

# INFORMATION TECHNOLOGY, COMPUTER SCIENCE AND MANAGEMENT ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ



UDC 004.89

Original Empirical Research

<https://doi.org/10.23947/2687-1653-2025-25-3-221-232>

## Reconstructing a Full-Body Model from a Limited Set of Upper-Limb Motion Data

 Artem D. Obukhov , Daniil V. Teselkin 

Tambov State Technical University, Tambov, Russian Federation

✉ [obukhov.art@gmail.com](mailto:obukhov.art@gmail.com)

EDN: HLYDVW

### Abstract

**Introduction.** Accurate reconstruction of the human body model is required when visualizing digital avatars in virtual simulators and rehabilitation systems. However, the use of exoskeleton systems can cause overlapping and shielding of sensors, making it difficult for tracking systems to operate. This underlines the urgency of the task of reconstructing a human body model based on a limited set of data on arm movements, both in the field of rehabilitation and in sports training. Existing studies focus on either large-scale IMU networks or full video monitoring, without considering the issue of reconstructing a body model based on arm motion data. The objective of this research is to develop and test machine learning methods aimed at reconstructing body model coordinates using limited data, such as arm position information.

**Materials and Methods.** To conduct the study, a virtual simulation environment was created in which a virtual avatar performed various movements. These movements were recorded by cameras with a first-person and side view. The positions of the keypoints of the body model relative to the back point were saved as reference data. The regression task considered was to reconstruct the user's arm positions in a full body model in five different variations, including keypoint coordinates extracted from a video and a virtual scene. The task also involved comparing different regression models, including linear models, decision trees, ensembles, and three deep neural networks (DenseNN, CNN-GRU, Transformer). The accuracy was estimated using MAE and the mean Euclidean deviation of body segments. Experimental studies were conducted on five datasets, whose size varied from 25 to 180 thousand frames.

**Results.** The experiments showed that ensembles (LightGBM) were best-performing in most situations. Among neural network models, the CNN-GRU-based model provided the lowest error. Training models on a sequence of 20 frames did not provide significant improvement. Using the inverse kinematics module on a number of scenarios allowed reducing the error to 3%, but in some cases worsened the final result.

**Discussion.** The analysis of the results obtained showed low reconstruction accuracy when using computer vision datasets, as well as the lack of superiority of complex models over simpler ensembles and linear models. However, the trained models allowed, with some error, for the reconstruction of the position of the user's legs for a more reliable display of the digital model of his body.

**Conclusion.** The data obtained showed the complexity of solving the problem of reconstructing a human body model using a limited amount of data, as well as a large error in a number of machine learning models. The comparison of models on different datasets proved low applicability of first-person data that did not contain information on the distance to the arm s. On the other part, using absolute values of arm positions as input information provided for the reconstruction of the body model with significantly less error.

**Keywords:** reconstruction of the human body model, machine learning, virtual simulators, limited data




**Acknowledgements.** The authors would like to thank the head of the scientific project, M.N. Krasnyansky, Dr.Sci.(Engineering), Professor, Rector of TSTU for organizing the research process.

**Funding Information.** The research is done with the financial support from the Ministry of Education and Science of the Russian Federation within the framework of the project “Development of an Immersive Virtual Reality Interaction System for Professional Training Based on an Omnidirectional Platform” (124102100628-3).

**For Citation.** Obukhov AD, Teselkin DV. Reconstructing a Full-body Model from a Limited Set of Upper-Limb Motion Data. *Advanced Engineering Research (Rostov-on-Don)*. 2025;25(3):221–232. <https://doi.org/10.23947/2687-1653-2025-25-3-221-232>

Оригинальное эмпирическое исследование

## Подход к реконструкции модели тела на основе ограниченного набора данных о двигательной активности рук

А.Д. Обухов  , Д.В. Теселкин 

Тамбовский государственный технический университет, г. Тамбов, Российская Федерация

 [obuhov.art@gmail.com](mailto:obuhov.art@gmail.com)

### Аннотация

**Введение.** Точная реконструкция модели тела человека крайне важна для визуализации цифровых аватаров в виртуальных тренажерах и реабилитационных системах. Однако использование экзоскелетных систем может привести к перекрытию и экранированию датчиков, что затрудняет работу систем отслеживания. Это подчеркивает актуальность задачи реконструкции модели тела человека на основе ограниченного набора данных о движениях рук, как в сфере реабилитации, так и в спортивной подготовке. Существующие исследования сосредоточены либо на масштабных IMU-сетях, либо на полном видеоконтроле, не рассматривая вопрос реконструкции модели тела на основе данных о движениях рук. Цель данной работы заключается в разработке и тестировании методов машинного обучения, направленных на восстановление координат модели тела с использованием ограниченных данных, например, информации о положении рук.

**Материалы и методы.** Для проведения исследования была сформирована виртуальная имитационная среда, в которой виртуальный аватар выполнял различные движения. Эти движения фиксировались камерами с видом от первого лица и боковой. В качестве эталонных данных сохранялись положения ключевых точек модели тела относительно точки спины. Рассматривалась задача регрессии, целью которой было восстановление положения рук пользователя в полной модели его тела в пяти различных вариациях, включающих координаты ключевых точек, извлеченные из видео и виртуальной сцены. Задача также подразумевала сравнение различных моделей регрессии, среди которых были линейные модели, деревья решений, ансамбли, а также три глубокие нейронные сети (DenseNN, CNN-GRU, Transformer). Точность оценивалась с использованием MAE и среднего Евклидова отклонения сегментов тела. Проведены экспериментальные исследования на пяти наборах данных, размер которых варьировался от 25 до 180 тысяч кадров.

**Результаты исследования.** Эксперименты показали, что ансамбли (LightGBM) наиболее эффективны в большинстве ситуаций. Среди нейросетевых моделей наименьшую погрешность обеспечила модель на базе CNN-GRU. Обучение моделей на последовательности из 20 кадров не дало значительного улучшения. Применение модуля инверсной кинематики на ряде сценариев позволяет снизить погрешность до 3 %, но в ряде случаев ухудшает итоговый результат.

**Обсуждение.** Анализ полученных результатов показал низкую точность реконструкции при использовании наборов данных от компьютерного зрения, а также отсутствие превосходства сложных моделей перед более простыми ансамблями и линейными моделями. Тем не менее, обученные модели позволяют с некоторой погрешностью восстанавливать положение ног пользователя для более достоверного отображения цифровой модели его тела.

**Заключение.** Полученные данные показывают сложность решения задачи реконструкции модели тела человека при использовании ограниченного объема данных, а также большую погрешность у ряда моделей машинного обучения. Сравнение моделей на различных наборах данных показало низкую применимость данных от первого лица, не содержащих информацию о расстоянии до рук. С другой стороны, использование в качестве входной информации абсолютных значений положения рук позволяет осуществить реконструкцию модели тела со значительно меньшей погрешностью.

**Ключевые слова:** реконструкция модели тела человека, машинное обучение, виртуальные тренажеры, ограниченные данные

**Благодарности.** Авторы благодарят руководителя научного проекта М.Н. Краснянского, доктора технических наук, профессора, ректора Тамбовского государственного технического университета за организацию научно-исследовательского процесса.

**Финансирование.** Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках проекта «Разработка иммерсивной системы взаимодействия с виртуальной реальностью для профессиональной подготовки на основе всенаправленной платформы» ([124102100628-3](https://doi.org/10.23947/2687-1653-2025-25-3-221-232)).

**Для цитирования.** Обухов А.Д., Теселкин Д.В. Подход к реконструкции модели тела на основе ограниченного набора данных о двигательной активности рук. *Advanced Engineering Research (Rostov-on-Don)*. 2025;25(3):221–232. <https://doi.org/10.23947/2687-1653-2025-25-3-221-232>

**Introduction.** Virtual simulators integrated with controlled exoskeletons allow simulating physical activity and rehabilitation exercises in a controlled environment [1]. To achieve the maximum immersion effect, precise tracking of the user's entire body kinematics is required to form a virtual avatar that corresponds to real human movements. However, upper exoskeletons can block sensors that demand direct visual control (e.g., HTC Vive Tracker), as well as create electromagnetic interference and limit the view of external cameras, which requires the use of additional markers [2]. In these cases, traditional tracking systems, such as infrared markers and multiple capture cameras, cannot consistently provide a complete set of position data for all body segments [3]. Thus, researchers are faced with the task of reconstructing a full body model based on limited information, such as arm or hand position data.

One of the possible solutions is to use wearable sensors such as inertial measurement units (IMU). However, a sufficient number of sensors (at least 11, and often up to 18 elements) are required to fully reconstruct the body model [4]. As the number of sensors reduces, the accuracy of the data decreases sharply. At the same time, computer vision technologies are actively developing and are increasingly used in virtual reality systems to track hands and fingers, which makes the task of reconstructing a body model from limited data obtained only from the user's hands particularly topical [5].

The task of reconstructing a full human skeletal model from a limited set of visual data (e.g., arm movements) has a considerable practical and scientific value. Classic marker-based motion capture systems used infrared cameras and joint markers, while computer vision-based algorithms are being developed and successfully implemented for markerless solutions. Modern convolutional neural networks, such as OpenPose, BlazePose, and MediaPipe Pose, are capable of detecting 2D positions of body keypoints without additional labels [6]. These methods effectively detect visible points (arms, shoulders, pelvis, etc.), but without depth information from a single camera, the distance to the body is not restored, which complicates the full 3D reconstruction of the body model. A solution to this problem can be found by means of stereo cameras and triangulation methods [7]. Using such approaches, modern models (e.g., MediaPipe Pose) can track up to 33 keypoints with an error of about 1–2 cm. Under real conditions, that allows obtaining 3D coordinates of major joints (e.g., hands, elbows, knees) through combining data from several cameras and minimizing projection error. However, such tracking systems often prove unsuitable if the cameras only see the hands, and it is necessary to estimate the rest of the skeleton based on the movements of the hands without direct visual control. This is a hot issue in virtual reality systems, where cameras are present only on the headset and record only the user's hands in the working area. In this regard, it is required to consider existing approaches to solving this problem.

The main area of work on this task involves the use of regression-based methods or neural networks that are able to complement the pose relying only on partial data [4]. For example, regression models trained on video pairs with partially masked bodies and arms can reconstruct missing body parts under complex conditions [8, 9]. This indicates that modern models are indeed able to display full body pose from partial visual information about the arms. Other studies use neural network architectures that focus on motion sequences, such as recurrent networks (LSTM/GRU) and specifically Transformer [10, 11]. For example, paper [12] describes AvatarPoser, a Transformer-based model that predicts a full 3D pose of the body (including legs and torso) from the position of the head and arms. This system extracts deep features from incoming motion signals and separates global body movement and local joint orientations. For accurate pose matching, limb optimization is also performed by the inverse kinematics method [12]. Furthermore, the idea of improving the robustness of predictions in the absence of visibility is implemented in the EgoPoser model, which also relies on the Transformer mechanisms to account for intermittent arm motion data, ensuring stable predictions [13]. It is worth noting that training these models requires marking full poses, which leads to the need to use large datasets, such as Human3.6M, CMU MoCap/AMASS, MPI-INF-3DHP and others, where there is synchronized video and a 3D skeleton [14, 15]. However, existing datasets that compare the first-person view with a full body model are insufficient, which makes the task of collecting and comparing such data hot issues. The formation of such a dataset can be organized by recreating human movements in a virtual scene, where you can flexibly adjust the position of virtual cameras for recording video, and obtain accurate coordinates of body points with the required frequency [4].

In this study, the main subject for the implementation of the obtained scientific results is a virtual simulator system based on an upper controlled exoskeleton. Modern models of VR-headsets are focused on positioning by cameras, assuming that the main source of information on arm movement comes from the built-in headset camera used to recognize arms. In addition, to expand the experimental base and identify patterns in human movements, it is assumed that there is

an external frame-by-frame capture system that records the user's body position as a whole. The objective of this research is to develop and test machine learning methods for reconstructing body coordinates based on partial arm position data. In conclusion of the study, it is planned to compare both classical regression models and neural network models, including modern architectures based on attention mechanisms, which will allow us to evaluate the advantages of each approach in various experimental scenarios.

**Materials and Methods.** First, the procedure for collecting and primary processing the data was considered. The data was collected in a virtual environment, where the process of using a VR-headset with a camera was imitated. All data (first-person view, side camera view) were tracked by virtual cameras. Then, the videos were processed by MediaPipe library models, which allowed for hand detection to isolate 21 keypoints of the palm, as well as extracting data on 33 key points of the body model from the side virtual camera. In parallel, the true metric coordinates of all body segments (18 key points of the standard digital avatar model specified in the Unity game engine), including hand position points, were recorded in the virtual space. These real coordinates (reference) formed the target  $Y$  set for most scenarios. Using a virtual camera made it possible to bypass the limitations of physical sensors and obtain reference information about the body pose. MediaPipe was selected as the main hand tracking framework due to its modular processing graph system and ready-made ML models (palm detector and full body model). We introduced the abbreviation “CV” for the data obtained during processing by computer vision and MediaPipe models (denoted as  $cx_i, cy_i, cz_i$ ), and the “reference” was understood as the metric coordinates of body points (denoted as  $vx_i, vy_i, vz_i$ ), recorded in the virtual scene relative to the user's back.

Next, we considered the data preparation procedure for various regression scenarios. To analyze the machine learning models and their capabilities, 5 datasets (experiments) were formed, differing in which features  $X$  were used and which target variables  $Y$  were predicted:

1) Set 1 “Arms (first-person view)  $\rightarrow$  Arms (reference)”:  $X = \{(cx_i, cy_i, cz_i)\} \in \mathbb{R}^{63}, i = 1-21$  — coordinates of keypoints of arms in FPV (63 values),  $Y = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{18}, i = 1-6$  — metric coordinates of the same points of arms (18 values).

2) Set 2 “Arms (first-person view)  $\rightarrow$  Body (reference)”:  $X = \{(cx_i, cy_i, cz_i)\} \in \mathbb{R}^{63}, i = 1-21$  — coordinates of the arms (obtained from FPV, 63 values),  $Y = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{54}, i = 1-18$  — metric coordinates of all points of the body (54 values). This way, a complete reconstruction of the body is performed on the basis of the arms data.

3) Set 3 “Arms (first person view)  $\rightarrow$  Body (CV)”:  $X = \{(cx_i, cy_i, cz_i)\} \in \mathbb{R}^{63}, i = 1-21$  — coordinates of arms (FPV based on CV, 63 values),  $Y = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{99}, i = 1-33$  — coordinates of 33 body points from an additional side view video (99 values). It differs from the previous task in that regressions are performed exclusively on CV data.

4) Set 4 “Body (CV)  $\rightarrow$  Body (reference)”:  $X = \{(cx_i, cy_i, cz_i)\} \in \mathbb{R}^{99}, i = 1-33$  — coordinates of body points (side camera view, 99 values),  $Y = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{54}, i = 1-18$  — metric coordinates of all body points (54 values). The task is to check the accuracy of direct data conversion from CV to metric values of 18 keypoints.

5) Set 5 “Arms (reference)  $\rightarrow$  Body (reference)”:  $X = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{18}, i = 1-6$  — metric coordinates of arm points (18 values),  $Y = \{(vx_i, vy_i, vz_i)\} \in \mathbb{R}^{54}, i = 1-18$  — metric coordinates of the entire body (54 values). It differs from set 2 in that only reference data are used. Thus, the very fact of reconstructing movement based on a limited set of points is verified.

Next, we consider the models used to solve the five regression problems mentioned above. The architectures of all the models in different problems are similar; the differences for each set are only in the input and output dimensions. In total, two classes of models are considered: classic regression models from the Scikit-Learn library (as well as XGBoost and LightGBM models) and neural network models based on the Keras framework [16, 17].

Classical models include: LinearRegression, ElasticNet (with L1/L2 regularization), ensembles of trees (RandomForestRegressor, HistGradientBoostingRegressor), boosting (XGBRegressor, LightGBMRegressor), and KNN regressor. Since the target variable includes multiple outputs (point coordinates), the models are wrapped in a MultiOutputRegressor, which allows predicting all parameters simultaneously. All tree-type models are configured with 100 trees and a depth of 5, while boosted models have learning\_rate = 0.05.

Next, we consider neural network architectures.

A Fully Connected Network (we denote it as DenseNN). The input layer corresponds to the feature dimension  $X$  (from now on, it depends on the dataset), followed by 4 fully connected layers: 256, 512, 1024, and 128 neurons with ReLU activation and Dropout sparsification layers (25%). The model ends with an output layer of dimension  $Y$  (also depends on the dataset). Batch normalization (BatchNorm) and the Adam optimizer ( $lr = 1e-3$ ) with the MSE loss function are used.

Convolutional Recurrent Network (CNN-GRU). After the input layer, 1D convolution (128 filters, kernel = 3) and BatchNormalization are applied. This is followed by a GRU layer (128 units) with sequence return. The attention mechanism is implemented: a dense layer with tanh activation above the GRU output produces frame weights, which are then multiplied with the GRU output using softmax and summed. Then there is a fully connected layer of 128 neurons with activation of ReLU and Dropout (30%), after which the output layer comes. Adam optimizer ( $lr = 1e-3$ ) with MSE loss function.



Transformer. First, several 1D convolutions (kernel = 3, dilation\_rate 1, 2 and 4) are applied to create local context. Then, a Squeeze-and-Excite layer is added for adaptive filtering of channels [18]. Furthermore, trainable positional embeddings and 3 encoder blocks of the transformer are introduced, each implementing MultiHeadAttention (4 heads, key in 64/4 size), subsequent summation and normalization, and then — a two-layer dense network (size 256, 64) with Dropout — again summation and normalization. After the encoders, GlobalAveragePooling1D is performed, then — a fully connected layer of 128 neurons with ReLU activation and Dropout (25%), followed by a linear output. Adam optimizer ( $lr = 1e-3$ ) with MSE loss function.

Based on the conducted review and the existing experience in this area, we can propose an approach to solving the regression problems under consideration. A dataset of animation of typical human movements is formed, which is applied to a virtual avatar in a simulation scene. The movements are recorded using several virtual cameras: one of them is located at the avatar's eye level (FPV), and the second watches it from the side (side camera), covering its entire height. Additionally, metric values of 18 points of the body model are recorded. The final data view for each source is shown in Figure 1.

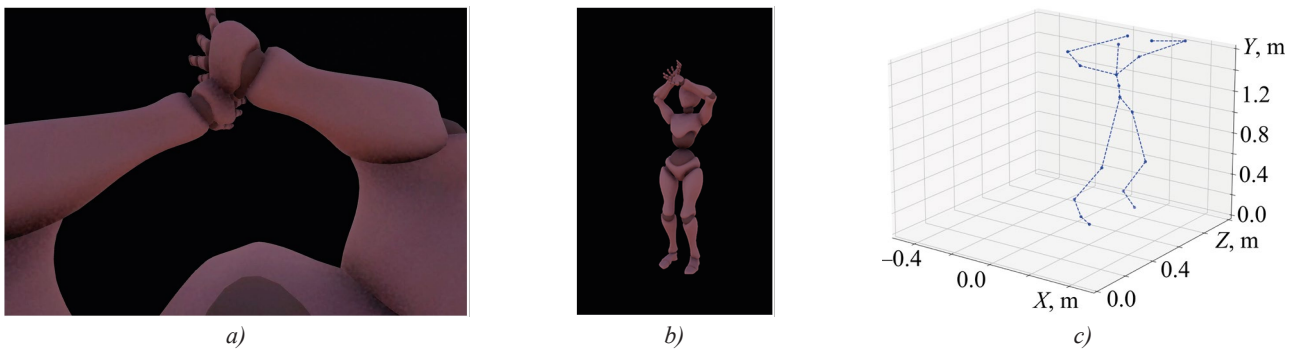


Fig. 1. Source data: *a* — frame from first-person camera; *b* — frame from side camera; *c* — skeleton constructed using reference data

Video data is processed by the corresponding models (MediaPipe Pose/Hands), after which the point coordinates are saved in arrays. Then, within the framework of the proposed approach, machine learning models are trained, which, based on the initial data alone (e.g., information about the arms), form a complete 3D configuration of the body. After predicting the pose, the elbow and knee joints can be further adjusted so that the segment lengths and limb positions better match the arm signatures. Also, to assess the contribution of the temporal context to the accuracy of the reconstruction problem, it is proposed to conduct an additional experiment to solve the regression problem for each set not based on the data of one frame, but on a certain sequence of  $N$ -frames.

In this paper, there is no focus on correcting the body model after reconstruction according to the inverse kinematics rules. The major objective is to train and compare a set of regressors (linear, tree-type, KNN models) and neural networks (DenseNN, CNN-GRU and Transformer) to determine the most accurate model. The selection is based on the metrics of the mean absolute error (MAE), the total deviation (Euclidean distance) for all points of the model from the reference ones, as well as on assessing the computational complexity (prediction time). This solves the problem of reconstructing a body model based on a limited set of information about arm movements. In addition, other regression options are considered within the experimental section. The calculation is made using the following formulas:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

$$\Delta = \frac{1}{MJ} \sum_{n=1}^M \sum_{j=1}^J \|y_{n,j} - \hat{y}_{n,j}\|, \quad \|v\| = \sqrt{v_x^2 + v_y^2 + v_z^2},$$

where  $y_i$  — true value;  $\hat{y}_i$  — model prediction,  $N$  — number of compared values,  $M$  — number of frames;  $J$  — number of joints (keypoints),  $y_{n,j}, \hat{y}_{n,j} \in \mathbb{R}^3$  — true and predicted 3-D position vector of  $j$ -th joint in  $n$ -th frame.

**Research Results.** In accordance with the described methodology, data was collected on 11 types of various complex animations, including body movements, jumps and active movements. Nine types were used for training, and two — for validation (data from them were not used in the training process). The total volume was 239968 records, but at each stage, filtering and selection of records was performed in the event that one of the sources did not return correct values (most often, this involved obtaining arm coordinates using computer vision). Thus, for sets 1–3, 25 and 8 thousand records were selected for training and validation, for sets 4 and 5 — 183 and 56 thousand, respectively. During the learning process, the training sample was further divided in the ratio 75/25. The dimension of the data for each experiment was indicated above when describing the corresponding sets. Figure 2 shows the comparative results of all models for all data sets using the MAE metric. Figure 3 shows the total deviation metric. And Figure 4 shows a comparison of models based on the calculation time of one forecast. Next, we compare the results obtained.

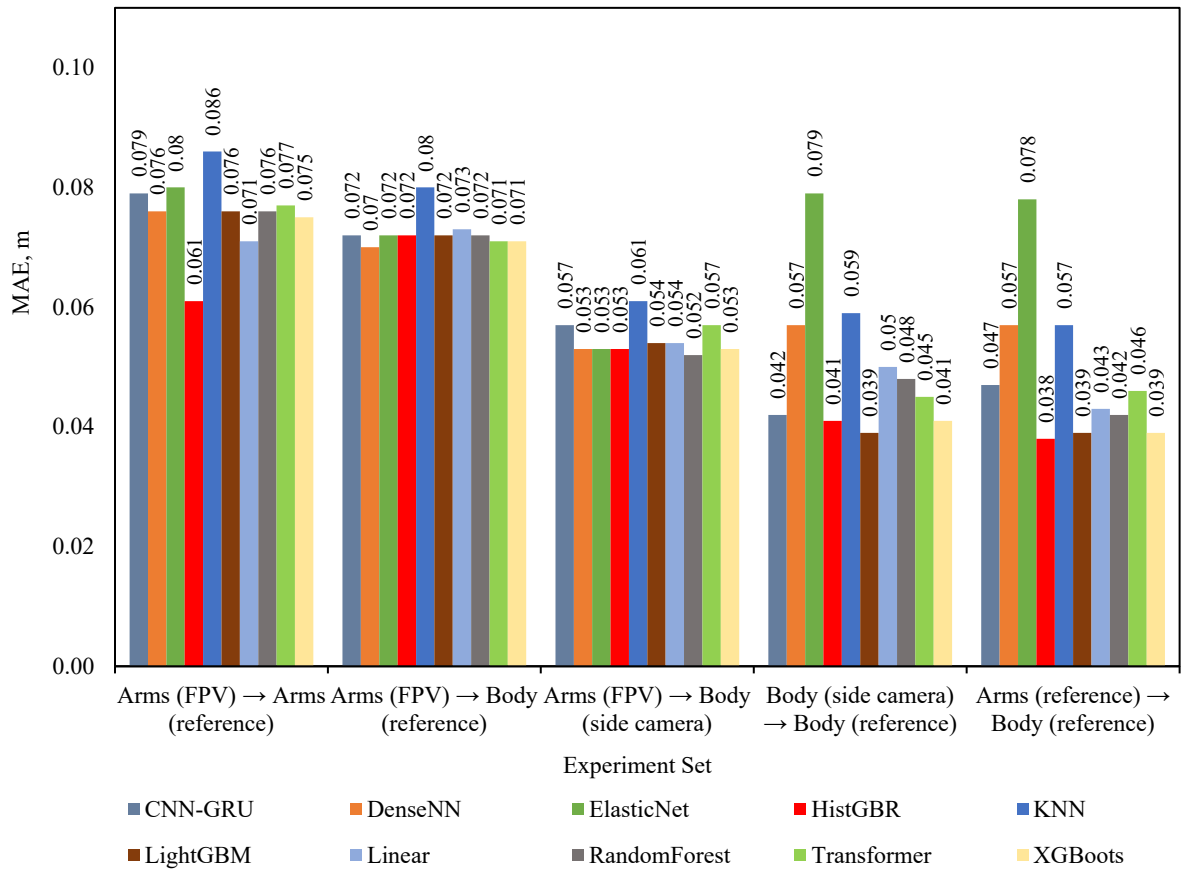


Fig. 2. Comparison of models by MAE metric

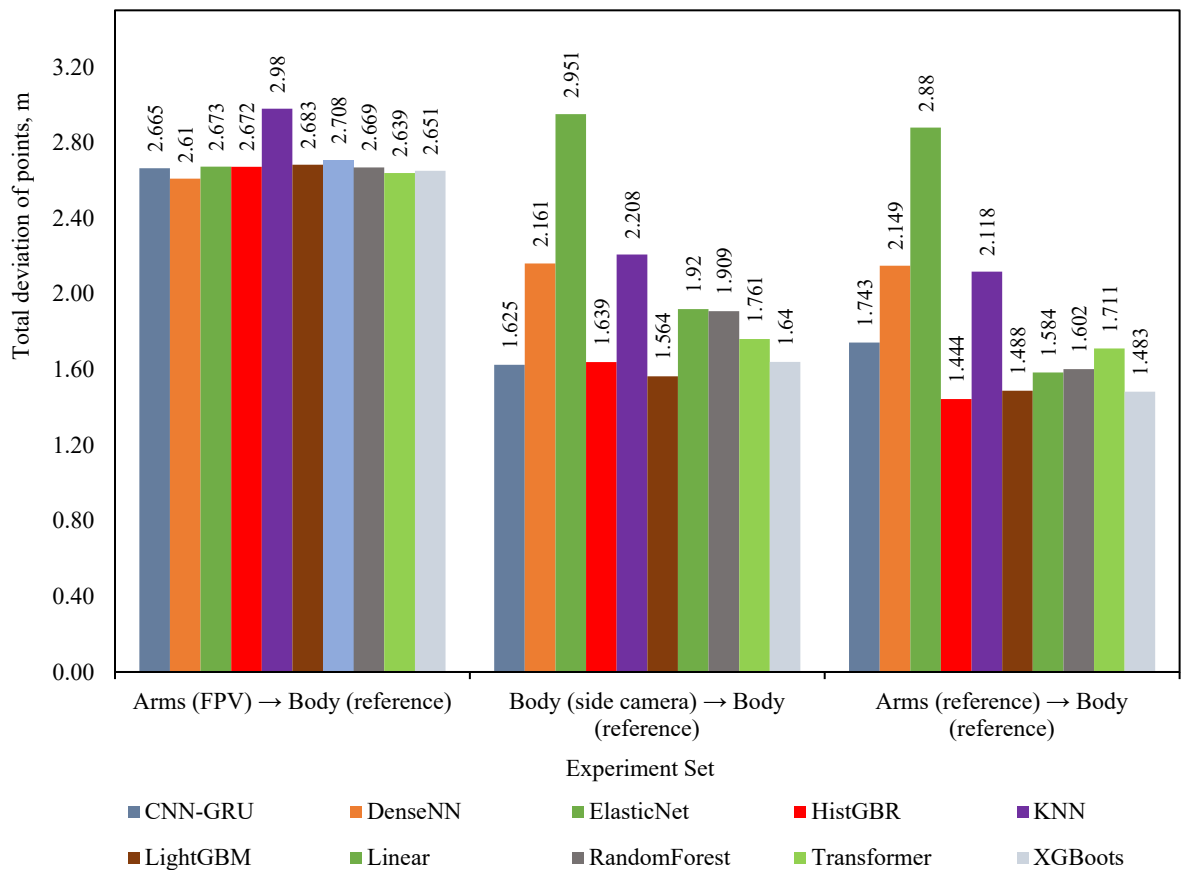


Fig. 3. Comparison of models by total deviation

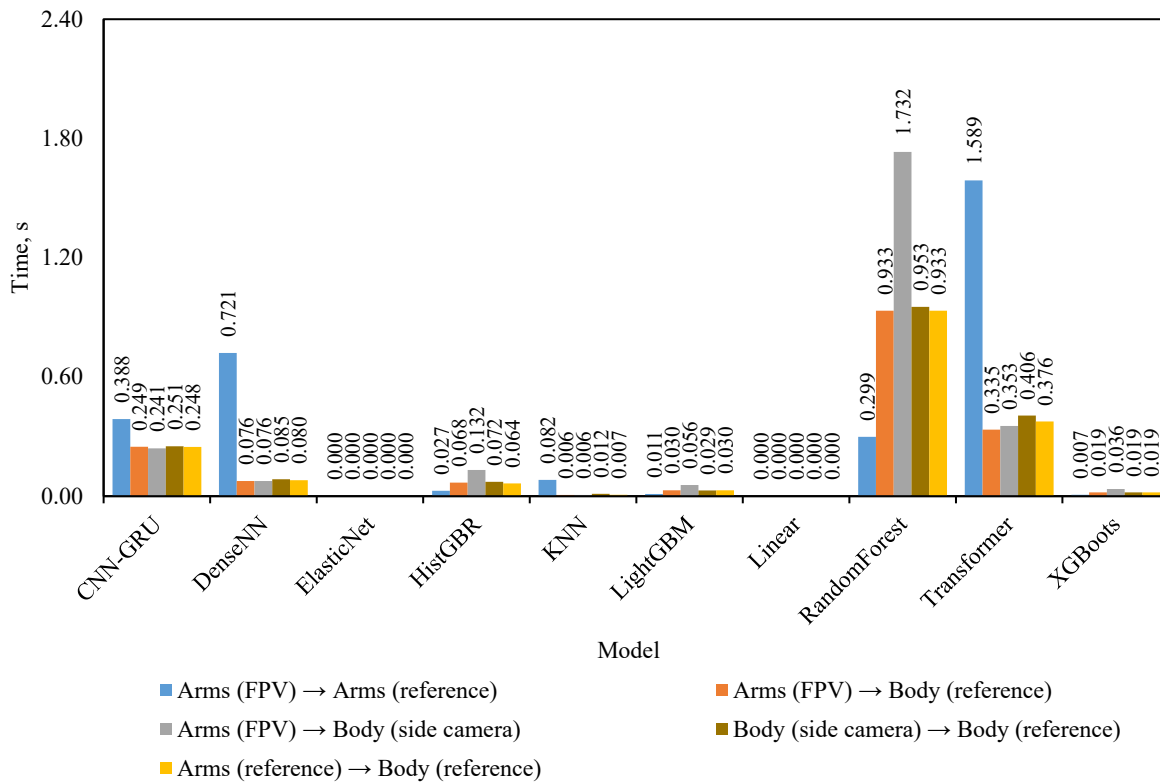


Fig. 4. Comparison of models by performance

The analysis of the data obtained shows the heterogeneity of the behavior of the models when changing the source of input information and on various metrics. In most scenarios, gradient ensembles (HistGBR, LightGBM, XGBoost, and RandomForest) demonstrate the lowest MAE error. Neural network models perform worse, especially when complex tasks of reconstructing a body based on arm data (CV) are considered. Nevertheless, if we evaluate all models according to MAE, then it is not possible to single out a clear leader. On the other hand, the total deviation of all points (Fig. 3) clarifies significantly the situation when solving three regression problems. Ensembles are superior, as in the previous case, but CNN-GRU is the best among neural network models. The obtained values of the total deviation, ranging from 1.4 to 3.5 meters, indicate low efficiency of solving the regression problem by all models, especially on the set of “Arms (FPV) → Body (reference)”. When evaluating the performance of models by computation time, it can be noted that classical machine learning models (linear and ensembles) have sufficient performance for real-time use. At the same time, CNN-GRU, Transformer and specifically Random Forest, are extremely computation-consuming, which makes them applicable only in offline (not real-time) systems. For DenseNN, long calculations are often observed at the first call of the model.

Given our experience in body reconstruction tasks, it is important to evaluate models not only by the specified metrics, but also visually. For this purpose, we reconstruct body skeletons for sets 2, 4, and 5 using the LightGBM and CNN-GRU models. This comparison (Fig. 5) allows us to evaluate how the most accurate architecture (LightGBM) differs visually from the more complex one (CNN-GRU).

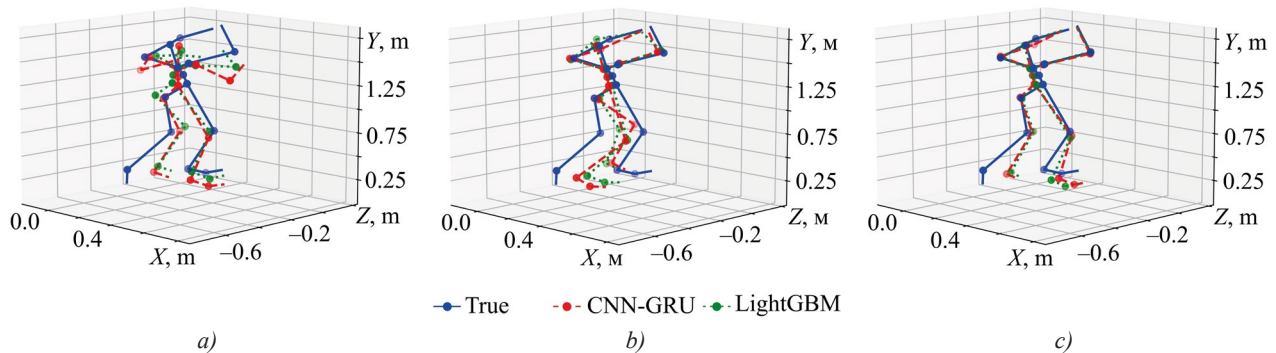


Fig. 5. Visual comparison of CNN-GRU and LightGBM models:  
 a — on the set “Arms (FPV) → Body (reference)”; b — on the set “Body (CV) → Body (reference)”; c — on the set “Arms (reference) → Body (reference)”

Visual comparison shows that there is a noticeable difference between the CV data and the real position, since the first-person camera is not able to accurately determine the real depth and distance to the arms. This results in an approximate position of the upper body (first graph — Fig. 5). When using the full-body CV data, there is also a significant error, although the pose matches to some extent. The third set, based on the arms from the reference (which can be obtained by extracting coordinates from the VR controllers or absolute position sensors), shows that the upper body is reconstructed quite accurately, while the legs are only approximately reconstructed, with a large error. Thus, for all three sets and both models, we can talk only about an approximate reconstruction, which generally corresponds to the results of the total deviation metrics in Figure 3.

Next, an experiment was conducted to train the listed models not on a single frame, but on a sequence of 20 frames. This allowed us to identify some dynamic characteristics and increase the volume of initial information. As a visual comparison showed, since the determining metric was the total deviation, we considered only it (Fig. 6). In general, the use of a frame sequence slightly reduced the total deviation; some models showed even worse results. From a visual point of view (Fig. 7), there was a certain improvement for the LightGBM model, where the reconstruction quality increased significantly, even when reconstructing the body based on arm data (FPV). This also concerned the other two datasets. However, for the neural network model as a whole, no significant improvements were found.

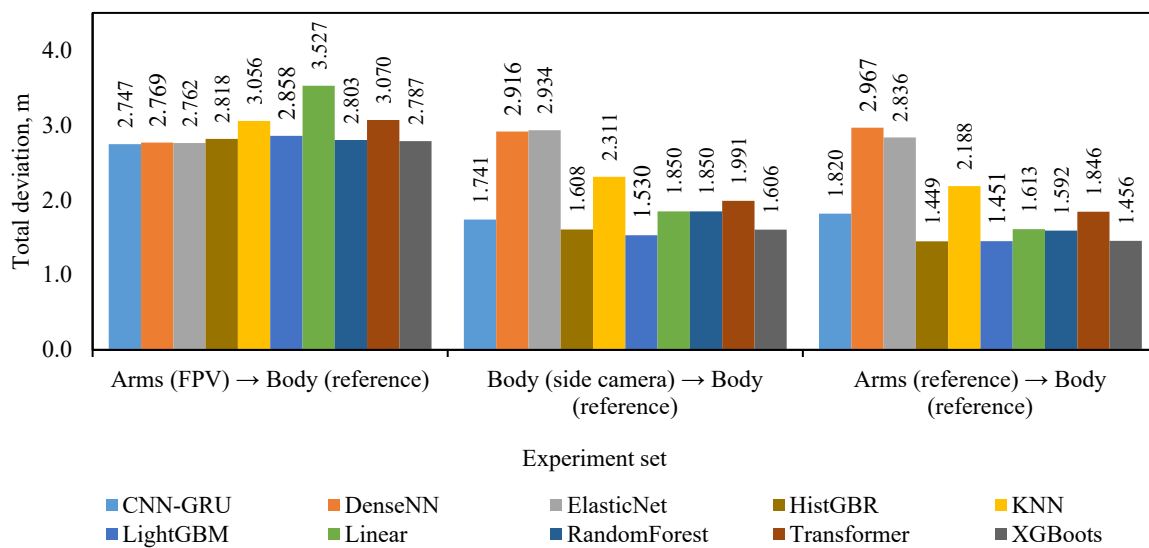


Fig. 6. Comparison of models by total deviation (trained on a sequence of frames)

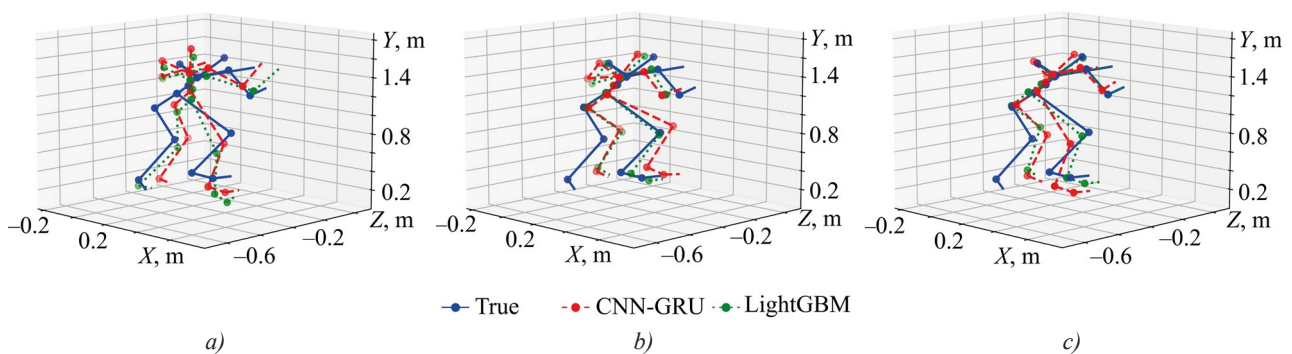


Fig. 7. Visual comparison of CNN-GRU and LightGBM models (trained on a sequence of frames):  
a — on set “Arms (FPV) → Body (reference)”; b — on set “Body (CV) → Body (reference)”;  
c — on set “Arms (reference) → Body (reference)”.

At the end of the experiment, a test was conducted on the implementation of point correction based on the inverse kinematics (IK) model. For this, after predicting body points using machine learning models, the developed IK module was used, which first corrected the end links (hands and feet) using the FABRIK method [19, 20], taking into account the angular limitations of the elbows and knees. Then, the module redistributed the resulting displacements between the pelvis and the thoracic region, automatically aligning the spinal axis. The results of this module are presented in Figure 8.



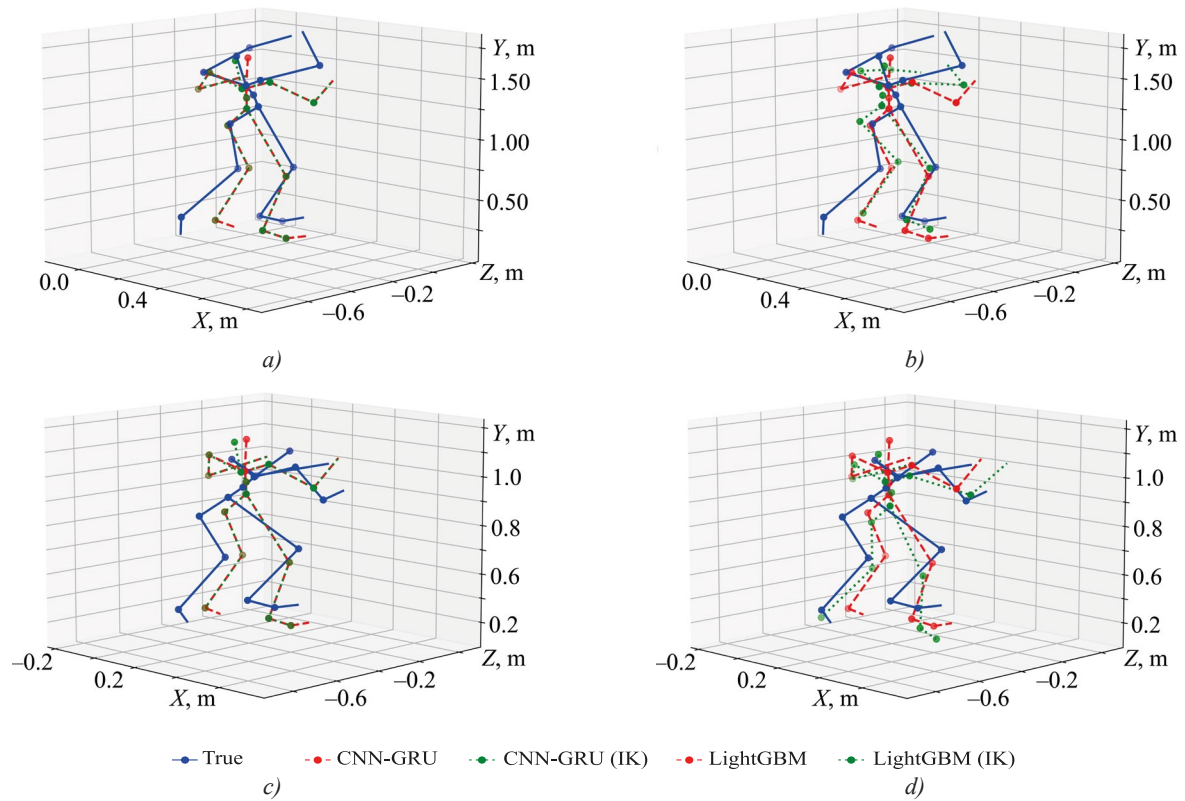


Fig. 8. Visual comparison of models with and without inverse kinematics correction (indicating total deviation before and after correction) on “Arms (FPV) → Body (reference)” dataset: *a* — CNN-GRU (before = 3.511, after = 3.436 m); *b* — LightGBM (before = 3.183, after = 3.112 m); *c* — CNN-GRU on a sequence of frames (before = 2.952, after = 2.991 m); *d* — LightGBM on a sequence of frames (before = 3.261, after = 3.306 m)

The resulting visualizations and numerical evaluations demonstrate that the implementation of the proposed two-pass inverse kinematics generally reduces the total Euclidean deviation of joints from the reference for single frames, but the effect varies depending on the model type and body position. In the first experiment, for the CNN-GRU model, the total deviation decreased from 3.511 to 3.436 meters, and for LightGBM — from 3.183 to 3.112 meters, which corresponded to an improvement of about 2–3%. Graphically, this is manifested in a more natural head alignment and a reduction in the “bends” in the elbows and knees. A different picture is observed in the second experiment, based on 20 frames and a different animation: for CNN-GRU, the error increased from 2.952 to 2.991 meters, and for LightGBM — from 3.261 to 3.306 meters. It is noted that the correction procedure tends to straighten the skeleton, which in this case only worsens the situation. This indicates that geometric constraints applied post factum can improve static anatomical plausibility, but in complex animations, worsen the current pose.

**Discussion.** The conducted research has revealed several patterns. First, reconstructing a full body model from a limited dataset is possible, especially, when the input and output data are from the same source. This was evidenced by the high-quality reconstruction of the body model based on the arm position. However, significant problems were identified in reconstructing the user's leg position, as there was insufficient information about arm movements to predict complex animation. Third, using arm positions from a first-person video stream obtained through computer vision to reconstruct a full body model resulted in high errors due to the lack of distance data to the arms, having only their position relative to the user's eyes. Pre-processing of data simulated in a virtual environment also showed difficulties in recognizing arms in complex animations, which negatively affected the learning process.

When comparing different machine learning architectures for this task, it is worth noting that simpler linear models show good results in predicting the position of body segments, since there are clear dependences between the input and output data that can be approximated by these models. Complex neural network models also solve a similar problem, showing greater flexibility in working with complex input data, but they are not characterized by high performance, and the process of their training is expensive. In a visual comparison, neural network models did not show high efficiency, demonstrating results comparable or even worse.

The experiment shows that the use of a data source with a very limited information value (information about the position of the arms from the computer vision system is just such a source) causes a significant error in solving the regression problem. Firstly, the tracking object often goes out of sight and is not recognized by the model (this is clearly seen in the reduction in

the volume of training data in sets 1 and 2). Secondly, the lack of correct data about the depth, i.e., the distance to the arms, complicates their absolute positioning. In VR systems, this aspect is mitigated by triangulation using data from multiple cameras, but in the simulations conducted, the neural network model for arm recognition did not reflect the correct coordinates along Z axis. A potential solution to the problem and the topic of further research could be obtaining data directly from VR-headsets equipped with integrated cameras. This would expand the training set with natural motion data and provide better quality arm capture in the virtual frame, as the headset's capture system could return metric space coordinates to the digital model, providing a type 5 set ("Arms (reference) → Body (reference)").

When analyzing the topicality of the study within the subject area, it should be compared to existing works. The main difference is the limitation on the use of arm movement data, since a more accurate approach is considered to be the use of at least 5 body points for further reconstruction [21]. This is proved by our previous studies [4], in which the optimal number of points for reconstruction is indicated as at least 5–7, obtained using a reference tracking system.

It is important to note that many VR applications and games implement a tracking system based only on controllers and a headset, followed by reconstruction of a simplified body position using IK algorithms, which allows the arm movements to be extended to the entire body. As the authors [21] emphasize, in such systems, the same readings from sensors located on the arms can correspond to numerous different full poses. This points to the need for additional tuning of inverse kinematics to avoid artifacts, and trained models should select a plausible option. Therefore, the complexity of the task without additional sources of information about the position of the legs or torso remains high. The conducted study highlights this problem, indicating the need to search for and collect additional sources of information to achieve, at a minimum, the mapping of "Arms (reference) → Body (reference)", and ideally — to recognize the entire trajectory of movement, which will help to more accurately predict the position of other body parts. A promising direction here may be the use of not only pre-trained neural networks (e.g., MediaPipe), but also the capture of all information about the surrounding world, which will allow for better segmentation of the user's arms, and perhaps, the torso and legs.

Another limitation of the study is the lack of an assessment of the impact of the training sample size on the quality of the models. In this paper, data from 11 different animation types were collected, two additional types were used for validation, but given the volumes and variability of movements, the set should be significantly larger. However, the study aimed at comparing models within a given task, which demonstrated the ambiguity of their efficiency compared to classic linear models and ensembles. This also indicates the need for further improvement of the model architecture.

Finally, the stage of body model correction based on the kinematic model, implemented through the imposition of anatomical constraints and re-evaluation of the pose, gave ambiguous results — in one pose, it reduced the total deviation, and in another, on the contrary, it increased it. On the other hand, it should be taken into account that the IK module should work with already distorted data on the arms and head in the case of set 1. Therefore, the transition to a higher-quality dataset can reduce the error in the kinematics module.

**Conclusion.** Thus, as a result of the conducted research, an approach to predicting the body model based on a limited set of points was developed, including the stages of data processing, solving regression problems and using the IK module to correct the body model. The corresponding experimental studies were conducted, which showed that LightGBM-type models (among ensembles) and CNN-GRU with an attention mechanism (among neural network models) demonstrated the best results for the selected metrics. The comparison also showed low accuracy of the body model reconstruction when using models (ElasticNet, KNN, DenseNN), which indicated their weak generalization ability. During the visual comparison, contradictions were revealed in the quality of skeleton reconstruction when performing complex animation, since the position of the arms was insufficient to determine the position of the legs and head. In addition, the use of correction based on inverse kinematics is not always justified for complex poses, since the imposition of anatomical constraints and overestimation of the pose can cause additional distortions.

Comparison of the developed models also allows us to draw conclusions about the degree of their applicability: models trained on a first-person data set do not provide reliable reconstruction of the body model, showing a high visual error, which limits their use to only theoretical comparison; while models trained on real arm positions (set 5) show more reliable predictions of body position, which may be in demand in virtual simulators without a sufficient set of sensors. Since the models trained on set 5 work with absolute arm positions, this provides their versatility when selecting a tracking system, because arm position data can be obtained not only using a computer vision system, but also virtual reality controllers or inertial sensors that track arm position.

This study forms several directions for further research within the framework of the body model reconstruction task. The conducted comparative experiments of machine learning models have shown that in order to successfully solve the task, it is required to collect more information about human movements, expand the dataset, and implement more effective learning models with greater generalization ability.

## References

1. Tiboni M, Borboni A, Vêrité F, Bregoli Ch, Amici C. Sensors and Actuation Technologies in Exoskeletons: A Review. *Sensors*. 2022;22(3):884. <https://doi.org/10.3390/s22030884>
2. Vélez-Guerrero MA, Callejas-Cuervo M, Mazzoleni S. Artificial Intelligence-Based Wearable Robotic Exoskeletons for Upper Limb Rehabilitation: A Review. *Sensors*. 2021;21(6):2146. <https://doi.org/10.3390/s21062146>
3. Zihe Zhao, Jiaqi Wang, Shengbo Wang, Rui Wang, Yao Lu, Yan Yuan, et al. Multimodal Sensing in Stroke Motor Rehabilitation. *Advanced Sensor Research*. 2023;2(9):2200055. <https://doi.org/10.1002/adsr.202200055>
4. Obukhov A, Dedov D, Volkov A, Teselkin D. Modeling of Nonlinear Dynamic Processes of Human Movement in Virtual Reality Based on Digital Shadows. *Computation*. 2023;11(5):85. <https://doi.org/10.3390/computation11050085>
5. Kuan Cha, Jinying Wang, Yan Li, Longbin Shen, Zhuoming Chen, Jinyi Long. A Novel Upper-Limb Tracking System in a Virtual Environment for Stroke Rehabilitation. *Journal of NeuroEngineering and Rehabilitation*. 2021;18:166. <https://doi.org/10.1186/s12984-021-00957-6>
6. Jen-Li Chung, Lee-Yeng Ong, Meng-Chew Leow. Comparative Analysis of Skeleton-Based Human Pose Estimation. *Future Internet*. 2022;14(12):380. <https://doi.org/10.3390/fi14120380>
7. Obukhov AD, Dedov DL, Surkova EO, Korobova IL. 3D Human Motion Capture Method Based on Computer Vision. *Advanced Engineering Research (Rostov-on-Don)*. 2023;23(3):317–328. <https://doi.org/10.23947/2687-1653-2023-23-3-317-328>
8. Islam MdM, Nooruddin Sh, Karray F, Muhammad G. Human Activity Recognition Using Tools of Convolutional Neural Networks: A State-of-the-Art Review, Data Sets, Challenges, and Future Prospects. *Computers in Biology and Medicine*. 2022;149:106060. <https://doi.org/10.1016/j.compbiomed.2022.106060>
9. Obukhov A, Dedov D, Volkov A, Rybachok M. Technology for Improving the Accuracy of Predicting the Position and Speed of Human Movement Based on Machine-Learning Models. *Technologies*. 2025;13(3):101. <https://doi.org/10.3390/technologies13030101>
10. Titarenko DYU, Ryzhkova MN. Possible Neural-Network Use for Error Recognition during Physical Exercising. *Radio Engineering and Telecommunications Systems*. 2024;(3):62–72. <https://doi.org/10.24412/2221-2574-2024-3-62-72>
11. Hung Le Viet, Han Le Hoang Ngoc, Khoa Tran Dinh Minh, Son Than Van Hong. A Deep Learning Framework for Gym-Gesture Recognition Using the Combination of Transformer and 3D Pose Estimation. *Cybernetics and Physics*. 2024;13(2):161–167. <https://doi.org/10.35470/2226-4116-2024-13-2-161-167>
12. Jiaxi Jiang, Paul Strelí, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, et al. AvatarPoser: Articulated Full-Body Pose Tracking from Sparse Motion Sensing. In book: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds). *Computer Vision — ECCV 2022*. Cham: Springer; 2022. P. 443–460. [https://doi.org/10.1007/978-3-031-20065-6\\_26](https://doi.org/10.1007/978-3-031-20065-6_26)
13. Jiaxi Jiang, Paul Strelí, Manuel Meier, Christian Holz. EgoPoser: Robust Real-Time Egocentric Pose Estimation from Sparse and Intermittent Observations Everywhere. In book: Leonardi A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds). *Computer Vision — ECCV 2024*. Cham: Springer; 2024. P. 277–294. [https://doi.org/10.1007/978-3-031-72627-9\\_16](https://doi.org/10.1007/978-3-031-72627-9_16)
14. Baradel F, Groueix Th, Weinzaepfel Ph, Brégier R, Kalantidis Y, Roges G. Leveraging MoCap Data for Human Mesh Recovery. In: *Proc. IEEE/CVF Conference on 3D Vision (3DV)*. New York City: IEEE; 2021. P. 586–595. <https://doi.org/10.1109/3DV53792.2021.00068>
15. Seong Hyun Kim, Sunwon Jeong, Sungbum Park, Ju Yong Chang. Camera Motion Agnostic Method for Estimating 3D Human Poses. *Sensors*. 2022;22(20):7975. <https://doi.org/10.3390/s22207975>
16. Kumar S, Srivastava M, Prakash V. Advanced Hybrid Prediction Model: Optimizing LightGBM, XGBoost, Lasso Regression and Random Forest with Bayesian Optimization. *Journal of Theoretical and Applied Information Technology*. 2024;102(9):4103–4115. URL: <https://jaitit.org/volumes/Vol102No9/32Vol102No9.pdf> (дата обращения: 01.06.2025).
17. Nidhi Dua, Shiva Nand Singh, Vijay Bhaskar Semwal. Multi-Input CNN-GRU Based Human Activity Recognition Using Wearable Sensors. *Computing*. 2021;103(7):1461–1478. <https://doi.org/10.1007/s00607-021-00928-8>
18. Vosco N, Shenkler A, Grobman M. Tiled Squeeze-and-Excite: Channel Attention with Local Spatial Context. In: *Proc. IEEE/CVF International Conference on Computer Vision Workshops*. New York City: IEEE; 2021. P. 345–353. <https://doi.org/10.1109/ICCVW54120.2021.00043>
19. Kolpashchikov DYU, Gerget OM, Danilov VV. FABRIK-Based Comparison of the Inverse Kinematic Algorithms Operation Results for Multi-Section Continuum Robots. *BMSTU Journal of Mechanical Engineering*. 2022;753(12):34–45. <https://doi.org/10.18698/0536-1044-2022-12-34-45>
20. Lamb M, Lee S, Billing E, Högberg D, Yang J. Forward and Backward Reaching Inverse Kinematics (FABRIK) Solver for DHM: A Pilot Study. In: *Proc. 7th International Digital Human Modeling Symposium*. 2022;7(1):26. <https://doi.org/10.17077/dhm.31772>
21. Qiang Zeng, Gang Zheng, Qian Liu. DTP: Learning to Estimate Full-Body Pose in Real-Time from Sparse VR Sensor Measurements. *Virtual Reality*. 2024;28(2):116. <https://doi.org/10.1007/s10055-024-01011-1>

**About the Authors:**

**Artem D. Obukhov**, Dr.Sci. (Eng.), Associate Professor of the Department of Automated Decision Support Systems, Tambov State Technical University (112, Michurinskaya Str., Tambov, 392000, Russian Federation), [SPIN-code](#), [ORCID](#), [ResearchGate](#), [ScopusID](#), [ResearcherID](#), [obuhov.art@gmail.com](mailto:obuhov.art@gmail.com)

**Daniil V. Teselkin**, Assistant Professor of the Department of Automated Decision Support Systems, Tambov State Technical University (112, Michurinskaya Str., Tambov, 392000, Russian Federation), [SPIN-code](#), [ORCID](#), [ResearchGate](#), [ScopusID](#), [dteselk@mail.ru](mailto:dteselk@mail.ru)

**Claimed Contributorship:**

**AD Obukhov:** conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft preparation, writing – review & editing.

**DV Teselkin:** data curation, formal analysis, software, validation, visualization, writing – original draft preparation.

**Conflict of Interest Statement:** the authors declare no conflict of interest.

**All authors have read and approved the final version of manuscript.**

**Об авторах:**

**Артём Дмитриевич Обухов**, доктор технических наук, доцент кафедры «Системы автоматизированной поддержки принятия решений» Тамбовского государственного технического университета (392000, Российская Федерация, г. Тамбов, ул. Мичуринская, 112), [SPIN-код](#), [ORCID](#), [ResearchGate](#), [ScopusID](#), [ResearcherID](#), [obuhov.art@gmail.com](mailto:obuhov.art@gmail.com)

**Даниил Вячеславович Теселкин**, ассистент кафедры «Системы автоматизированной поддержки принятия решений» Тамбовского государственного технического университета (392000, Российская Федерация, г. Тамбов, ул. Мичуринская, 112), [SPIN-код](#), [ORCID](#), [ResearchGate](#), [ScopusID](#), [dteselk@mail.ru](mailto:dteselk@mail.ru)

**Заявленный вклад авторов:**

**А.Д. Обухов:** разработка концепции, получение финансирования, проведение исследования, разработка методологии, научное руководство, написание черновика рукописи, написание рукописи, предоставление ресурсов, административное руководство исследовательским проектом.

**Д.В. Теселкин:** курирование данных, формальный анализ, разработка программного обеспечения, валидация результатов, визуализация, написание черновика рукописи.

**Конфликт интересов:** авторы заявляют об отсутствии конфликта интересов.

**Все авторы прочитали и одобрили окончательный вариант рукописи.**

**Received / Поступила в редакцию** 24.06.2025

**Reviewed / Поступила после рецензирования** 20.07.2025

**Accepted / Принята к публикации** 31.07.2025