Tobias Vogelsang*, Fabian B. Fahlbusch, Anna-Lena Behr, and Sebastian Zaunseder

# Real-Time Pose Estimation of Preterm Infants Using Depth Images

https://doi.org/10.1515/cdbme-2025-0211

**Abstract:** Early diagnosis of neurodevelopmental disorders in infants relies on accurate analysis of spontaneous movements. Achieving this requires fast and precise pose estimation methods tailored to infant-specific anatomy and motion. This study evaluates several pretrained YOLOv11-pose models for pose estimation in depth video recordings of preterm neonates and infants using the open source babyPose data set database. The fastest model (YOLOv11n-pose) has a inference time of 0.007 seconds. Considering a previously proposed data split without subject-wise separation between training and testing data, the most accurate model (YOLOv11m-pose) has a median root mean squared distance (RMSD) of 2.15. The median Dice Similarity Coefficient (DSC) and Recall (R) of the joints are 0.85 and 0.86, while the median DSC and R of the joint connections are 0.90 and 0.91. Considering a subject-wise separation of training and testing data, the results noticeably degrade, e.g. to a median DSC and R of the joints of 0.79 and 0.81, while the median DSC and R of the joint connections are 0.75 and 0.79. The present work demonstrates a fast and, copared to the literature, accurate approach to depth-based pose estimation in preterm neonates and infants paving the way for automated movement analysis as a clinical tool for early detection of developmental impairments. Particularly in semi-automated settings where subject-specific annotations can be provided, the results are convining. Regarding the abilities to generalize, more work is required to improve the results.

**Keywords:** preterm infants; pose estimation; YOLO; depth videos

## 1 Introduction

The analysis of spontaneous movement patterns in preterm neonates and infants plays a crucial role in identifying acute neurological conditions [1]. Certain movement abnormalities, such as seizure-like events, excessive restlessness due to withdrawal or asymmetric movements suggestive of perinatal stroke, may indicate underlying acute pathology [2, 3]. In contrast, the assessment of General Movements (GMs) [4] in the preterm and early postterm period provides a validated early biomarker for later neurodevelopmental outcomes, including cerebral palsy. Abnormal or absent GMs are strongly associated with adverse motor development, making their evaluation critical for early prognosis. Therefore, a comprehensive movement analysis, encompassing both acute movement abnormalities and predictive GM assessment, is essential for both clinical decision-making and the development of advanced diagnostic tools. Early identification of movement abnormalities enables the timely initiation of motor and neurodevelopmental therapy, potentially improving functional outcomes in affected infants.

Recent advances in deep learning (DL) detection or pose estimation have had a significant impact on the field of medical research, particularly in the domain of motion analysis [5, 6, 7]. YOLOv11-Pose is a feed-forward model that integrates real-time object detection with pose estimation to facilitate automated tracking of limb movements. Thereby, YOLOv11-pose is offering both speed and precision [8]. These features make it particularly well suited for neonatal monitoring, where both speed and precision are essential.

Yin et al. [9] demonstrate the effectiveness of a YOLO based DL model on 2D and 3D pose estimation of infants. The authors also show promising results on GMs classification especially for 3D pose estimations. There are also studies that succesfully implement DL models [10, 11, 12] and feature based machine learning approaches [13] for real-time pose estimation of infants with depth videos.

To the best of our knowledge, pose estimation using a YOLO model on infant depth images has not been done before. This contribution explores the implementation of YOLOv11-pose for neonatal pose estimation of depth videos. Thereto, fine tuned YOLOv11-pose networks have the potential to improve movement analysis and support early diagnosis and treatment strategies for preterm infants.

***Corresponding author: Tobias Vogelsang,** Chair of Diagnostic Sensing, Faculty of Applied Computer Science, University of Augsburg, Augsburg, Germany, e-mail: tobias.vogelsang@uni-a.de
**Fabian B. Fahlbusch, Anna-Lena Behr,** Neonatology and Pediatric Intensive Care, Faculty of Medicine, University of Augsburg, Augsburg, Germany
**Sebastian Zaunseder,** Chair of Diagnostic Sensing, Faculty of Applied Computer Science, University of Augsburg, Augsburg, Germany

# 2 Materials and Methods

## 2.1 Yolov11-Pose

YOLOv11-Pose is an advanced DL model designed for real-time human pose estimation. It extends the YOLOv11 architecture by incorporating a dedicated keypoint detection head, enabling the precise localization of anatomical landmarks [8].

The YOLOv11-pose architecture is composed of three fundamental components. First, the backbone serves as the primary feature extractor, utilizing convolutional neural networks to transform raw image data into multi-scale feature maps. The second component is the neck, which functions as an intermediate processing stage. This component employs specialized layers to aggregate and enhance feature representations across different scales. The third component, designated as the "head," fulfills the function of predicting object locations and classifications. This is achieved by leveraging the refined feature maps generated by the neck component to formulate the final outputs for object localization and classification.

The model predicts normalized bounding boxes along with keypoints for each detected object, allowing simultaneous object detection and pose estimation. In addition, the model predicts visibility scores whether keypoints are visible or not.

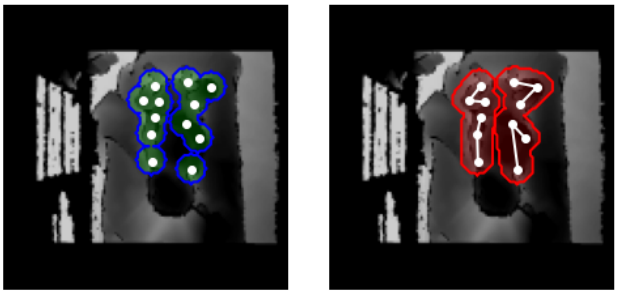
There are different YOLOv11-pose models with respect to size, prediction time and accuracy. YOLOv11n-Pose is the smallest and fastest version, while YOLOv11s-Pose and YOLOv11m-Pose are each one slower but more accurate.

To fine-tune and test the YOLOv11-Pose models, we use the open source babyPose data set [14]. The data set was developed to investigate the relationship between short- and long-term preterm birth complications and to explore models for non-contact monitoring of infant movements. The dataset contains 5 minute depth videos of 16 preterm infants with autonomous breathing hospitalized in the neonatal intensive care unit (NICU) of the G. Salesi Hospital (Ancona, Italy). This makes a total of 16.000 frames with a size of $640 \times 480$. Additionally, there is a version of the data set that is not free available and is employed in related works by the babyPose author. It contains 27.000 frames. The frames were captured by RGB depth camera (Orbbec® camera, Troy, Michigan, U.S.A.) placed over the open crib. The keypoints that represent the 14 pose coordinates are annotated by clinical partners of the authors [11] and stored in a *.xlsx file . The visual quality of the frames varies, making the babyPose dataset more variable and adapted to a real application. Fig. 1 shows two examples.

## 2.2 Preprocessing

In order to allow pose estimation with YOLOv11-pose, images and labels need to be transformed in a proper way and saved in a specified path structure. It is imperative that the images' width and height are equal. Thereto, the upper and lower parts of the images are extended by $640 \times 80$ zero matrices, so that the new image size is $640 \times 640$. Subsequently, the labels of babyPose are transformed and standardized into a YOLOv11 specified format. In addition, synthetic bounding box coordinates were added to the labels, as YOLOv11-Pose is based on object detection and therefore requires bounding boxes for proper operation. These coordinates were derived by calculating the minimum and maximum keypoint positions in each spatial direction (top, bottom, left, right), enclosing all visible joints. In order to guarantee that all pixels belonging to the infants are contained within the synthetic bounding boxes, it is necessary to extend them by 60 pixels. We are allowed to do so because the pose estimation of the YOLOv11 is not severely influenced by more precise bounding boxes.

To reduce the time and memory requirements of the training process, it was necessary to resize all frames to $128 \times 128$. Joints and joint connections with a radius of 6 pixels were extracted as a mask as in previous works [11, 10, 12, 13] by dilating the pose localizations. Example images of joints and joint connections are presented by Fig. 1.



**(a)** Example image of the joint mask.

**(b)** Example image of the joint connection mask.

**Fig. 1:** Example images of the joint and joint connection masks. The white points denote the joint location while the white lines denotes the joint connections. The colored area indicates the 6-pixel radius.

## 2.3 Metrics

A variety of metrics can be used to assess the effectiveness of pose estimation, allowing for a comprehensive evaluation of the results. The Root Mean Square Distance (RMSD) is a common metric for evaluating spatial accuracy in pose estimation. It calculates the mean Euclidean distance between predicted and annotated keypoints. The metric provides an intuitive indication of the average localization error and is particularly

suitable for applications where precise point positions are critical. The following equation defines the RMSD

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{y}}_\mathbf{i} - \mathbf{y}_\mathbf{i}\|^2} \tag{1}$$

$N$ describes the number of pixels in the depth images, while $\hat{\mathbf{y}}_\mathbf{i}$ and $\mathbf{y}_\mathbf{i}$ describe vectors that contain the x- and y-coordinate of the i-th pixel.

Two other metrics which are used in previous works are the Dice Similarity Coefficient (DSC) and the Recall (R). Both are defined as follows:

$$\text{DSC} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2}$$

$$\text{R} = \frac{TP}{TP + FN} \tag{3}$$

In this context, True Positive (TP) denote the true joint or joint connection, while False Positive (FP) denotes background pixels, respectively, that have been classified as joints. False Negative (FN) represents pixels belonging to a joint or joint connection connection erroneously designated as background. This means that we compute the DSC and R of the joints ($DSC_j$, $R_j$) and joint connections ($DSC_{jc}$, $R_{jc}$).

We also analyze the speed of the DL models by calculating the mean inference of the estimation.

## 2.4 Training

To train the YOLOv11-pose network, we used the Ultralytics framework [15], built on PyTorch. The learning rate is 0.002 and we used the ADAM optimizer. The batch size was 64 and the number of epochs was set to 300. Default data augmentation methods provided by Ultralytics were applied, including random flipping, translation, and scaling. For more efficient training, we used an NVIDIA GeForce RTX 3070 with an 8 GB GDDR6 VRAM.

For training and testing, we used the proposed split by the authors of babyPose data set, the babyPose split. The babyPose split does not pursue a separation of infants between training and test set, but is also used by related studies, which are mentioned in section 4. To quantify the ability of our models to generalize, we additionally implemented a leave-one-subject-out cross validation scheme (LOSOCV) where the model is trained each time on all infants but one and tested on this remaining subject.

# 3 Results

The pose estimation results and the mean inference time on the test data of babyPose data set are presented in Tab. 1. Tab. 2 shows the results of the joints $j$ and the joint connections $jc$.

**Tab. 1:** Pose estimation median [mean] results of the different Yolov11-pose models. The RMSD is given in pixels and the inference time in seconds.

| Model | $\text{RMSD}_{\text{babyPose}}$ | $\text{RMSD}_{\text{LOSOCV}}$ | Inference Time |
|---|---|---|---|
| YOLOv11n | 2.55 [11.86] | 12.81 [16.82] | 0.007 |
| YOLOv11s | 2.32 [11.64] | 11.16 [17.89] | 0.008 |
| YOLOv11m | 2.15 [11.48] | 9.46 [20.06] | 0.008 |

**Tab. 2:** Pose estimation median [mean] results of the joint and joint connections of different Yolov11-pose models.The DSC and R are unitless quantities. The upper half uses the babyPose split and the lower part uses the LOSOCVS.

| Model | $DSC_j$ | $DSC_{jc}$ | $R_j$ | $R_{jc}$ |
|---|---|---|---|---|
| YOLOv11n | 0.84 [0.83] | 0,89 [0.80] | 0.85 [0.84] | 0.90 [0.78] |
| YOLOv11s | 0.85 [0.83] | 0.89 [0.81] | 0.85 [0.84] | 0.90 [0.78] |
| YOLOv11m | 0.85 [0.83] | 0.90 [0.81] | 0.86 [0.85] | 0.91 [0.78] |
| YOLOv11n | 0.71 [0.68] | 0,70 [0.67] | 0.73 [0.69] | 0.68 [0.65] |
| YOLOv11s | 0.77 [0.71] | 0.71 [0.64] | 0.74 [0.72] | 0.77 [0.65] |
| YOLOv11m | 0.79 [0.75] | 0.75 [0.72] | 0.81 [0.78] | 0.79 [0.68] |

# 4 Discussion and Conclusion

As evidenced in Tab. 1 the smallest network is also the fastest one. Nevertheless, all three models show a fast estimation of more than 110fps and can be interpreted as real-time applications.

**Results using babyPose split:** It is evident that the median RMSD in Tab. 1 is lower than the mean RMSD of all models. This finding suggests the presence of significant outliers, which merit further investigation and analysis. The DSC and R in Tab. 2 shows a similar effect. However, it is less distinct for the joints than for the joint connections.

Tab. 3 compares the results of the present study with those of related studies. It shows that YOLOv11-pose outperform every other model by the RMSD. One possible explanation for this improved performance is the difference in training protocols. Specifically, the present study employed data augmentation techniques, which were not used in the compared studies, as well as a larger batch size and an extended training duration of 300 epochs (versus 100). These factors may have contributed to improved model performance, although potential overfitting cannot be ruled out and should be explored in future investigations. The Feature Based method and the BabyPoseNet are both less complex models. Therefore, they are not vulnerable to overfitting. Another significant difference is the fine tuning of pretrained weights. This impact should also be analyzed.

Comparing the DSC and R between the present work and the related ones, Tab. 3 shows that the YOLOv11 models outperform the results of the joint connection but BabyPoseNet shows slightly better results for the joint. By comparing the results with the models using more frames than the presented work we notice the same but the DSC and R of YOLOv11 models are sightly less than the best results of the DeA model.

**Results using LOSOVC:** As in Tab. 1 and Tab. 2, the mean and median error of the metrics used is, as expected, greater with a LOSOVC than with the babyPose split. Nevertheless, the performance of YOLOv11m is comparable to the performance of the related work despite LOSOCV. In terms of RMSD, YOLOv11m-pose is even better than the related works.

**Summary:** In summary, the YOLOv11-pose models are well suited for estimating neonatal and infant poses in depth video recordings and perform at least as well as models reported in related studies. Furthermore, the functionality of YOLOv11-pose can be extended by incorporating additional objects for detection such as, e.g., respiratory devices, feeding tubes or intravenous catheters. This enhancement may help prevent misclassification or displacement of these medical devices during pose estimation.

**Tab. 3:** Comparison of median pose estimation and joint connection results between the present work and related studies. The upper part uses the small data set (16000 frames) and middle part uses the bigger data set (27000 frames). Both parts uses babyPose split and the lower part shows the models validated with LOSOVC.

| Model | $DSC_j$ | $DSC_{jc}$ | $R_j$ | $R_{jc}$ | RMSD |
|---|---|---|---|---|---|
| YOLOv11n | 0.84 | 0.89 | 0.85 | 0.90 | 2.55 |
| YOLOv11s | 0.85 | 0.89 | 0.85 | 0.90 | 2.32 |
| YOLOv11m | 0.85 | 0.90 | 0.86 | 0.91 | 2.15 |
| BabyPoseNet [12] | 0.89 | 0.89 | 0.87 | 0.86 | 11.27 |
| Feature Based [13] | - | - | - | - | 11.20 |
| DeA [10] | 0.90 | 0.90 | - | 0.90 | 10.79 |
| EDANet [11] | 0.79 | 0.81 | 0.69 | 0.70 | - |
| TwinEDA [11] | 0.89 | 0.88 | 0.86 | 0.83 | - |
| TwinEDAv0 [11] | 0.80 | 0.83 | 0.72 | 0.73 | - |
| TwinEDAv1 [11] | 0.88 | 0.87 | 0.85 | 0.81 | - |
| YOLOv11n | 0.71 | 0.70 | 0.73 | 0.68 | 16.82 |
| YOLOv11s | 0.77 | 0.71 | 0.74 | 0.77 | 11.16 |
| YOLOv11m | 0.79 | 0.75 | 0.81 | 0.79 | 9.46 |

# References


[1] K. Fry-Hilderbrand, Y. P. Chen, and A. Howard. “Validating a System to Monitor Motor Development of At-Risk Infants in Black Communities: A Case Study.” In: *2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. Philadelphia, PA, USA: IEEE, Dec. 2021, pp. 1–6.

[2] A Basu, J Baggaley, and K Simpson. “G106(P) Case notes review of perinatal stroke in term and preterm infants in the Northern region over a 10 year period.” en. In: *Arch Dis Child* 101.Suppl 1 (Apr. 2016), A60.2–A60.

[3] Karin B Nelson and John K Lynch. “Stroke in newborn infants.” en. In: *The Lancet Neurology* 3.3 (Mar. 2004), pp. 150–158.

[4] Christa Einspieler et al. “The qualitative assessment of general movements in preterm, term and young infants — review of the methodology.” en. In: *Early Human Development* 50.1 (Nov. 1997), pp. 47–60.

[5] Paolo Bonato et al. “Position paper on how technology for human motion analysis and relevant clinical applications have evolved over the past decades: Striking a balance between accuracy and convenience.” en. In: *Gait & Posture* 113 (Sept. 2024), pp. 191–203.

[6] Merryn D. Constable, Hubert P. H. Shum, and Stephen Clark. “Enhancing surgical performance in cardiothoracic surgery with innovations from computer vision and artificial intelligence: a narrative review.” en. In: *J Cardiothorac Surg* 19.1 (Feb. 2024), p. 94.

[7] Johanna Gleichauf et al. “Sensor Fusion for the Robust Detection of Facial Regions of Neonates Using Neural Networks.” en. In: *Sensors* 23.10 (May 2023), p. 4910.

[8] Rahima Khanam and Muhammad Hussain. *YOLOv11: An Overview of the Key Architectural Enhancements*. arXiv:2410.17725 [cs]. Oct. 2024.

[9] Wang Yin et al. “A self-supervised spatio-temporal attention network for video-based 3D infant pose estimation.” en. In: *Medical Image Analysis* 96 (Aug. 2024), p. 103208.

[10] Lucia Migliorelli, Emanuele Frontoni, and Sara Moccia. “An accurate estimation of preterm infants’ limb pose from depth images using deep neural networks with densely connected atrous spatial convolutions.” en. In: *Expert Systems with Applications* 204 (Oct. 2022), p. 117458.

[11] Lucia Migliorelli et al. “TwinEDA: a sustainable deep-learning approach for limb-position estimation in preterm infants’ depth images.” en. In: *Med Biol Eng Comput* 61.2 (Feb. 2023), pp. 387–397.

[12] Sara Moccia et al. “Preterm infants’ limb-pose estimation from depth images using convolutional neural networks.” In: *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Siena, Italy: IEEE, July 2019, pp. 1–7.

[13] Sara Moccia et al. “Preterm Infants’ Pose Estimation With Spatio-Temporal Features.” In: *IEEE Trans. Biomed. Eng.* 67.8 (Aug. 2020), pp. 2370–2380.

[14] Lucia Migliorelli. *BabyPose_Dataset*. June 2020.

[15] Glenn Jocher, Jing Qui, and Ayush Chaurasia. *Ultralytics YOLO*. Jan. 2023.