A skeleton-based method for exercise recognition based on 3D coordinates of human joints

Nataliya Bilous, Oleh Svidin, Iryna Ahekian, Vladyslav Malko
Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine

Article Info

Article history:

Received May 15, 2023
Revised Nov 13, 2023
Accepted Dec 20, 2023

Keywords:

3D coordinates of joint
ARKit
BlazePose
Conjunction of statements
Exercise
Human pose
Movement

ABSTRACT

The aim of the work is to develop the method of identification and comparison of poses and exercises performed by a person that will have a low sensitivity to data errors. This method uses their formal descriptions in the form of conjunctions of logical statements and should work regardless of the shooting angle at which the video was taken and the proportions of the person on it. Each statement describes the position of the joints relative to each other along one of the axes. The joint coordinates are corrected by taking into account the length of the bones that connect them that eliminates the necessity to process outliers and it also improves the accuracy of joints positioning. Removal of errors out of the data using the method of averaging the graph along each axis at every step. In order to do this, consecutive points are grouped so that the difference between the maximum and the minimum does not exceed the error. The groups are then filtered to leave only those in which both are smaller or both are larger. The proposed method of identification requires just a modern smartphone and has no restrictions on how to take video of exercises.

Corresponding Author:

Nataliya Bilous
Department of Software Engineering, Kharkiv National University of Radio Electronics
14 Nauky Avenue, Kharkiv 61166, Ukraine
Email: nataliya.bilous@nure.ua

1. INTRODUCTION

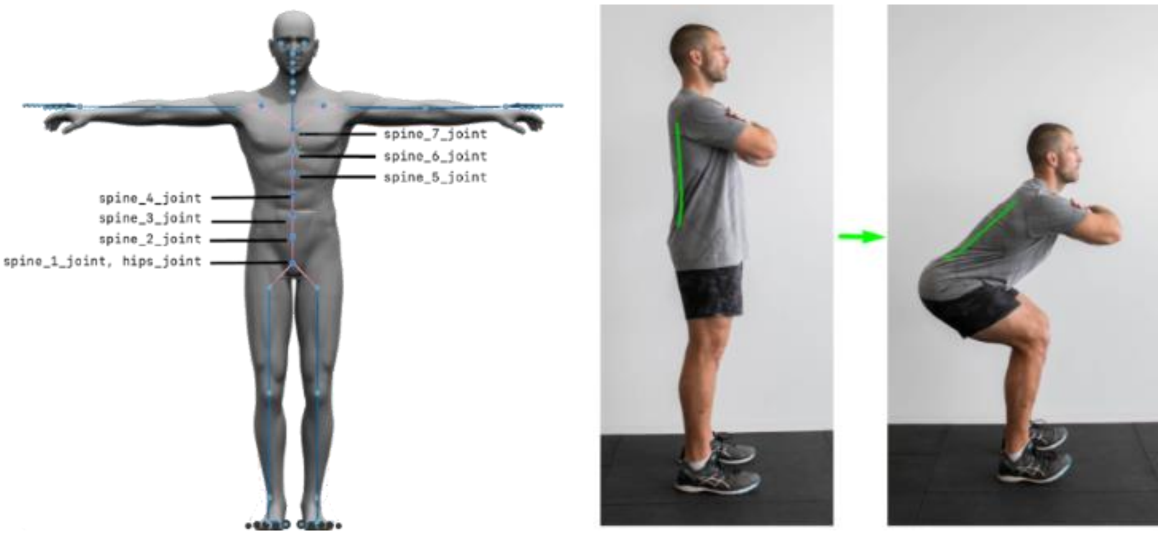
There are practical tasks in medicine, sports and some other fields that require the analysis of human movements. For example, a patient during rehabilitation is not always able to be under the doctor's supervision to check how well the exercises are performed. To solve this problem, it is necessary to digitalize human movements into three-dimensional space. It will help to analyze what poses they consist of and to compare with trainer's poses. There are known solutions for digitalizing human movements. They are divided into 2 types: i) marker type that uses specialized equipment, such as suits with sensors; and ii) markerless type which requires only the external observation through the camera.

Marker systems are not suitable to solve practical problems. Potential users will not buy and carry the necessary equipment. Markerless systems are suitable as they do not require any equipment except a camera. They have their problems, such as significant inaccuracies in the data obtained. Good examples of markerless solutions are OpenPose and ARKit. The first one is an open source product. It can obtain the coordinates of human joints (further-pose) into three-dimensional space synchronizing a video from some cameras and processing it with neural networks. ARKit does almost the same things but requires just one camera. The disadvantage is that it is currently limited to a narrow range of devices.

Almost all markerless technologies produce a list of joints with their coordinates after digitizing human movements. Markerless types of human positioning in space can be used in various fields of activity, for example, in medicine to determine the correctness of physical therapy exercises for people who are rehabilitating after injuries with minimal intervention by the doctor, which facilitates his work and can make patients perform exercises at home, after which the data of the performed exercises will be saved and transferred to the doctor for analysis and further correction of loads if necessary. In robotics, markerless types of positioning can be used to identify a person in space and, in connection with face recognition, can be used for personal identification systems, for example, for robotic assistants (robotic nannies) for children or people with disabilities, for example, when determining which person is who is being observed, has been injured or is in unnatural position or does not show signs of life, notify the person responsible for it or call emergency services, also this approach to implementation can be used as a car assistant when a position analysis is performed and can help to automatically notify emergency services in the event of a road accident or stop the car is in autopilot mode if a person shows signs of uncontrollable condition (falling asleep at the wheel or driving in an unnatural position).

A schematic model of the joints location can be seen on the Figure 1(a) where joints are marked with blue dots, and for the spine joints their names were added. The particular attention should be paid to the joint lying in the pelvic area. It is considered to be the root joint and all the other joints are positioned relative to it. Different technologies produce different number of joints, so OpenPose produces the information about 24 joints and ARKit does it for 91 joints. Of course, there are alternatives where a three-dimensional model of a human body is produced. However, it is difficult to use for analysis due to the lack of an easy way to obtain the position of body key points relative to each other. ARKit was selected as the data provider for analysis due to all the above reasons.

After obtaining the data on the position of the joints, it is necessary to developed a method that will allow you to identify the positions in which the body was during the recording. Movement is always a change of position. It can be called a sequence of poses. A certain sequence of poses is an exercise. Therefore, to identify poses is enough to identify exercises. For example, a squat exercise consists of two poses as shown in Figure 1(b).



(a) (b)

Figure 1. The schematic model of the joints location, (a) produced by ARKit and (b) key poses during squats

The object of the research is the process of identification of poses and exercises performed by a person and recorded on camera. The subject of the research is the poses and exercises identification and comparison method, based on the data obtained using a markerless motion capture system ARKit motion tracking (MT). The aim of the work is to develop the method of the identification of poses and exercises performed by a person.

It is necessary to compare the coordinates of the joints to solve the problem of poses identification. Not all the joints are involved in movement and are important for its identification and it should be taken into account. Existing works already consider various approaches to the search for the most important joints, the movement of which is important to describe the action. Ofli *et al.* [1] identified methods for searching the most informative joints on the basis of joint angle, rotational velocity of the joint.

Batabyal *et al.* [2] observed that joints that are not directly involved in the sequence of actions (for example, the ankle joint in the case of a clap) have a consistently low frequency of toothed noises. It has been observed that interarticular variations in terms of individual coordinates (for example, the change in x-coordinate of one joint relative to the change in z-coordinate of another joint during a movement) are more important than variations in joint position as a whole (for example, left knee position relative to right knee position). This fact is ignored in previous works such as Ofli *et al.* [1] or Yang and Tian [3]. The usage of the embedding of a sparse covariance matrix in a linear collector of low dimensionality was proposed to solve the noise problem in [4], [5], and the usage of support vector machine (SVM) was proposed to classify actions. Most previous approaches use the positions of 3D point clouds, representations of common angles or their variants reduced by principal components analytics (PCA) [6]–[9].

The usage of Euclidean distances was discussed in detail by Liu and Müller [10], [11]. They conducted experiments and compared the same human movements shot from different angles. Their experiments showed that it was possible to achieve good results only if both videos were shot from a similar angle, and the movements on them differed only in speed. They used approaches for processing motion information which were previously used only for sound processing [12], [13]. One of the problems of such approaches is their sensitivity to postural deformations, which can occur in logically related movements, as well as their sensitivity to inaccuracies in the data. To identify the exercises in the video, it is important to understand how many poses they consist of and when these poses begin and end. The proposed algorithms are described in [14], [15], they allow to search for local extremes of the motion graph.

The methods described above are for the two-dimensional data. A lot of research on how to achieve maximum results with the limitations of two-dimensional space can be seen in existing works. For example, there is a work where a method for searching the direction of vision is proposed. It allowed to understand in which projection to the camera the person was shot [16]. There is no need in it while using a data provider that delivers results in 3D. In this case, the first step is not to find a way to adjust the proportions of the distorted projection of the camera, but to find approaches to the correct interpretation of the information about the z coordinate. Cong *et al.* [17] proposed a method that exploits the natural geometric constraints of the point cloud for self-observation and uses 2D keypoints in images for weak observation with a large data set. Such a method, called 3D multi-person pose estimation (MPE), demonstrates the advantage and ability to generalize data. The method first detects people and then estimates the 3D pose for each person according to the cropped image and the point cloud. The whole process of the method contains two important components, including the spotlight image module. The first component combines information from two different data modalities to take full advantage of the 3D point cloud geometry and image features, and another component uses temporal cues present in sequence data to improve pose accuracy by learning dynamic human motion rules and matching pose space signs of large dimensions [17]. Having analyzed skeleton models in human activity recognition studies, Mohottala *et al.* [18] conducts an indepth analysis of child action recognition on graph convolutional networks (GCN). When performing the pose-based GCN analysis, it was extended by comparing red-green-blue (RGB) modality and skeleton modality in GCN. As a result, reliability values by classes were obtained, which show that the unlimited nature of the video strongly limits the pose estimation process. Therefore, the result was obtained that the spatial-temporal graph convolutional networks (ST-GCN) model is able to achieve higher performance than the long-term recurrent convolutional network model. Despite differences in pose between adults and children, these results suggest that when actions are movement-oriented, the skeleton modality can perform on par with the RGB modality. This opens up future research directions in creating an optimal skeleton graph that improves pose estimation in complex scenarios and multi-modality set development [18].

For modeling plausible 2D human poses, the GFPose framework is proposed for various applications. At the core of GFPose is a time-dependent estimation network that estimates the gradient at each body joint and gradually damps the perturbed 3D human pose according to the task specification. After conducting experiments, H. Ci empirically proved that GFPose as a multi-hypothesis pose estimator outperforms the existing state-of-the-art (SOTA) dataset by 20% on the Human 3.6 M dataset and can produce reasonable and realistic samples for denoising, completion and pose generation tasks. Although GFPose shows great potential in many applications, the backsampling process requires repeated model inferences. The total inference time is proportional to the number of sampling steps, which limits the use of very large deep networks in real-time scenarios [19]. For 3D pose assessment of several people based on unlabeled data, Criado *et al.* [20] used a hybrid system for pose assessment is proposed, which includes a three-stage conveyor system containing the following components: i) skeleton detector, ii) a graph neural network (GNN) corresponding to several types of frames, and iii) multilayer perceptron pose estimation.

The reliability of the proposed pose assessment system for several people was experimentally investigated. The system was tested using the Carnegie Mellon University Panoptic Studio dataset and showed an average joint accuracy error of 29.79 mm. The result of the test can be considered less than acceptable for many applications. Furthermore, the recall/accuracy ratio is very close to unity, demonstrating the robustness of the mapping framework model. To solve the problem of low efficiency of traditional methods of assessing human posture, the general graph optimization (G2O) position based on graph optimization is proposed [21]. This method provides real-time performance using the following algorithms: i) three-dimensional reconstruction of bone proportions based on two-dimensional key points, ii) three-dimensional classification of human orientation based on weighted two-dimensional characteristics of joints, and iii) inverse special correction based on heuristic search and inverse joint suppression algorithms based on the angles of rotation of human joints.

The method is less accurate than traditional methods, but the speed is much higher. We researched and analyzed the methods of recognizing human emotions to determine the state of his health [22]. One of the simplest methods is the method based on the classification of key points (landmarks of the face), the coordinates of which can be obtained using the following algorithms: Point distribution model (PDM), comparing machine learning (CML) algorithm, active appearance model (AAM), deep learning partial differential equation model (DPM) or convolutional neural network (CNN). It is also possible to use a method that consists in randomly transferring a sequence of frames taken from a video with a certain step to the 3D-CNN. Neural networks such as CNN use three-degree-of-freedom convolutions that transform four-dimensional maps into three-dimensional feature maps. We compared two methods of pain detection: detection of pain based on the presence of strong negative emotions and using a dataset with images of people divided into the classes "painful" - "non-painful" for neural network training. It is experimentally found that the specialized dataset "PAB-F" performs better despite the relatively small number of images, and "fer2013" has a high false positive rate.

For the task of real-time pose estimation, Liu *et al.* [23] proposed a light dynamic model of a huge convolution. The model improves the problems of reducing narrow-channel information destruction by using the ReLU activation function, expanding the effective recessive field of the model, and increasing the model complexity without excessively increasing the computational cost. Kuhnke and Ostermann [24] proposes relative pose consistency and a semi-supervised learning strategy for consistency-based head pose estimation. Evaluation of the approaches used in the domain adaptation scenario and across datasets shows that this approach outperforms SOTA. But ultimately there are gaps in methods that are trained on real datasets similar to the target domain.

A model that can minimize the negative log-likelihood loss is proposed when investigating the problem of miscalibration in 3D pose estimation of 3D pose with multiple hypotheses. The model shows that unlike existing methods, it can learn a well-calibrated posterior distribution with little loss of overall accuracy [25]. Analyzing self-monitoring methods for human pose estimation, Zhang *et al.* [26] proposes a self-monitoring approach that learns directly from large-scale videos of real 6D category-level poses in the wild. The framework developed and proposed in the paper reconstructs the canonical 3D shape of an object category and learns the exact correspondence between input images and the canonical shape using surface embedding. For training, new geometric losses of coherence of cycles are proposed, which create cycles in 2D and 3D space in different cases and time steps. In experiments, their self-supervised approach can perform at par with or even better than current methods for category-level 6D object pose estimation and keypoint transfer tasks [26]. A new self-monitoring method was also proposed to estimate human monocular 3D pose from unlabeled multiangle images without camera calibration [27]. For this purpose, multi-image coherence was used to separate 2D estimates into canonical predictions, i.e., 3D pose and camera rotation, which were used to refine the errors of the 2D estimates and re-project the 3D pose onto the 2D self-determining neural network training.

Various human pose estimations (HPE) have been analyzed in areas such as activity recognition, animation and gaming, virtual reality, video tracking, and present an analytical study of deep learning techniques for both 2D and 3D HPE domains [28]. Several deep learning-based datasets have been used to study and estimate human pose, such as max planck institute for informatics (MPII); common microsoft objects in context (COCO); frames marked in cinema (FLIC); Human3.6M, leeds sports pose (LSP), and 3D wild poses (3DPW).

Research by Bilous *et al.* [29], we explored the task of determining human body positions in video streams, a key problem in the field of computer vision. Knowing that using such specific hardware as motion capture systems is costly, the challenge is to evaluate the effectiveness of using existing computer vision libraries for body pose detection without additional hardware and with reduced computational requirements. A set of libraries was analyzed, including OpenPose, PoseNet, and BlazePose, for their ability to recognize and track body parts and movements in real-time video. Our evaluation focused on performance, accuracy, and computational efficiency. They also examined various pose comparison algorithms to determine their speed and accuracy in pose determination. The results show that BlazePose combined with the weighted distance method provides superior performance in pose recognition, demonstrating high accuracy and robustness in different scenarios. The F1 index indicated that this combination proved to be powerful and effective for the task at hand [29]. Research by Nguyen *et al.* [30] focuses on improving skeleton-based human action recognition methods. These methods have gained popularity due to their compactness and robustness against appearance variations. They discuss the double-feature double-motion network (DD-Net), a lightweight structure that excels in speed, and performance for hand and body actions in skeleton-based recognition. However, DD-Net struggles with actions weakly connected to global trajectories. To eliminate this restriction, the authors introduce the triple-feature double-motion network (TD-Net), an enhanced version of DD-Net. TD-Net incorporates a new branch using normalized coordinates of joints (NCJ) to enrich spatial information. This approach has shown superior performance over the DD-Net and other SOTA methods across multiple datasets, including MSR-Action3D, CMDFall, JHMDB, FPHAB, and NTU RGB+D. Additionally, the authors deployed an application on an edge device to demonstrate real-time human action recognition capabilities. This application can process up to 40 frames per second for pose estimation using MediaPipe and recognizes actions from skeleton sequences in just 0.04 milliseconds [30].

Qin *et al.* [31] presents a new approach in the area of action recognition using skeleton sequences. These sequences are considered as ideal for action recognition on edge devices due to their lightweight and compact nature. Traditional methods in the area of skeleton-sequence recognition use graph neural networks to extract spatiotemporal signals from 3D joint coordinates, focusing mainly on first- and second-order features (joint and bone representations), which has achieved high accuracy. However, these models struggled with actions having similar motion trajectories. To eliminate this restriction, they introduced a new concept of fusing higher-order features in the form of angular encoding into the GNN architectures. This angular encoding captures the relative movements between body parts while maintaining invariance against different human body sizes, allowing for more precise action recognition. This method showed significant improvements in distinguishing actions with similar motion trajectories. Integrating angle coding into existing action recognition architectures, such as ST-GCN and decoupling GCN, allowed the model to achieve high accuracy in two major benchmarks, including NTU60 and NTU120. Additionally, this method required fewer parameters and reduced runtime, enhancing its efficiency. The research demonstrates the effectiveness of incorporating angular features into ST-GCN for skeleton-based action recognition. This innovation not only improves the accuracy of recognizing complex actions but also supports real-time action recognition on edge devices due to its efficient processing [31]. Duan *et al.* [32] presented a new approach to action recognition based on skeleton sequences. This approach, called PoseConv3D, uses volumetric heat maps instead of graph sequences as the basic representation of human skeletons. Unlike methods based on GCNs, PoseConv3D is more efficient at learning spatio-temporal features, more robust to noise in pose estimation, and better generalisable across different datasets. PoseConv3D is also able to handle scenarios with multiple faces without additional computational costs. This approach achieved top results on five of the six standard skeleton-based action recognition benchmarks and all eight multimodal action recognition benchmarks [32].

Feng and Meunier [33] provide an in-depth review of the field of human action recognition using skeleton graphs and GNNs. They acknowledge that the human skeleton as a compact representation of human action is gaining increasing attention in areas such as video surveillance and human-computer interaction, and significantly improves performance in these areas. The researchers analysed previous works that used GCNs to process skeleton sequences, but found that GCN-based methods have limitations in terms of robustness, interaction between different modalities, and scalability. This has led to the development of alternative approaches, such as PoseConv3D, which uses 3D heat maps instead of graph sequences as the underlying representation of human skeletons. They also note that skeleton GNN-based methods are important because of their ability to focus on the action and compactness. Since skeleton sequences only include pose information, they are immune to external contextual influences such as changes in lighting or background [33]. Vishwakarma and Jain [34] focused on developing a method that uses skeletal pose data from depth sensors to create a so-called "motion polygon". This technique aims to improve the accuracy and efficiency of human action recognition systems. He uses special algorithms to process skeletal pose data, which enables efficient recognition and classification of different types of movements and actions [34].

2. PROBLEM STATEMENT

The problem is to find the videos digitalized using ARKit and exercises and to evaluate their correctness. To assess the correctness, it is necessary to have the sample regarding which the exercises will be compared. This sample will be the exercises performed by the trainer.

Human movements can be described as the vector of poses $\overline{M} = \{m_1, m_2, \dots, m_n\}$, where n is the number of shots in the video. Each pose m_i is described by the matrix where the columns are the coordinates along one of the axes, and the rows are the coordinates of a particular joint. In total, it has 91 rows, which corresponds to the number of joints that can track ARKit.

$$m_i = \begin{pmatrix} x_{hips} & y_{hips} & z_{hips} \\ x_{shoulder} & y_{shoulder} & z_{shoulder} \\ \vdots & \vdots & \vdots \\ x_{wrist} & y_{wrist} & z_{wrist} \end{pmatrix} \quad (1)$$

The movement of the trainer is also a vector of poses identical in structure to the other movements, but each pose in it is considered to be correct, let's denote it as $\overline{E}$. A set of key poses L that is sufficient to describe the movements must be found for a person who is not a trainer.

$$f_1 : \overline{M} \to L, L \subset M \quad (2)$$

The poses of a person who repeats the exercises are not absolutely correct in comparison with the poses of the trainer. Therefore, among all the key poses it is necessary to find the right poses CL. In case when all poses are performed correctly, the set CL can be identical to the set L.

$$f_2 : L \to CL, CL \subseteq L \quad (3)$$

A set of key poses is also chosen for the trainer in the set $\overline{E}$. As the trainer performs all the poses correctly, let's denote it as L_{ref}. It is necessary to find the percentage of the correct poses Q that a person performs in relation to the poses of the trainer L_{ref}.

$$Q = \frac{|L|}{|L_{ref}|} \quad (4)$$

The main requirements for the method are the insensitivity to the data inaccuracies and the ability to compensate for the disparity between the speed of the compared videos. If the requirements are not met, the system cannot be considered to be of high quality. In this case, it needs to be improved or new approaches to solving the problem should be found. Quality assurance measures are essential to ensure that the system meets the desired standards. Continuous monitoring and evaluation are crucial components of maintaining a high-quality system. Moreover, seeking feedback from users and stakeholders can provide valuable insights for identifying areas of improvement. It's important to consider both technical and user-oriented perspectives when striving for system excellence.

3. METHOD

Information about the pose of the human body is returned in the form of many joints. This information may contain material errors and should be processed in order to correct inaccuracies for the further analysis. Therefore, the human body must be represented as a hierarchical structure as shown in Figure 2(a), which connects all the joints together. Then the distance between the connected joints can be estimated. Defining the distance between the joints as the length of the vector, we can correct the position of the more mobile joints relative to the less mobile ones. Therefore, a collinear vector of the required length should be found as shown in Figure 2(b).

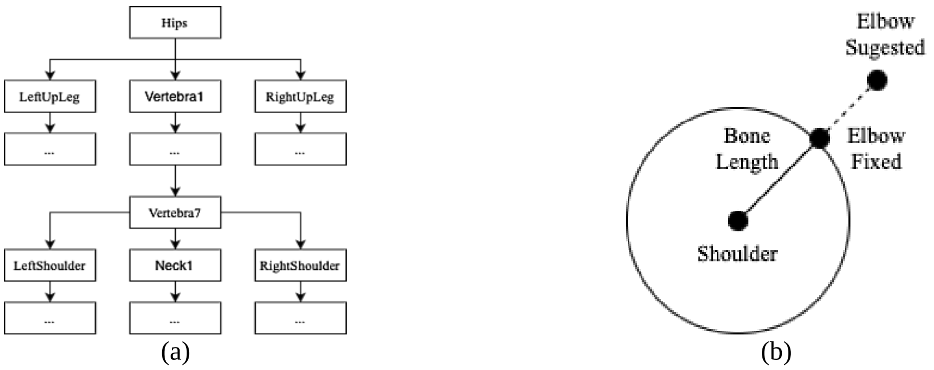
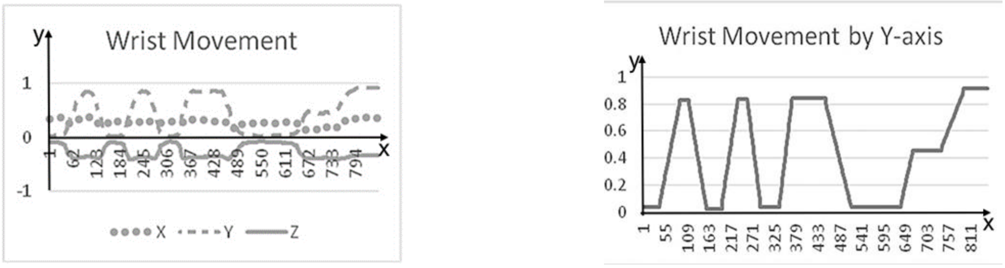


Figure 2. Structure of joints, correction of vector (a) tree diagram of joints and (b) correction of the position of a joint

Every variant of coordinates in the next shot of the video is a new pose m_i, but most of these poses are not of interest because they are not decisive for movement $\overline{M}$. The key poses of motion L are the local extremes of motion, but they are difficult to find because, although the outliers are filtered, the diagram still has a toothed appearance as shown in Figure 3(a). These teeth appeared due to inaccuracies in skeleton recognition. It may be different for different data providers, for ARKit it does not exceed 0.05 m. To solve this problem, it was decided to use an algorithm for averaging values with a given step, which is equal to the error of the data providers. The averaging algorithm is used separately for motion along each axis. Its aim is to go through all the points along each of the axes and combine successive points into groups. The difference between the maximum and minimum within a group should not exceed the error value. After dividing the graph into groups, they are filtered. The following groups are selected: groups where the movement of the graph changed its direction to the opposite and groups where their length exceeds the specified value l.

The second condition is necessary if we want to keep information about long pauses that happened while performing the movement. If the value is too small, we will leave many intermediate values, and only the same records will be compared. The value of the other groups is reduced to their average value as shown in Figure 3(b). On this graph, groups are straight lines parallel to the x-axis. To search the poses, the proposed method of identification works in the following way: each pose can be formally described as the relative position of the joints.



(a) (b)

Figure 3. The wrist movement in three directions: (a) the x-axis is the distance in meters and the y-axis is the shot number and (b) the results of the preparation of wrist movement data for further analysis

For example, the right hand is above the right shoulder and the left hand is below the left shoulder. This method cannot show exactly the closeness of two poses to each other, but it allows us to claim that these are the same known poses. It is very resistant to data errors. A pose, for example, squats T_{squat} can be described as a conjunction of statements.

$$T_{squat} = f(p_i) = (x_{wrist} > x_{shoulder}) \wedge (y_{knee} > y_{hips}) \wedge (y_{hips} > y_{foot}) \wedge \cdots \wedge (z_{hips} > z_{knee}) \quad (5)$$

We can get a set of correctly performed CL poses that are present in motion by processing the key poses with this function. Then the video with a trainer can be processed and the set of poses obtained from it L_{ref} can be compared with the one that was found before. Comparing these values, we can calculate what percentage of Q poses corresponded to the poses of the trainer.

4. EXPERIMENTS

For the experiments, it is suggested to use two advanced technologies for tracking human movement: ARKit for iOS devices and BlazePose for desktop systems. The choice of these technologies is dictated not only by their SOTA functionality and accuracy, but also by their wide accessibility for consumers and researchers. ARKit, developed by Apple, is part of the iOS infrastructure and uses cameras, accelerometers, and gyroscopes built into iPhones to accurately track human movements in three dimensions. This technology allows us to collect detailed information about the location and movement of joints, which is critical to our research. On the other side, BlazePose, developed by Google, is an advanced human pose tracking system focused on desktop applications. It uses advanced machine learning and computer vision algorithms to determine a person's position in space, providing high accuracy and reliability in recognising complex movements. This makes BlazePose an ideal technology for desktop systems. The use of these technologies opens up great opportunities for analysing human movements in a variety of scenarios. ARKit, for example, allows us to use mobile devices to collect data in different physical spaces without the need for specialised hardware. BlazePose provides the opportunity for more detailed analysis in a controlled laboratory environment on desktop systems, where more powerful computing resources and sophisticated data processing algorithms can be used.

The mobile devices used were the iPhone XR, iPhone 12, and iPhone 12 pro, each running iOS 14.2. The choice of these devices is due to their high-performance cameras, built-in accelerometers, and gyroscopes, which are an integral part of ARKit technology. While ARKit is an advanced combination of computer vision and machine learning, BlazePose, used on desktop systems, is based on deep learning algorithms to accurately track human poses. To implement BlazePose using a desktop computer based on Intel Core i5-9300 processor, 16 GB RAM and NVIDIA GeForce GTX 1050 graphics card.

For experimental research, selected a number of datasets that include a variety of human motion videos, including UTKinect-Action3D [35], SBU-Kinect-Interaction v2.0 [36], Florence3D [37], JHMDB [38], and NTU RGB+D 120 [39]. These datasets were selected for their representativeness and diversity in the context of human activity and motion, which are key to validating and verifying the effectiveness of tracking algorithms across different platforms. Each of these datasets represents a unique dataset, allowing to hold a comprehensive analysis and evaluation of our tracking algorithm. In addition to using these datasets, an in-house dataset was also collected using ARKit and BlazePose, allowing for a comparison of different motion tracking approaches and a deeper and more comprehensive analysis.

The model testing protocol included recording three basic exercises: squats, push-ups, and forward bends. Each exercise was recorded in several variations, reflecting different styles of performance - from accurate to completely incorrect. Exercises were also recorded at different angles relative to the participant's face, increasing the angle by 45 degrees for each new recording to assess how perspective affects movement tracking accuracy. Each session began and ended with the participant assuming a standard posture - standing upright with arms at their sides. This starting and ending posture serves as a calibration reference, ensuring consistency between recordings and facilitating a more accurate assessment of movement accuracy. Each recording was repeated ten times to provide a deep and reliable data set for further analysis.

Strict error thresholds for the system were established, defining a maximum tolerance of 0.05 m based on empirical tests as an upper error bound for the ARKit system. An important parameter for analysis was also the detection of significant pauses in movement that exceeded 24 frames at a recording rate of 24 frames per second, as this played a key role in studying the smoothness and rhythm of the exercises. This methodical approach to data collection and analysis allowed us not only to evaluate the accuracy and reliability of ARKit in various use cases, but also to determine the potential of using consumer technology for professional-level motion analysis. This methodology provided us with the opportunity to conduct a comprehensive comparison of two advanced motion tracking technologies in different environments.

5. RESULTS AND DISCUSSION

Initially, a significant amount of data from datasets such as UTKinect-Action3D, SBU-Kinect-Interaction v2.0, Florence3D, JHMDB, and NTU RGB+D 120 was used. These datasets provided us with a large number of videos and images covering a wide range of human actions and activities. From the NTU RGB+D 120, 3,000 of the more than 114480 available videos were selected, including various types of exercises and movements. This data was used to evaluate the ability of our algorithms to recognize and classify different human movements and poses. In addition, an in-house dataset was collected using ARKit technology on different iPhone models. Three basic exercises (squats, push-ups, and inclines) performed at different angles and speeds were recorded to collect data that accurately reflects real-life human movements. In total, about 500 unique videos were collected, each of which was carefully analyzed.

The analysis of ARKit data showed that this system is characterized by high motion tracking accuracy even in complex scenarios with minimal errors. BlazePose, used on desktop systems, also demonstrated high tracking accuracy. The results indicate the high potential of both motion tracking systems for a variety of applications including fitness, rehabilitation, health monitoring, and for use in complex software systems for monitoring human posture and position. The research shows that ARKit and BlazePose have their unique advantages and limitations. Here is a comparative analysis of both systems Table 1.

ARKit demonstrates high tracking accuracy and high adaptability to different scenarios, while BlazePose demonstrates slightly lower accuracy and adaptability, but still gives good results, especially in stationary environments. ARKit proved particularly useful in mobile scenarios with variable perspectives, while BlazePose demonstrated its strengths in stationary environments. Each exercise consisted of two poses. The exercise was repeated 10 times in total on each record. The ideal result of the algorithm will be the recognition of 20 poses. After checking the method on video, where all the exercises were performed correctly at different angles, the following values for each angle were obtained with ARKit as shown in Table 2 and BlazePose as shown in Table 3.

Table 1. Comparison of ARKit and BlazePose

Criterion	ARKit (iOS)	BlazePose (Desktop)
Tracking accuracy	95%	92%
Error detection	5%	8%
Adaptability to Scenarios	High	Middle

Table 2. Percentage of poses that were recognized on the record at different angles using ARKit

Angle\Exercise	Squat (%)	Push-up (%)	Bend (%)
0	90	85	95
45	100	95	100
90	100	100	100
135	95	100	100
180	90	85	90

Table 3. Percentage of poses that were recognized on the record at different angles using BlazePose

Angle\Exercise	Squat (%)	Push-up (%)	Bend (%)
0	92	83	92
45	99	96	98
90	100	99	99
135	94	98	97
180	88	84	92

Then there is a check for false positive and false negative results. For these 4 videos are recorded: exercises performed correctly; exercises are performed very quickly; exercises performed partially incorrectly (not all posture requirements were met); and exercises were not performed, but there were movements that are similar to the necessary exercises. The number of poses found is shown in Table 4 for ARKit and in Table 5 for BlazePose. The measurement of the similarity between the four videos, where the same exercises were performed was done in the next experiment. The first video was taken as a sample shot by the trainer, and the others were comparable to him. The result is in Table 6 using ARKit and Table 7 using BlazePose. During the following trials the same exercises were identified, but using the other existing identification methods for ARKit as shown inTable 8 and BlazePose as shown in Table 9. Videos shot at an angle of 90 degrees to the human face were compared in all experiments, except the first one for ARKit as shown in Table 10 and BlazePose as shown in Table 11.

Table 4. Percentage of poses that were recognized while performing exercises with different levels of correctness. ARKit

	Squat (%)	Push-up (%)	Bend (%)
Correct	100	100	100
Accelerated	95	90	100
partially correct	0	5	0
Incorrect	0	0	0

Table 5. Percentage of poses that were recognized while performing exercises with different levels of correctness. BlazePose

	Squat (%)	Push-up (%)	Bend (%)
Correct	99	99	98
Accelerated	96	92	98
partially correct	1	8	0
Incorrect	0	0	0

Table 6. The comparison of the similarity of exercises. ARKit

	Squat (%)	Push-up (%)	Bend (%)
1	85	90	75
2	85	75	85
3	65	85	80

Table 7. The comparison of the similarity of exercises. BlazePose

	Squat (%)	Push-up (%)	Bend (%)
1	84	92	76
2	86	74	84
3	67	84	82

Table 8. The comparison of the similarity of exercises, using the method of comparing angles. ARKit

	Squat (%)	Push-up (%)	Bend (%)
1	32	90	86
2	45	78	81
3	48	86	78

Table 9. The comparison of the similarity of exercises, using the method of comparing angles. BlazePose

	Squat (%)	Push-up (%)	Bend (%)
1	31	89	85
2	46	80	79
3	46	88	75

Table 10. The comparison of the similarity of exercises, using the method of intervals. ARKit

	Squat (%)	Push-up (%)	Bend (%)
1	40	82	84
2	55	72	82
3	44	73	79

Table 11. The comparison of the similarity of exercises, using the method of intervals. BlazePose

	Squat (%)	Push-up (%)	Bend (%)
1	38	80	86
2	57	74	80
3	45	75	81

The result of the tests, it was proved that to obtain the most accurate result you need to shoot in the planes where the most significant changes occur. The reason for it is the following: it is difficult for the data provider to estimate the depth of the image and sometimes it returns completely incorrect data that cannot be corrected even using the methods built into the method. However, the method as a whole demonstrates the high resistance to inaccuracies. It is caused the results of the first test, which shows that the best results were achieved when the shooting angle was 90 degrees. From this point of view, almost all changes occur without the involvement of the z-axis. It is necessary to choose the correct limit of inaccuracy of the provided data before applying the method of identification. It is also necessary to choose the length values for intervals which are not extremes and should not be ignored. In case when these options are chosen incorrectly, you can get two extreme cases: too many intermediate poses are found or too many key poses are missed.

In addition, this method is quite good at determining the similarity of movements in the video as a whole and in each movement separately. Thus, the method recognizes more than 90% of the poses even when the movements speed changes significantly. The disadvantage of a high percentage of recognition is the presence of false-positive results. They occur due to the fact that a person can perform completely different movements and fall into extremely similar poses. The probability value of such coincidence is very small as too many factors must come together for this. Compared to the other identification methods, the developed method has almost the same results when the data has a minimum number of errors and recognizes approximately by 30% more poses in cases where the data is saturated with errors. As we can see from the tests of the application of our method for fitness, it has proven itself positively, and therefore, in the future, the method will be tested in other areas, such as monitoring the driver's condition using a camera in the cabin, as well as in the role of a video nanny for monitoring human positions. We believe that the used method can prove itself quite well for these areas, and can also be used as one of the components in a composite software system for monitoring the condition and position of a person in connection with other methods.

6. CONCLUSIONS

The developed identification method shows sufficient results for practical tasks, such as helping patients in the recovery process or people engaged in fitness. Given that this method was launched on the basis of data obtained from a telephone camera, this opens up broad prospects for its use. The advantages of the method are resistance to inaccuracies in the data, as well as the fact that it sets the position in a human-understandable form, it allows the additional analysis. The division into poses allows to search special exercises according to their key points, as well as look for movement patterns, for example, to find a rep of the same movement in the video. On the other hand, the method has some disadvantages. The pose will not be detected if it is performed a little bit incorrectly. To solve the problem, it makes sense to try reducing the requirements for the pose gradually until it is identified. Another disadvantage of the method is that it can be easily deliberately deceived. For example, it will not be able to distinguish circular and diamond-shaped movements of the hands, because it focuses on changing the vectors of motion. However, this is not important given the area of use. One of the main problems observed during the experiments is the incorrect results caused by incorrect values of the z coordinate. Although the method performed better by 30% than others, it is still an important cause of error. Incorrect depth value is difficult to correct but it is likely that in the future, due to the wider use of leaders in phones, inaccuracies can be avoided.

ACKNOWLEDGEMENTS

The work was carried out within the framework of the scientific directions of the Department of Software Engineering and the research laboratory "Information Technologies in Learning and Computer Vision Systems" of the Kharkiv National University of Radio Electronics with the support of scientists from the Technical University of Applied Sciences Wildau, Brandenburg, Germany and the Volkswagen Foundation (Contract № Az 9C042).

REFERENCES

[1] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition,” *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 24–38, 2014, doi: 10.1016/j.jvcir.2013.04.007.

[2] T. Batabyal, T. Chattopadhyay, and D. P. Mukherjee, “Action recognition using joint coordinates of 3D skeleton data,” *Proceedings - International Conference on Image Processing, ICIP*, vol. 2015, pp. 4107–4111, 2015, doi: 10.1109/ICIP.2015.7351578.

[3] X. Yang and Y. L. Tian, “EigenJoints-based action recognition using Naïve-Bayes-Nearest- neighbor,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 14–19, 2012, doi: 10.1109/CVPRW.2012.6239232.

[4] C. Y. Chiu, S. P. Chao, M. Y. Wu, S. N. Yang, and H. C. Lin, “Content-based retrieval for human motion data,” *Journal of Visual*

Communication and Image Representation, vol. 15, no. 3, pp. 446–466, 2004, doi: 10.1016/j.jvcir.2004.04.004.

[5] L. Kovar and M. Gleicher, "Automated extraction and parameterization of motions in large data sets," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 559–568, 2004, doi: 10.1145/1015706.1015760.

[6] E. Keogh, T. Palpanas, V. B. Zordan, D. Gunopulos, and M. Cardle, "Indexing large human-motion databases," *Proceedings 2004 VLDB Conference: The 30th International Conference on Very Large Databases (VLDB)*, pp. 780–791, 2004, doi: 10.1016/B978-012088469-8.50069-3.

[7] Y. Sakamoto, S. Kuriyama, and T. Kaneko, "Motion map: image-based retrieval and segmentation of motion data," *Computer Animation 2004-ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 259–266, 2004, doi: 10.1145/1028523.1028557.

[8] K. Forbes and E. Fiume, "An efficient search algorithm for motion data using weighted PCA," *Computer Animation, Conference Proceedings*, pp. 67–76, 2005, doi: 10.1145/1073368.1073377.

[9] E. Hsu, K. Pulli, and J. Popović, "Style translation for human motion," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 1082–1089, 2005, doi: 10.1145/1073204.1073315.

[10] G. Liu, J. Zhang, W. Wang, and L. McMillan, "A system for analyzing and indexing human-motion databases," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 924–926, 2005, doi: 10.1145/1066157.1066290.

[11] M. Müller, T. Röder, and M. Clausen, "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 677–685, 2005, doi: 10.1145/1186822.1073247.

[12] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," *Deep Learning Approach for Natural Language Processing, Speech, and Computer Vision*, pp. 99–125, 2023, doi: 10.1201/9781003348689-5.

[13] L. Kovar and M. Gleicher, "Flexible automatic motion blending with registration curves," *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2003*, pp. 214–224, 2003.

[14] G. Shcherbakova, V. Krylov, O. Logvinov, and N. Bilous, "Adjustement of wavelet function parameters for analysis of non-stationary periodic signals with multistart optimization," *2017 4th International Scientific-Practical Conference Problems of Infocommunications Science and Technology, PIC S and T 2017-Proceedings*, vol. 2018, pp. 110–112, 2017, doi: 10.1109/INFOCOMMST.2017.8246361.

[15] O. Hramm, N. Bilous, and I. Ahekian, "Configurable cell segmentation solution using hough circles transform and watershed algorithm," *Proceedings of the International Conference on Advanced Optoelectronics and Lasers, CAOL*, vol. 2019, pp. 602–605, 2019, doi: 10.1109/CAOL46282.2019.9019493.

[16] A. O. Rakova and N. V. Bilous, "Reference points method for human head movements tracking," *Radio Electronics, Computer Science, Control*, no. 3, pp. 121–128, 2020, doi: 10.15588/1607-3274-2020-3-11.

[17] P. Cong *et al.*, "Weakly supervised 3D multi-person pose estimation for large-scale scenes based on monocular camera and single LiDAR," *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, vol. 37, pp. 461–469, 2023, doi: 10.1609/aaai.v37i1.25120.

[18] S. Mohottala, S. Abeygunawardana, P. Samarasinghe, D. Kasthurirathna, and C. Abhayaratne, "2D Pose estimation based child action recognition," *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2022, 2022, doi: 10.1109/TENCON55691.2022.9977799.

[19] H. Ci *et al.*, "GFPose: Learning 3D human pose prior with gradient fields," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 4800–4810, 2023, doi: 10.1109/CVPR52729.2023.00465.

[20] D. R. -Criado, P. Bachiller, G. Vogiatzis, and L. J. Manso, "Multi-person 3D pose estimation from unlabelled data," *arXiv-Computer Science*, pp. 1-9, 2022.

[21] H. Sun, Y. Zhang, Y. Zheng, J. Luo, and Z. Pan, "G2O-pose: Real-time monocular 3D human pose estimation based on general graph optimization," *Sensors*, vol. 22, no. 21, pp. 1-22, 2022, doi: 10.3390/s22218335.

[22] N. V. Bilous, O. V. Rassokha, I. A. Ahekian, and O. V. Gramm, "Research on methods for development of software system for emotions recognition and state of human health determination (in Ukranian: Дослідження методів для розробки програмної системи розпізнавання емоцій та визначення стану здоров'я людини)," *Bionics of intelligence*, vol. 94, pp. 65–70, doi: 10.30837/ bi.2020.1(94).10.

[23] Z. Liu, S. Liu, Z. Liu, H. Wang, and Q. Jin, "Lightweight dynamic large convolution model for real-time human pose estimation," in *Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)*, 2022, pp. 523-527, doi: 10.1117/12.2662876.

[24] F. Kuhnke and J. Ostermann, "Domain adaptation for head pose estimation using relative pose consistency," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 3, pp. 348–359, 2023, doi: 10.1109/TBIOM.2023.3237039.

[25] P. A. Pierzchlewicz, R. J. Cotton, M. Bashiri, and F. H. Sinz, "Multi-hypothesis 3D human pose estimation metrics favor miscalibrated distributions," *arxiv- Computer Science*, pp. 1-16, 2022.

[26] K. Zhang, Y. Fu, S. Borse, H. Cai, F. Porikli, and X. Wang, "Self-supervised geometric correspondence for category-level 6D object pose estimation in the wild," in *International Conference on Learning Representations*, pp. 1-19, 2022.

[27] H. W. Kim, G. H. Lee, M. S. Oh, and S. W. Lee, "Cross-view self-fusion for self-supervised 3D human pose estimation in the wild," in *Computer Vision – ACCV 2022*, Cham: Springer, pp. 193–210, 2023, doi: 10.1007/978-3-031-26319-4_12.

[28] P. Kumar, S. Chauhan, and L. K. Awasthi, "Human pose estimation using deep learning: review, methodologies, progress and future research directions," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 4, pp. 489–521, 2022, doi: 10.1007/s13735-022-00261-6.

[29] N. V. Bilous, I. A. Ahekian, and V. V. Kaluhin, "Determination and comparison methods of body positions on stream video," *Radio Electronics, Computer Science, Control*, no. 2, pp. 52-60, 2023, doi: 10.15588/1607-3274-2023-2-6.

[30] T. T. Nguyen, D. T. Pham, H. Vu, and T. L. Le, "A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, vol. 16, no. 8, pp. 709–726, 2022, doi: 10.1049/cvi2.12119.

[31] Z. Qin *et al.*, "Fusing higher-order features in graph neural networks for skeleton-based action recognition," in *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-15, 2022, doi: 10.1109/TNNLS.2022.3201518.

[32] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 2959–2968, 2022, doi: 10.1109/CVPR52688.2022.00298.

[33] M. Feng and J. Meunier, "Skeleton graph-neural-network-based human action recognition: a survey," *Sensors*, vol. 22, no. 6, pp. 1-52, 2022, doi: 10.3390/s22062091.

[34] D. K. Vishwakarma and K. Jain, "Three-dimensional human activity recognition by forming a movement polygon using posture skeletal data from depth sensor," *ETRI Journal*, vol. 44, no. 2, pp. 286–299, 2022, doi: 10.4218/etrij.2020-0101.

[35] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–27, 2012, doi: 10.1109/CVPRW.2012.6239233.

[36] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–35, 2012, doi: 10.1109/CVPRW.2012.6239234.

[37] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 479–485, 2013, doi: 10.1109/CVPRW.2013.77.

[38] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3192–3199, 2013, doi: 10.1109/ICCV.2013.396.

[39] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale Benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020, doi: 10.1109/TPAMI.2019.2916873.

BIOGRAPHIES OF AUTHORS

Nataliya Bilous is a professor at Department of Software Engineering of the Kharkiv National University of Radio Electronics, Ph.D. In addition, she is serving as the Director of Outsourcing Scientific Training and Production Center (OSTPC) (2013–present) and the head of the Research Laboratory of Information Technologies in Learning and Computer Vision Systems (ITLCVS) (2005–present). She is laureate of the State Prize in the field of Science and Technology (2008). Now she is also researcher of Technical University of Applied Sciences, Wildau, Brandenburg, Germany. She has published more than 300 various research papers in international journals and conferences, author of 12 textbooks and scientific monographs, and 14 intellectual property certificates. She is co-organizer of ICCUBITO (2013), IEEE ITIB (2015) conferences. Her research interests are in computer vision systems, medical and diagnostic automated systems, artificial intelligence, pattern recognition, and body position. She can be contacted at email: bilous.n.v@gmail.com or nataliya.bilous@nure.ua.

Oleh Svidin is a master in Department of Software Engineering of the Kharkiv National University of Radio Electronics. His research interests are in artificial intelligence, computer vision, and body positioning. Currently, he is a senior developer on flutter and xamarin in computer vision and artificial intelligence. He can be contacted at email: oleh.svidin@nure.ua.

Iryna Ahekian is a senior lecturer at Department of Software Engineering of the Kharkiv National University of Radio Electronics. In addition, she is serving as the researcher of the laboratory "Information Technologies in Learning and Computer Vision Systems". She holds a Master of Software Engineering, a Bachelor of Accounting and Audit, Master of Taxation. She has experience in computer vision projects as researcher and developer of software. She is published more than 40 various research papers in international journals and conferences, author of 4 textbooks, and 4 intellectual property certificates. Her research interests are in computer vision systems, artificial intelligence, pattern recognition, body position, and learning technologies. She can be contacted at email: iryna.ahekian@nure.ua.

Vladyslav Malko is a Ph.D. student at Department of Software Engineering in Kharkiv National University of Radio Electronics. He holds a Master of Software Engineering, specializes in data mining, computer vision, and the development and protection of web-based applications of any complexity. In addition, he is serving as an assistant at Department of Software Engineering. His research interests are artificial intelligence, computer vision, pattern recognition, and body position in 3D. He can be contacted at email: vladyslav.malko@nure.ua.