Exploring motion capture algorithms in computer vision using intel depth camera

M.D.Ollaberganova, T.A. Xudaybergenov

The analysis of existing approaches to tracking the human body revealed the presence of problems when capturing movements in a three-dimensional coordinate system. The promise of motion capture systems based on computer vision is noted. Existing research on markerless motion capture systems only considers positioning in 2D space. Therefore, the goal of the study was to improve the accuracy of determining the coordinates of the human body in three-dimensional coordinates by developing a motion capture method based on computer vision and triangulation algorithms.

Keywords: motion capture, virtual reality, triangulation, computer vision, machine learning

Introduction. Significant progress has now been made in the field of computer vision. Technologies have been developed that allow solving the problems of detecting objects, determining their state, geometric assessment of the space depicted in the frame, and many others. Thanks to this, computer vision has become widespread in various fields of human activity, from healthcare and education to the entertainment sector. A fairly promising direction is the use of computer vision technologies for three-dimensional reconstruction and positioning of various objects, including people. There are quite a large number of systems for determining the absolute position of a person in space, which can be divided into the following categories:

□ systems that use inertial sensors and make it possible to determine the magnitude of their movement, as well as changes in angles between them, which involves the use of gyroscopes and accelerometers [1]. A well-known representative of this category is Intel Depth [2], which includes up to 32 inertial sensors;

□ laser positional tracking systems, based on the use of base stations installed on opposite sides of the room and emitting infrared rays, which make it possible to accurately determine the position and orientation of sensors in space. An example of such systems are Intel Depth virtual reality kits from HTC [3], which have an error of up to 0.1 mm;

□ systems using magnetic sensors [4], based on the use of a magnetic field to capture human movement, which involve the presence of wearable sensors on the user’s body. Intel Depth falls into this category.

- portable electromagnetic motion tracking system, considered one of the fastest (sampling frequency 240 Hz);

□ optical systems based on markers - determine the position of objects using markers using a set of cameras. An example is Intel Depth, which has a fairly low error: the average absolute marker tracking errors are 0.15 mm in static tests and 0.2 mm (with corresponding angular errors of 0.3°) in dynamic tests [5];

□ markerless optical systems based on the use of computer vision and machine learning. Examples of such technologies are OpenPose, MediaPipe, Intel Depth. With their help, human movements can be tracked with an accuracy of up to 30 mm [6].

Analyzing the listed categories of motion capture systems, we can conclude that most solutions used to recognize human actions and movements involve the presence of various wearable devices, such as sensors or gloves. The bulk of these devices are cumbersome due to the large number of sensors and the need for a wired connection. Some such systems have high accuracy, but cannot be used due to their size or the presence of electromagnetic interference [7]. Inertial systems have a number of problems associated with error accumulation, which limits their use to relative positioning in space only.

Therefore, optical systems for recognizing and tracking user actions are very popular. To obtain information about the user's actions and position, frames obtained from the camera are used. Among optical systems, it is worth noting those that use markers (the user may be dressed in special clothing or certain marks are attached to him), which makes their use in real conditions difficult and is more applicable to specially prepared premises (for example, film studios).

Systems that do not use any markers allow users to interact more freely with the environment and are more suitable for real-world applications. Significant disadvantages of systems in this area include relatively low accuracy, unreliability and low performance. This may be largely due to the shortcomings of computer vision algorithms used to recognize a person in a frame, as well as the following reasons: variability in a person’s appearance and lighting conditions, partial occlusions due to layering of objects in the scene, and the complexity of the human skeletal structure.

The operation of markerless motion capture systems is usually based on an algorithm for estimating human pose. Approaches to solving the problem of human pose estimation can be divided into top-down and bottom-up. In top-down approaches, people are first detected in the frame, then the pose of each detected person is estimated. Algorithms that belong to bottom-up approaches, at the first stage, search for body parts in the frame, then group them into poses. As a rule, convolutional neural networks are used for this task, such as YOLO (You Look Only Once) [8], SSD (Single Shot Detection) [9], R-CNN (Region CNN) [10] and others. They allow you to recognize many different objects, including a person or individual parts of the body, with high accuracy. However, one of the disadvantages of the solutions listed above is their low performance and slow operation. To solve this problem, there are special frameworks Intel Depth [11], MediaPipe [12], OpenPose [13]), which also use neural networks optimized for real-time operation.

It should be noted that the above algorithms, technologies and approaches of markerless motion capture systems allow positioning in two-dimensional space, which makes it difficult both to determine the distance to objects and their sizes, and to track complex movements when, for example, the user’s hands are hidden by his body. Existing stereo camera solutions can be effective, but are not very accurate when the subject is far away from the camera, which is what happens when tracking a person's entire body. In addition, they do not solve the problem of occlusions. Thus, a current research direction is the development of a motion capture method using multiple cameras and computer vision technologies. When implementing multi-camera motion capture systems, the problem of combining objects from several images inevitably arises, i.e. the need to perform triangulation. Among the triangulation methods, linear and iterative linear algorithms can be distinguished.

Linear triangulation is the most common approach to performing the reconstruction of objects in three-dimensional space, including methods such as linear eigenmethod, linear least squares method, direct linear transformation, differing in varying degrees of noise resistance [14].

Iterative linear methods are a more robust version of linear triangulation. Conventional linear methods may be less accurate when solving problems of triangulation of a set of points, since when solving systems, the minimized error has no geometric meaning (it does not take into account the shape of the skeleton and the rules for connecting points). The basic idea of iterative linear methods is to adaptively change the weights of linear equations such that the weighted equations correspond to the errors. Iterative linear methods include L2 and L∞ triangulation [15].

Thus, within the framework of this research, the following task is set: it is necessary to develop a method for capturing human movements that allows positioning the user’s body in three-dimensional coordinates with minimal error and using computer vision technologies. The proposed method can be used either as a replacement for existing motion capture systems or as part of other algorithms, for example, for subsequent classification of the human condition. The goal of this work is to improve the accuracy of determining the poses and coordinates of the human body in three-dimensional coordinates by developing motion capture methods based on computer vision. To achieve this goal, it is necessary to formalize the main stages of the process of capturing points of the human body from several cameras, integrate triangulation algorithms, choosing among them the optimal one in terms of accuracy, and implement a software implementation of the proposed method.

Materials and methods. Solving the problem of three-dimensional positioning of a person in space includes the following main stages:

□ preliminary calibration of a set of cameras;

□ implementation of procedures for detecting a person in a frame and calculating skeletal points;

□ calculation of three-dimensional reconstruction of a human body model.

Let's look at them in more detail.

The calibration process involves the camera system taking several pictures of a calibration template from which key points with their known relative positions in space can be easily identified. Afterwards, internal and external parameters are calculated for each camera. Internal parameters are constant for a specific camera, external parameters depend on the location of the cameras relative to each other [16]. Therefore, this step must be completed before using the camera system for the first time in a given location.

To calculate the coordinate values of a point in three-dimensional space, it is necessary to know the coordinates of its projections on images and the projective matrices of cameras [10]. The projective matrix P of some camera can be represented as a combination of matrices A (containing internal parameters of the camera) and R (rotation), as well as a displacement vector T, which describe the change in coordinates from the world coordinate system to the coordinate system relative to the camera:

$$P = A[R \mid T] = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix},$$

where (,) x y are the coordinates of the projection of a three-dimensional point on the image in pixels; (,) x y c c — coordinates of the camera’s central point; (,) x y f f — focal length in pixels.

At the second stage, it is necessary to directly obtain key (skeletal) points of the human body on each of their cameras. To extract skeletal body points from a frame, it is possible to use various machine learning technologies, for example, Intel Depth, MediaPipe, OpenPose and others [8]. As part of this study, it is proposed to use the highly efficient and productive Pose module from the MediaPipe library. MediaPipe Pose uses machine learning to provide highly accurate human body pose tracking, 3D landmark detection, and full-body background segmentation masks from RGB video frames. This approach allows you to track up to 33 points and provides real-time operation on most modern devices.

At the third stage, the positions of key skeletal points in three-dimensional space are calculated. To obtain data on the position of human skeletal points in space, triangulation is performed - finding the coordinates of a three-dimensional point from the coordinates of its projections. Triangulation is one of the most important tasks in computer vision; its solution is a decisive step in 3D reconstruction and affects the accuracy of the entire result [9].

The three-dimensional reconstruction of object points based on the position values of point projections on images from all cameras is based on epipolar geometry. Its main idea is that 3D points in the scene are projected onto lines in the image plane of each camera - epipolar lines. These lines correspond to the intersection of the image plane with the plane passing through the centers of the cameras and the 3D point. This idea provides a condition for finding pairs of corresponding points in two images: if it is known that a point x on the plane of the first image corresponds to a point x ' on the plane of another image, then its projection must lie on the corresponding epipolar line.

Since X is a homogeneous representation of coordinates in three-dimensional space, to calculate them it is necessary to obtain i x and Pi for at least two cameras. To solve the system of equations (7), 4 algorithms were considered [4]:

□ direct linear transfer (DLT);

□ linear least squares method;

□ L2 triangulation; - optimal (polynomial) method.

DLT refers to linear triangulation algorithms, the main advantage of which is the simplicity of its implementation. For example, in the Intel Depth computer vision library there is a ready-made implementation of this algorithm in the triangulatePoints method.

The linear least squares method also refers to linear ones and consists in the fact that the system of homogeneous equations (7) is reduced to a system consisting of inhomogeneous equations, for solving which the least squares method is used.

When using a two-camera system, to minimize error (9), the following sequence of actions must be performed:

□ parameterize the bundle of epipolar lines in the first image using the parameter t.

Thus the epipolar line in the first image can be expressed as 0 □ ( )t ;

□ using the fundamental matrix F, calculate the corresponding epipolar line 1 □ ( )t in the second image;

□ express the distance function (9) as a function of t;

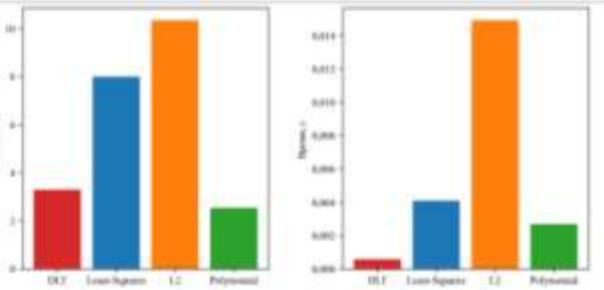
□ search for the value of t at which (9) tends to a minimum.

Using elementary calculus methods, we can reduce the solution of the minimization problem to finding the roots of a sixth-order polynomial. The estimated spatial point is calculated using the Direct Linear Transfer (DLT) method [7].

Research results. The solution to the optimization problem (11) is carried out by triangulating two-dimensional object points obtained from images of several cameras, in the framework of this study - from two cameras using various algorithms listed in the previous section.

The listed triangulation methods were implemented using the Intel Depth and NumPy libraries. For comparison, the algorithms were integrated into software that implements a 3D motion capture method. An example of the method for reconstructing the entire human skeleton is shown in Fig. 1.

Then, these algorithms were compared by the value of the reprojection error function (11) for all skeletal points from two images. A comparison of the selected triangulation methods was carried out in terms of the magnitude of the error, as well as in terms of the time to obtain a solution (computational complexity) for the entire set of skeletal points. Summary comparison diagrams are presented in Fig. 2.



For the selected triangulation methods, a series of experimental tests were also carried out, during which, for each approach, the calculated lengths of the user's limbs and the absolute deviation of the obtained values from the real ones were measured. The comparison is presented in Table 1.

Table 1

Body segment	DLT	Least-Squares	L2	Polynomial	Реальное значение
Forearm	25,2 ± 1,6	30,8 ± 0,2	26,6 ± 0,5	24,3 ± 0,4	26
Shin	42,2 ± 2,0	65,3 ± 1,1	44,6 ± 0,7	38,7 ± 1,8	41
Hip	45,7 ± 2,7	59,5 ± 0,49	48,7 ± 1,3	44,1 ± 0,6	45
Average deviation	2,43	14,58	2,26	1,67	0
Presented are the average values (in centimeters) after a sample of 10 measurements ± standard deviation in the sample					

The developed software includes the following modules:

□ for working with input devices (cameras);

□ to perform calibration and obtain basic camera parameters;

□ for synchronizing several cameras;

 for object recognition (user’s body and hands);

- to analyze the location of the found skeletal points;

 to build visualization in real time

When implementing the software, the Python programming language, Intel Depth and Matplotlib libraries were used. The system operates in several threads: one is responsible for receiving data from cameras, the second is for visualization, and the third is for sending the received world coordinates of the human body to external systems or modules. The use of a unified protocol with a data package in JSON format allows you to integrate the software into third-party systems (for example, game development environments Unity, Unreal Engine, etc.) [2].

Discussion and conclusion. Let us analyze the results of comparing triangulation algorithms based on selected metrics, presented in Fig. 2 and in table 1.

During the comparison, it was found that the optimal algorithm for 3D reconstruction is the polynomial method. The error value is about 2.55 pixels. In real tests, when determining a person's height, the error was no more than 3%, taking into account the fact that MediaPipe Pose does not fix the top point of the head and it is calculated approximately based on the position of the eyes. When measuring the limbs, the error ranged from 0.9 cm to 2.3 cm, the average was 1.67 (Table 1). Thus, real tests confirm the correctness of the choice of the polynomial method.

Next, we compare the results obtained with existing studies, for example, those described in [2]. The authors also use trained networks (OpenPose) to implement a markerless human recognition system, a camera calibration procedure, and skeletal point extraction, but place the cameras next to each other to simulate stereo vision. This key difference allows this study to recognize human postures where some parts of the body overlap others. In addition, using MediaPipe Pose allows you to track 33 skeletal points, rather than 18 as in the Intel Depth-based method. The obtained error values generally correspond to existing studies (the best result in [2] is 2 cm), which allows us to conclude that the proposed approach can be used in practice. Other markerless systems, for example, based on Kinect [3], also show comparable results in terms of measurement error (2–5 cm). Thus, the resulting solution generally corresponds in accuracy to existing developments.

Comparison of point set calculation time shown in Fig. 2 on the right shows that the DLT algorithm provides the best performance, however, all algorithms show acceptable results (to ensure performance of 30 and even 60 frames per second). Therefore, this metric is not decisive.

The developed software can be used in various subject areas, primarily as a replacement for motion capture systems based on inertial sensors. The advantages of the proposed solution are low economic costs of implementation and availability (transition from highly specialized motion capture suits to common camera-based tools), the possibility of parallel capture of body models of several users [4].

The scientific novelty of the research lies in an integrated approach to formalizing the process of three-dimensional positioning of a person using computer vision technologies, including preliminary calibration of a set of several cameras, formalization of procedures for detecting a person in a frame using an arbitrary neural network to obtain skeletal points, as well as calculation of a three-dimensional reconstruction of a body model human using various triangulation algorithms. The study includes all the necessary calculation formulas and detailed steps to achieve the goal - increasing the accuracy of determining the poses and coordinates of the human body in three-dimensional coordinates using computer vision technologies. The presented theoretical results are quite universal and can be used for the practical implementation of motion capture systems based on various neural network models, not just MediaPipe Pose.

Bibliography

1. Lind C.M., Abtahi F., Forsman M. Wearable Motion Capture Devices for the Prevention of Work-Related Musculoskeletal Disorders in Ergonomics – An Overview of Current Applications, Challenges, and Future Opportunities. Sensors. 2023;23(9):4259.

2. Sers R., Forrester S., Moss E., Ward S., Ma J., Zecca M. Validity of the Perception Neuron Inertial Motion Capture System for Upper Body Motion Analysis. Measurement. 2020;149:107024.

3. Bauer P., Lienhart W., Jost S. Accuracy Investigation of the Pose Determination of a VR System. Sensors. 2021;21(5):1622.

4. Irshad M.T., Nisar M.A., Gouverneur P., Rapp M., Grzegorzek M. AI Approaches towards Prechtl's Assessment of General Movements: A Systematic Literature Review. Sensors. 2020;20(18):5321.

5. Merriaux P., Dupuis Y., Boutteau R., Vasseur P., Savatier X. A Study of Vicon System Positioning Performance. Sensors. 2017;17(7):1591.

6. Nakano N., Sakura T., Ueda K., Omura L., Kimura A., Iino Y., et al. Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose with Multiple Video Cameras. Frontiers in Sports and Active Living. 2020;2:50.

7. Coronado E., Fukuda K., Ramirez-Alpizar I.G., Yamanobe N., Venture G., Harada K. Assembly Action Understanding from Fine-Grained Hand Motions, a Multi-camera and Deep Learning Approach. In: Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). New York, NY: IEEE; 2021. P. 2628–2634.

8. Tausif Diwan, Anirudh G., Tembhurne J.V. Object Detection Using YOLO: Challenges, Architectural Successors, Datasets and Applications. Multimedia Tools and Applications. 2023;82(6):9243–9275.

9. Wei Liu, Anguelov D., Erhan D., Szegedy C., Reed S., Cheng-Yang Fu., et al. SSD: Single Shot MultiBox Detector. In book: Leibe B., Matas J., Sebe N., Welling M. (eds). Computer Vision – ECCV 2016. Cham: Springer. 2016;9905:21– 37.

10. Bharati P., Pramanik A. Deep Learning Techniques—R-CNN to Mask R-CNN: A Survey. In book: Das A., Nayak J., Naik B., Pati S., Pelusi D. (eds). Computational Intelligence in Pattern Recognition. New York, NY: Springer. 2020;999:657–668.

11. Bajpai R., Joshi D. MoveNet: A Deep Neural Network for Joint Profile Prediction across Variable Walking Speeds and Slopes. IEEE Transactions on Instrumentation and Measurement. 2021;70:1–11.

12. Ghanbari S., Ashtyani Z.P., Masouleh M.T. User Identification Based on Hand Geometrical Biometrics Using Media-Pipe. In: Proc. 30th International Conference on Electrical Engineering (ICEE). New York, NY: IEEE; 2022. P. 373–378.

13. Weijian Mai, Fengjie Wu, Ziqian Guo, Yuhan Xiang, Gensheng Liu, Xiaobin Chen. A Fall Detection Alert System Based on Lightweight Openpose and Spatial-Temporal Graph Convolution Network. Journal of Physics: Conference Series. 2021;2035:012036.

14. Szeliski R. Recognition. In book: Computer Vision: Algorithms and Applications. London: Springer; 2011. P. 575–640.