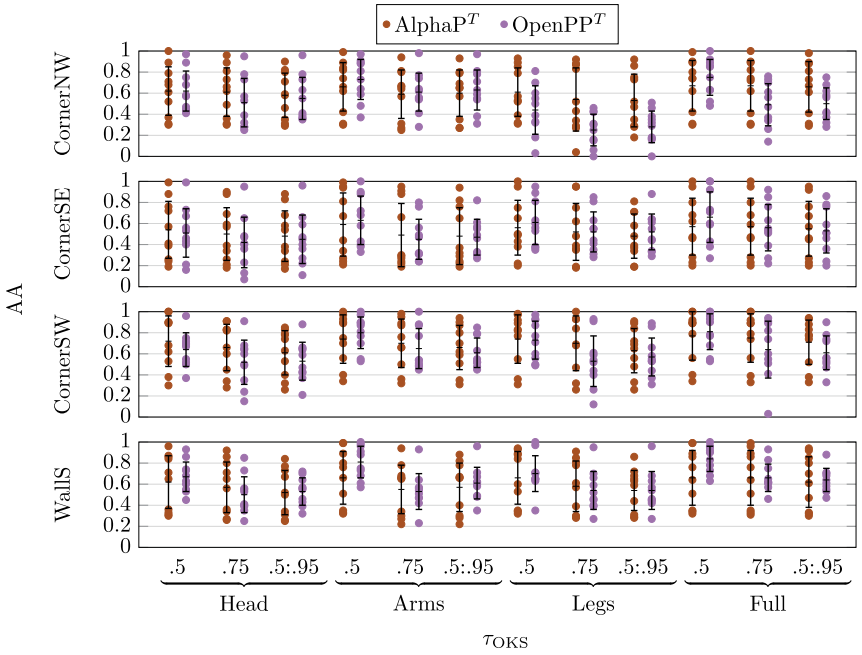


Fig. 7. Association Accuracy per viewpoint and subpose over the entire dataset where error bars show two standard deviations around mean results.

procedure scoring below the others. This video shows two people standing close together, dressed in loose medical aprons and facing away from the camera whilst the instrument table occludes their legs.

3.4. Association accuracy

Looking at tracking, Fig. 7 shows that AlphaP^T outperforms OpenPP^T on mean Full-pose AA from all viewpoints except WallS. Arms are tracked best in most situations, and Legs the worst. AlphaP^T shows little mean Full-pose scoring drop of up to 2 pp between low and high τ_{OKS}, which is 26 pp for OpenPP^T. On AA this drop is larger for OpenPP^T than for AlphaP^T for all subposes and viewpoints. However, OpenPP^T yields lower standard deviation than AlphaP^T for all subposes and the Full pose from all viewpoints.

3.5. Higher-order tracking accuracy

The integration of tracking and detection metrics with HOTA in Fig. 8 sees AlphaP^T outperform OpenPP^T everywhere except with low τ_{OKS} for some subposes and viewpoints. OpenPP^T still yields lower stan-

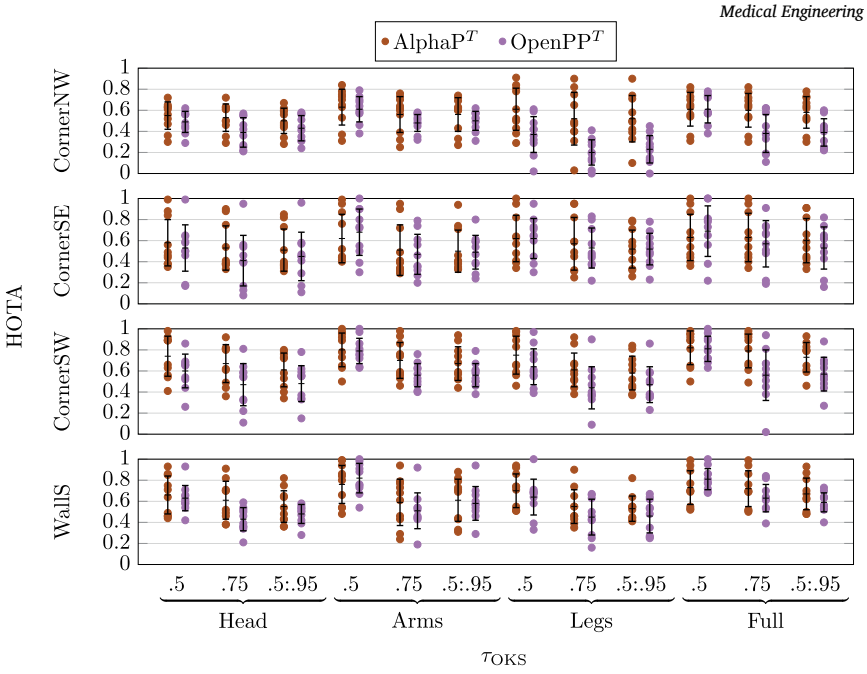


Fig. 8. Higher-Order Tracking Accuracy per viewpoint and subpose over the entire dataset where error bars show two standard deviations around mean results.

Table 2
p-value per algorithm pair from Hotelling's T-Squared with AP and PCKh as parameters.

	AlphaP^T	OpenPP	OpenPP^T	OpenP
AlphaP	0.9999	<0.0001	<0.0001	<0.0001
AlphaP^T		<0.0001	<0.0001	<0.0001
OpenPP			0.9103	0.0003
OpenPP^T				<0.0001

dard deviations except for the Head from CornerSE and CornerSW, Legs from CornerSW and WallS, and Full pose from CornerSW. The highest achieved Full-pose mean scores are 0.73 for AlphaP^T and 0.59 for OpenPP^T. Arms and CornerSW are preferred in most situations.

3.6. Hotelling's T-squared

Table 2 shows calculated p-values per pair of algorithms. The only non-significant differences occur when comparing tracking- and non-tracking versions of the same algorithm, in which case p-values approach 1.

When excluding AP from the test, conclusions remain the same except for a now statistically insignificant p-value between OpenPP and OpenP. Excluding PCKh instead gives insignificance between AlphaP^T and OpenP.

Testing on AA yields an insignificant difference between AlphaP^T and OpenPP^T. Adding AP and/or PCKh lowers the p-value back below our significance threshold.

3.7. Qualitative results

Fig. 9 shows example detections. In the first column people are standing close together. OpenPP^T merges the patient and cardiologist with 0.87 confidence. AlphaP(^{T}) mistakes the patient as part of the cardiologist at 0.72. The cardiologist and assistant who stands close are never merged. Only OpenPP and OpenP see the patient and merge no-one. OpenP detects most correct poses, but with the lowest confidence. All models except OpenPP^T are least confident about the cardiologist, who faces away from the camera.

The second column shows the cardiologist putting on an apron. AlphaP places a full pose with confidence 0.31 where only his Head is visible, and AlphaP^T sees nothing. OpenPP(^{T}) correctly detects the *shrs* and *hips* at 0.82, although *hip* placements seem off. OpenPP^T additionally sees a lower arm in the sleeve. Only OpenP detects the *nkls*. All algorithms detect the lab assistant where AlphaP, AlphaP^T and OpenP place the occluded *lnkl* wrongly with 0.84, 0.84 and 0.69 confidence.

In the third column the monitor reflects a lab assistant. All algorithms detect the reflection, where OpenPP is most confident at 0.83 and OpenP the least at 0.63. AlphaP, AlphaP^T and OpenP hallucinate two *knes* and/or *nkls*. These models detect the full cardiologist at 0.74, 0.74 and 0.60 confidence, where OpenPP(^{T}) detects all but his legs at 0.87 and 0.69. All models see the occluded assistant, where OpenPP(^{T}) is most confident at 0.95 and AlphaP(^{T}) the least at 0.62. Similarly to column two, AlphaP incorrectly detects a full pose around the head of the patient with 0.31 confidence. OpenPP correctly detects only their Head keypoints at 0.90.

The last column shows the instrument table with a sheet resembling clothing. AlphaP and OpenPP detect a pose here at 0.44 and 0.64 confidence. The occluding assistants are fully detected by AlphaP at 0.72 and 0.66 and merged by OpenPP at 0.77. AlphaP^T and OpenPP^T only detect the closest assistant at 0.72 and 0.90. OpenP detects both assistants partially at 0.73 and 0.57. It also sees Legs in the background bin with 0.22 certainty.

Fig. 10 shows tracking results from AlphaP^T and OpenPP^T. The first and third row show the cardiologist and assistant preparing, with a patient on the table. AlphaP^T detects the partial cardiologist at the bottom in 3 frames, which is only 1 for OpenPP^T. OpenPP^T detects the patient more consistently and with higher confidence. AlphaP^T moves ID 1 from the assistant to the cardiologist. OpenPP^T yields no such identity swaps.

In the remaining rows the cardiologist exits and re-enters the room, with the patient waiting in the hallway. After their return, OpenPP^T assigns the cardiologist a new ID whereas AlphaP^T recognizes them from before. The same happens for the patient after the cardiologist passes them in front. When watching frame by frame, AlphaP^T shows many identity swaps even with just one person visible.

4. Discussion

In this paper we introduced a dataset with footage from real CAG procedures in the Cath Lab, and provided benchmark results of several pose estimation- and tracking algorithms. Quantitative metrics were

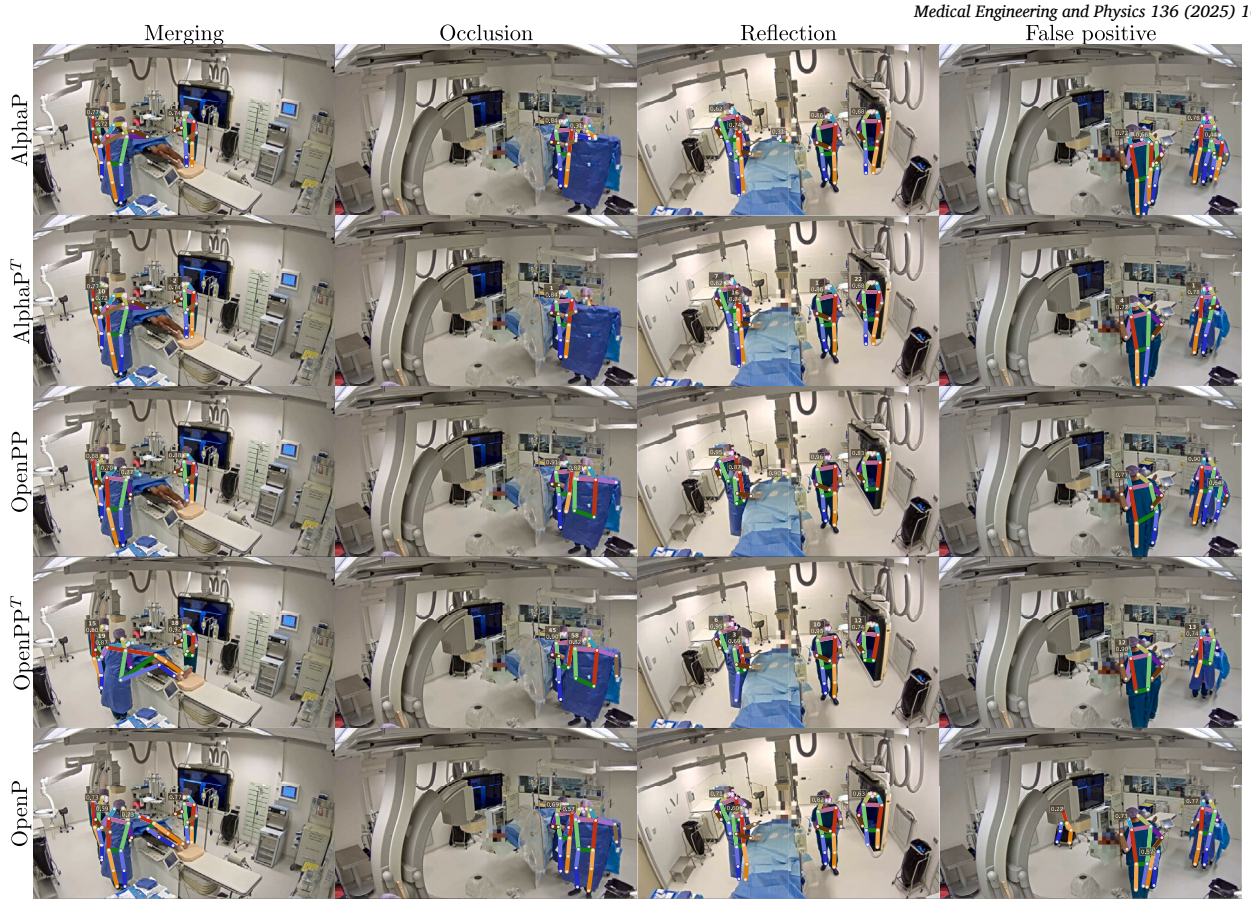


Fig. 9. Qualitative detections with confidences and tracking IDs.

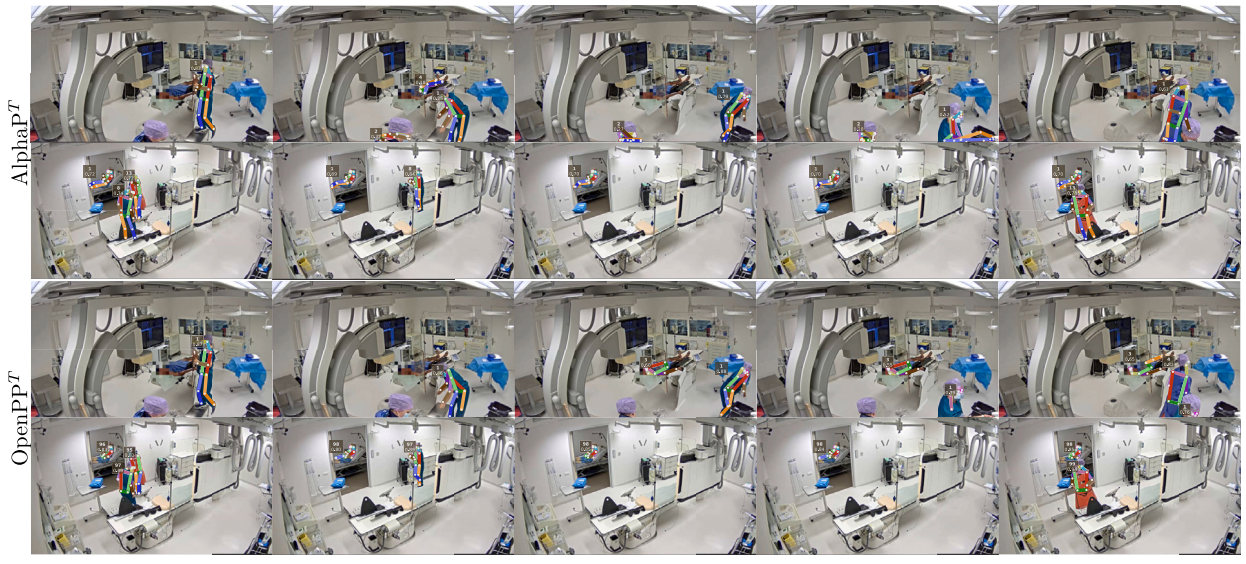


Fig. 10. Qualitative tracklets with confidences and tracking IDs.

evaluated on our annotated dataset per subpose and viewpoint, and qualitative observations were shown.

We observe that AlphaP(^{T}) produces the best AP, whereas OpenPP(^{T}) performs better on PCKh. As AP is calculated on (sub)poses and PCKh per keypoint, this suggests that OpenPP(^{T}) places keypoints more accurately and AlphaP(^{T}) connects them into poses better. This is in line with the top-down approach of AlphaP, which applies local restrictions on matchable keypoint pairs. Another explanation is that AlphaP(^{T}) could score poses more accurately in the Cath Lab, as AP considers confidence score. This could explain why metrics on AlphaP(^{T}) do not drop much between low and high τ_{OKS}, as accurate scoring might compensate cases of poor localization. Measured AP scores are higher in the Cath Lab than those reported on the MVOR dataset [23], although different recording methods render these results not directly comparable.

AP and PCKh show differences per viewpoint and keypoint. Legs especially are subject to occlusion in clinical settings. Head keypoints are

hidden behind masks and hairnets. Different viewpoints see different levels of such occlusion, and therefore show varying results. AlphaP(^{T}) tends to detect the Head better, possibly by imposing a prior through object detection. Qualitative results show a drawback of this approach, where priors encourage placement of full poses on partially visible humans or inanimate objects. We do observe that these incorrect keypoints receive low confidences, which is in line with the theory that AlphaP yields more accurate scores. Hence, in practice this drawback poses little issue if confidence is considered appropriately.

On AA, AlphaP^T scores better than OpenPP^T for our dataset. OpenPP^T tends to miss people in our setting or merge them; possibly due to the temporal limbs providing more matching paths to do so in close proximity. AlphaP^T still scored poorly with large variability between workflow phases. This could be due to its use of visual clues, which in combination with indistinguishable sterile clothing may have caused the many identity swaps. These issues could explain why the tracking models were outperformed by their non-tracking counterparts on AP, although this difference was only small and proved insignificant.

HOTA eases comparison by integrating detection and tracking performance. Here, AlphaP^T slightly outperforms OpenPP^T. When looking purely at this combined metric, using AlphaP^T from the CornerSW viewpoint seems to perform best in the Cath Lab.

Procedures are carried out with Arms, and Head orientation indicates where one is focusing. In our setting, Legs serve only to reposition oneself; something that can be inferred from other keypoints. Therefore for workflow analysis, Arms movement is probably the most descriptive followed by the Head and then Legs. Hence, we should prioritize subpose detections in that order.

Monitor reflections and occlusion pose problems for pose detection- and tracking in the Cath Lab. Reflections are a problem because the tested detectors are not trained to distinguish them from real human beings [40]. For workflow purposes the activity of persons is of interest, and their reflections serve only as noise. Occlusion renders persons invisible from individual views, causing False Negatives, or causing detected body joints to be connected incorrectly. Especially during tracking this presents an issue, as re-identification is difficult after losing- or wrongfully detecting a person. Tracking algorithms solve this problem through visual re-identification, but that does not work in the Cath Lab where everyone is dressed similarly.

CornerSW and WallS yield the best results in most situations. Although monitor reflections plague both, their limited occlusion and view of only the Cath Lab interior simplify the problem. CornerSE sees no reflections, but suffers from occlusions in the patient area by the monitor and operating table. CornerNW sees occlusion from the radiation shield and C-arm, reflections in the control room window and monitor, and people in the control room whose movements can be assumed to provide no relevant workflow information. The cardiologist facing away from CornerSW makes the use of this view for workflow analysis questionable. WallS, with its clear yet narrow view on the operating table surroundings, is an intuitive choice for workflow analysis during procedures. Before and after procedures, CornerSE provides a clearer view around the room entrances.

Human movement is descriptive of personnel activities [9–11], making reliable tracking important for workflow analysis. Unfortunately, no tested model yielded good tracking results on our dataset. AlphaP^T produced multiple identity swaps per minute and OpenPP^T merged or missed people.

Our model selection was limited with only one top-down algorithm and the dataset was relatively small. We did not annotate occluded keypoints which may have unfairly increased scores for more occluded viewpoints.

4.1. Future research

In following studies, more estimators [15,16,19] could be tested. Tracking should be done with a separate algorithm for better tracking performance. To overcome the visual differences between clothing in the Cath Lab and in general datasets like COCO, domain adaptation- or generalization methods could be explored [41]. With enough annotations, models could be re-trained for the Cath Lab or specific subposes. Interesting would be to annotate and detect keypoints in the C-arm, table, or lead screen. Expanding from single-view to multiple-view or 3D pose detection could help mitigate occlusion, as explored for the OR in [24–26]. It can be investigated how yielded poses can be used for automated recognition of e.g. personnel activities, workflow phases, or radiation exposure.

The insights from this work can aid the design of new computer vision setups in the Cath Lab or OR. For instance, cameras are best placed in a position which provides a clear view on personnel from the front, excluding reflective monitors or windows. As occlusion can rarely be avoided, a clear view of the Arms should be prioritized. When exploring pose detection, an algorithm can be chosen based on discussed trade-offs. In the design of a new pose detector one could focus on robustness against Cath Lab-specific occlusion, or distinguishing between real poses and reflections. When tracking, it is probably best not to use visual features due to similarities in appearance from sterile clothing.

The study shows that, considering confidence scoring and keypoint matching, AlphaP is the best-suited tested model in the Cath Lab. When only keypoint locations are sought, OpenPP could be a better choice. Due to identity swaps and pose merging, no tested tracker seems sufficient for use in workflow analysis. A new tracker should be developed based on the shortcomings highlighted in this paper. It should address the visual complexity of the Cath Lab specifically.

5. Conclusions

We annotated poses and identities in 2040 frames from ten CAG procedures. Detection- and tracking metrics AP, PCKh, AA and HOTA were calculated for the models from Table 1. Models showed significant performance differences, except when comparing different models of the same algorithm. The WallS and CornerSW viewpoints from Fig. 1 and the Arms keypoints were scored highest upon. The room coverage and decent results of CornerSE make this view a suitable alternative for workflow analysis, although its results vary with monitor positioning. OpenPP produced the most accurate keypoint locations in the Cath Lab. AlphaP(^{T}) yielded the best confidence scores, keypoint matching, and tracking results.

Declaration of competing interest

The authors who are affiliated with Philips (S. v R., C.B., B.H.W.H.) have financial interests in the subject matter, materials, and equipment, in the sense that they are employees of Philips.

None of the other authors have any financial relationship or competing interests.

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Acknowledgements

This research was funded by Rijksdienst voor Ondernemend Nederland grant number AI212005, and was sponsored in part by Philips Healthcare, study protocol NL71861.058.19.

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Medical Ethics Committee Leiden The Hague Delft (protocol code Z19.057, 30-10-2019) and by the board of the hospital where it was conducted. Informed consent was obtained from all subjects involved in the study.

References

[1] Timoh KN, Huaulme A, Cleary K, Zaheer MA, Lavoué V, Donoho D, et al. A systematic review of annotation for surgical process model analysis in minimally invasive surgery based on video. Surg Endosc 2023;37:4298–314. https://doi.org/10.1007/s00464-023-10041-w.

[2] Schouten AM, Flipse SM, van Nieuwenhuizen KE, Jansen FW, van der Eijk AC, van den Dobbelsteen JJ. Operating room performance optimization metrics: a systematic review. J Med Syst 2023;47:19. https://doi.org/10.1007/s10916-023-01912-9.

[3] Garrow CR, Kowalewski K-F, Li L, Wagner M, Schmidt MW, Engelhardt S, et al. Machine learning for surgical phase recognition: a systematic review. Ann Surg 2021;273(4):684–93. https://doi.org/10.1097/SLA.0000000000004425.

[4] Lalys F, Jannin P. Surgical process modelling: a review. Int J Comput Assisted Radiol Surg 2014;9:495–511. https://doi.org/10.1007/s11548-013-0940-5.

[5] Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. Nat Biomed Eng 2017;1:691–6. https://doi.org/10.1038/s41551-017-0132-7.

[6] Bkheet E, D'Angelo A-L, Goldbraikh A, Laufer S. Using hand pose estimation to automate open surgery training feedback. Int J Comput Assisted Radiol Surg 2023;18:1279–85. https://doi.org/10.1007/s11548-023-02947-6.

[7] Aksamentov I, Twinanda AP, Mutter D, Marescaux J, Padoy N. Deep neural networks predict remaining surgery duration from cholecystectomy videos. In: Med. image comput. comput.-assist. interv. - MICCAI 2017. Cham, Switzerland: Springer; 2017. p. 586–93.

[8] Berlet M, Vogel T, Ostler D, Czempiel T, Kähler M, Brunner S, et al. Surgical reporting for laparoscopic cholecystectomy based on phase annotation by a convolutional neural network (CNN) and the phenomenon of phase flickering: a proof of concept. Int J Comput Assisted Radiol Surg 2022;17:1991–9. https://doi.org/10.1007/s11548-022-02680-6.

[9] Saleem G, Bajwa UI, Raza RH. Toward human activity recognition: a survey. Neural Comput Appl 2023;35:4145–82. https://doi.org/10.1007/s00521-022-07937-4.

[10] Nguyen H-C, Nguyen T-H, Scherer R, Le V-H. Deep learning for human activity recognition on 3D human skeleton: survey and comparative study. Sens 2023;23(11):5121. https://doi.org/10.3390/s23115121.

[11] Wang C, Yan J. A comprehensive survey of RGB-based and skeleton-based human action recognition. IEEE Access 2023;11:53880–98. https://doi.org/10.1109/ACCESS.2023.3282311.

[12] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: common objects in context. In: Eur. conf. comput. vis. Cham, Switzerland: Springer; 2014. p. 740–55.

[13] Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: a survey. Proc IEEE 2023;111(3):257–76. https://doi.org/10.1109/JPROC.2023.3238524.

[14] Fang H-S, Li J, Tang H, Xu C, Zhu H, Xiu Y, et al. AlphaPose: whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans Pattern Anal Mach Intell 2023;45(6):7157–73. https://doi.org/10.1109/TPAMI.2022.3222784.

[15] Sun K, Xiao B, Liu D, Wang J. Deep high-resolution representation learning for human pose estimation. In: Proc. 2019 IEEE/CVF conf. comput. vis. pattern recognit. New York, USA: IEEE; 2019. p. 5686–96.

[16] Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, et al. ArtTrack: articulated multi-person tracking in the wild. In: Proc. 30th IEEE conf. comput. vis. pattern. recognit. New York, USA: IEEE; 2017. p. 1293–301.

[17] Cao Z, Hidalgo G, Simon T, Wei S-E, Sheikh Y. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 2021;43(1):172–86. https://doi.org/10.1109/TPAMI.2019.2929257.

[18] Kreiss S, Bertoni L, Alahi A. OpenPifPaf: composite fields for semantic keypoint detection and spatio-temporal association. IEEE Trans Intell Transp Syst 2022;23(8):13498–511. https://doi.org/10.1109/TITS.2021.3124981.

[19] Geng Z, Sun K, Xiao B, Zhang Z, Wang J. Bottom-up human pose estimation via disentangled keypoint regression. In: Proc. 2021 IEEE/CVF conf. comput. vis. pattern recognit. New York, USA: IEEE; 2021. p. 14671–81.

[20] Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: multi-object tracking by associating every detection box. In: Proc. 2022 eur. conf. comput. vis. Cham, Switzerland: Springer; 2022. p. 1–21.

[21] Berclaz J, Fleuret F, Turetken E, Fua P. Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell 2011;33(9):1806–19. https://doi.org/10.1109/TPAMI.2011.21.

[22] Bastian L, Wang TD, Czempiel T, Busam B, Navab N. DisguisOR: holistic face anonymization for the operating room. Int J Comput Assisted Radiol Surg 2023;18:1209–15. https://doi.org/10.1007/s11548-023-02939-6.

[23] Srivastav V, Issenhuth T, Abdolrahim K, de Mathelin M, Gangi A, Padoy N. MVOR: a multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. In: MICCAI-LABELS; 2018.

[24] Kadkhodamohammadi A, Gangi A, de Mathelin M, Padoy N. Articulated clinician detection using 3D pictorial structures on RGB-D data. Med Image Anal 2017;35:215–24. https://doi.org/10.1016/j.media.2016.07.001.

[25] Kadkhodamohammadi A, Gangi A, de Mathelin M, Padoy N. A multi-view RGB-D approach for human pose estimation in operating rooms. In: 2017 IEEE winter conf. appl. comput. vis. (WACV). New York, USA: IEEE; 2017. p. 363–72.

[26] Kadkhodamohammadi A, Padoy N. A generalizable approach for multi-view 3D human pose regression. Mach Vis Appl 2021;32:6. https://doi.org/10.1007/s00138-020-01120-2.

[27] Mirrashid N, Alibeiki Ę, Rakhtala SM. Development and control of an upper limb rehabilitation robot via ant colony optimization-pid and fuzzy-pid controllers. Int J Eng 2022;35(8):1488–93. https://doi.org/10.5829/ije.2022.35.08b.04.

[28] Fazli E, Rakhtala SM, Mirrashid N, Karimi HR. Real-time implementation of a super twisting control algorithm for an upper limb wearable robot. Mechatron 2022;84:102808. https://doi.org/10.1016/j.mechatronics.2022.102808.

[29] Wang Z, Butler R, van den Dobbelsteen JJ, Hendriks BHW, van der Elst M, Dauwels J. Towards robust object detection in unseen catheterization laboratories. In: IEEE int. workshop med. meas. appl. New York, USA: IEEE; 2024. p. 1–6.

[30] CVAT.ai Corporation. Computer vision annotation tool (CVAT). Available from: https://www.cvat.ai, 06 2023.

[31] Redmon J, Farhadi A. YOLOv3: an incremental improvement. Available from: arXiv: 1804.02767v1, 04 2018. https://doi.org/10.48550/arXiv.1804.02767.

[32] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 2015;37(9):1904–16. https://doi.org/10.1109/TPAMI.2015.2389824.

[33] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. 29th IEEE conf. comput. vis. pattern recognit. New York, USA: IEEE; 2016. p. 770–8.

[34] Ma N, Zhang X, Zheng H-T, Sun J. ShuffleNet V2: practical guidelines for efficient CNN architecture design. In: Proc. 15th eur. conf. comput. vis. Cham, Switzerland: Springer; 2018. p. 122–38.

[35] Padilla R, Passos WL, Dias TLB, Netto SL, da Silva EAB. A comparative analysis of object detection metrics with a companion open-source toolkit. Electron 2021;10(3):279. https://doi.org/10.3390/electronics10030279.

[36] Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: new benchmark and state of the art analysis. In: Proc. 2014 IEEE conf. comput. vis. pattern. recognit. New York, USA: IEEE; 2014. p. 3686–93.

[37] Kuhn HW. Variants of the Hungarian method for assignment problems. Nav Res Logist Q 1956;03(04):253–8. https://doi.org/10.1002/nav.3800030404.

[38] Luiten J, Ošep A, Dendorfer P, Torr P, Geiger A, Leal-Taixé L, et al. HOTA: a higher order metric for evaluating multi-object tracking. Int J Comput Vis 2021;129(2):548–78. https://doi.org/10.1007/s11263-020-01375-2.

[39] Hotelling H. The generalization of student's ratio. Ann Math Stat 1931;2(3):360–78.

[40] Park D, Park Y-H. Identifying reflected images from object detector in indoor environment utilizing depth information. IEEE Robot Autom Lett 2021;6(2):635–42. https://doi.org/10.1109/LRA.2020.3047796.

[41] Wang J, Lan C, Liu C, Ouyang Y, Qin T, Lu W. Generalization to unseen domains: a survey on domain generalization. IEEE Trans Knowl Data Eng 2023;35(8):8052–72. https://doi.org/10.1109/TKDE.2022.3178128.