Human Pose Estimation applying ANN while RGB-D Cameras video handling

Iakov Korovin[0000-0001-5192-8835], Donat Ivanov [0000-0002-2221-0231]

Southern Federal University, 2 Chehova st., 3479328, Taganrog, Russia
donat.ivanov@gmail.com

Abstract. In the paper we try to solve the problem of online human pose estimation. All over the world security systems of railway stations, airports and other facilities have control points. Control points are equipped with cameras and other devices for automated data acquisition. One of the tasks within security providing is to recognize the position of body parts of the human, passing the control point. In this paper, a method, based on the use of artificial neural networks. The features of the practical implementation of the proposed approach are described.

Keywords: Human Pose Estimation, Body Parts Recognition, Pose Recognition; Video Recognition.

1 Introduction

The task of pose estimation is relevant when solving a large number of practical problems. In the paper In the paper we describe a human pose estimation in the safety systems of railway stations, airports and other facilities.

Usually each visitor passes a control point. At the control point, a video camera is installed. One of the tasks of video analysis is to recognize the position of parts of the body of the visitor. In solving this problem, it is necessary to determine the position of the parts of the human body passing the control point, based on data from existing cameras.

The assessment of a human posture is an urgent task [1–7] due to different applications, such as motion capture, manipulation of objects in virtual environments, augmented reality, remote control of robotics means, etc. Evaluation of a person's posture and its change in time also allows to determine the actions performed by a person, or his behavior, which is an important task in the construction of video surveillance and security systems.

The process of assessing a person's posture is related to the search for posture parameters of a human body model that is most suitable for observations on one or more input images. While marker-based systems are already available in many commercial applications, marker-free posture assessment is still a complex research topic.

There are many algorithms that solve this problem with high accuracy from several input images [8, 9] or even one photograph [10, 11]. Often, such systems require

manual initialization and cannot process camera images in real time. Promising methods for assessing human posture in real time use a three-dimensional body model [12–14], time-of-flight depth cameras [13], or RGB-D cameras such as Microsoft Kinect, Intel RealSense, Asus Xonar, etc. [15].

In terms of the type of data used to construct the human skeleton, it can be divided into methods based on the use of only color (RGB) [16–20] and methods combining color and depth data (RGB-D) [21–25]. The definition of human actions can be based either on specified criteria for each type of action, or deep learning algorithms can be used to solve this problem. Regardless of the type of input data (RGB or RGB-D), the main purpose of these methods is to determine either the pose of the person or the actions performed by him.

Methods using only information about the color of the scene may have limited effectiveness due to factors such as camera movement, occlusion, the complex structure of the scene, or changing or low illumination of the scene.

Depth data are stable with respect to changes in the shooting conditions and background, and it makes it fairly easy to segment objects by depth. The use of depth sensors allows you to get a reliable assessment of a person's posture in real time, that is, they demonstrate high recognition accuracy and low time complexity, therefore they are quite popular in works on recognition of human actions. These methods are popular in studies of recognition of human actions [22–24, 26, 27]. However, at present, the accuracy and cost of depth sensors can significantly limit the areas and conditions of use of such systems.

There are three types of depth cameras that are commonly used in computer vision tasks: triangulation (using a pair of RGB cameras or a stereo camera), a time-of-flight camera (TOF, Time-of-Flight camera), and a camera based on structured backlighting.

Depth sensors based on structured illumination and TOF are subject to significant errors and low accuracy of measuring depth when working outdoors due to external illumination of the scene.

Stereo cameras have a lower cost and can give a higher accuracy of measurement, however, the calculation of depth requires more time and computational resources, and such systems cannot be used in poor or low-light scenes.

There is also the possibility of using laser scanners, however, these devices are very expensive and unsuitable for working in real time.

Therefore, when solving the problem of constructing a "skeleton" in order to determine the human posture in real time, the best option is to use RGB-D cameras. In this case, additional markers are not required, which cannot be installed on each person passing the control point. The limiting factors for processing only color streams or depth streams can be offset by a parallel stream. If the scene is poorly lit, you can use the more computationally expensive features for working with scene depth. If it is impossible to obtain a depth map using IR illumination, it is possible to use a pair of RGB-images of a stereo camera to obtain the depth of the scene.

From the approaches described above it follows that to solve the problem of constructing a human "skeleton" the most optimal solution is to use data on the color and depth of the scene. The use of existing methods based on these data streams is inef-

fective due to the presence of a priori information about the scene being processed (the passage of a control point by a person), since by adding optimizations for a specific task, one can achieve a significant increase in the speed and accuracy of the construction of the "skeleton" human and determination of marker poses for detecting non-standard behavior. For a quick preliminary search for "nodal" points of the skeleton, you can use a neural network, the coordinates and connection of which are already drowned on the depth map.

2 The Proposed Method for Human Pose Estimation from a Video Image

The principle of the proposed method is as follows: using an RGB-D camera, an RGB image and a depth map of the scene are captured. At the first stage, a camera with IR illumination is used, which allows you to quickly and with sufficient accuracy to get a depth map. However, there is no problem switching to using a stereo camera, since the method relies on processing the finished depth map, and where it was received from (ToF camera or stereo camera) does not matter.

After receiving the color image, it goes through the preprocessing process, and a correspondence is established between the pixels in the RGB image and the depth map. A depth map is necessary to obtain information about the three-dimensional coordinates of each part of the human body and the structure of the skeleton as a whole, and that is why the RGB-D camera is used in the method.

With the help of a trained neural network, a color image is used to search for the main parts of the human body (head, neck, trunk, shoulders, elbows, hands, knees, feet), and the correctness of the found parts is preliminary checked based on the conditions described above.

The centers of the elements found are projected onto a depth map, according to which the coordinates of the parts of the body are further refined, a "skeleton" is built, the correctness of its structure is checked, the resulting coordinates of each part are obtained, and a three-dimensional graph of the "skeleton" is constructed.

Thus, the following stages can be distinguished:

- Preparatory stage:
  - Collection and layout of the base of RGB-images with people in various poses
  - Neural network training
  - Calibrate RGB-D cameras
- Main stage:
  - Getting RGB images and depth maps
  - Distortion correction
  - RGB images and depth maps
  - Select a region of interest on a depth map
  - Detecting body parts in an RGB image and obtaining their three-dimensional coordinates
  - Validation of the received skeleton

– Search for skeleton elements by depth map

The most significant stages of the proposed method are considered on the example of practical implementation.

3 Practical implementation of the proposed method

3.1 Preliminary stage

The method of preliminary improvement of video sequence images [28] and the general algorithm of a target environment analyzer [29] are used to prepare images from the camera.

To train the neural network used for the preliminary search for body parts in the RGB image, a database of images was collected and marked out (70% of them for training and 30% for verification).

The database of images was collected by shooting people in various poses and performing some standard actions to obtain each part of the body from several angles and under different shooting conditions. Each image contains a set of traceable body parts, not necessarily complete. From frame to frame, changes were made to: scene illumination, angle, scale and rotation, so as not to train the neural network in specific shooting conditions.

To perform the marking of the image database, software was developed to accelerate the marking process. A screenshot of the program used is shown in Figure 1.

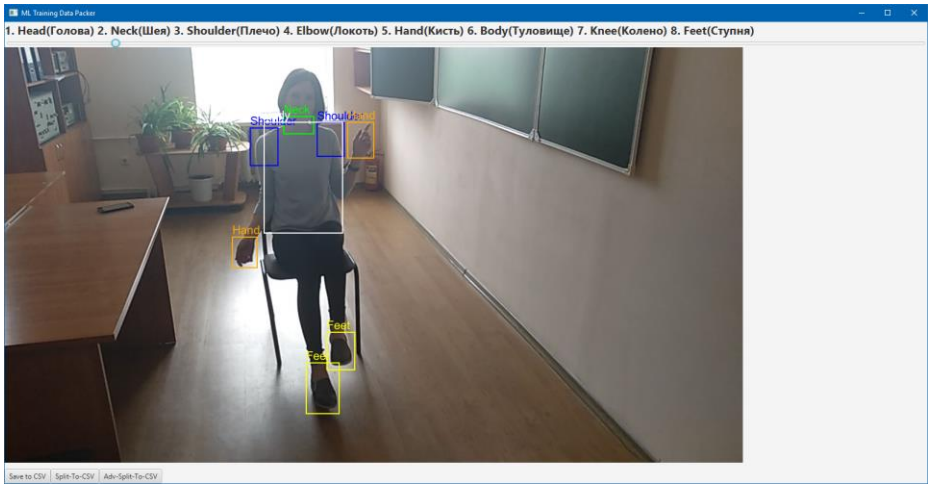

Fig.1. Image Markup Software

The developed software allows you to specify a list of labeled classes, switch between them, select the corresponding areas in the image, load and save markup data in

an xml file, prepare a batch data set for training a neural network using the Tensorflow framework [30, 31].

To train the neural network, the following set of classes was chosen: head, neck, center of the trunk, shoulder, elbow, hand, pelvis, knee, foot. If one of the classes was absent or was poorly visible on the marked image, then the corresponding part of the body was not marked. Examples of labeled images are shown in the following figure 2.

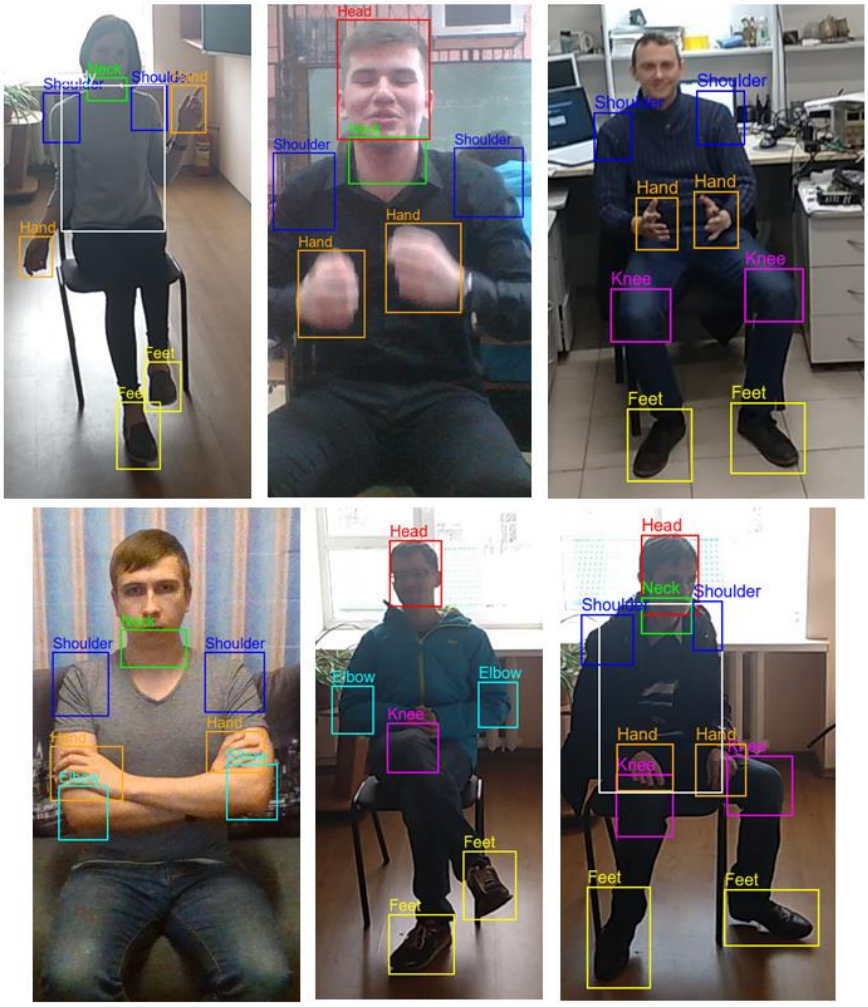

Fig.2. Tagged Image Examples

Based on the obtained set of tagged images and xml files with information about the markup, the neural network was trained. To solve the problem of searching and classifying objects on RGB images, we used the Google Object Detection API, which

allows you to fine-tune the future network model and select the necessary relationship between accuracy and speed.

Neural network training was conducted using Amazon Web Services, and more than 200,000 steps were taken to prepare the final neural network model. The use of cloud services is due to the high computing and time costs of network training.

When an image is received using an RGB camera due to the distortion effect, significant perspective distortions are observed at the image edges [32]. To correct distortion, it is necessary to calibrate the camera methods [33].

3.2 Main stage – building a "human skeleton"

Figure 3 shows a block diagram of the algorithm of the method for detecting a human skeleton by the data stream received from an RGB-D camera.

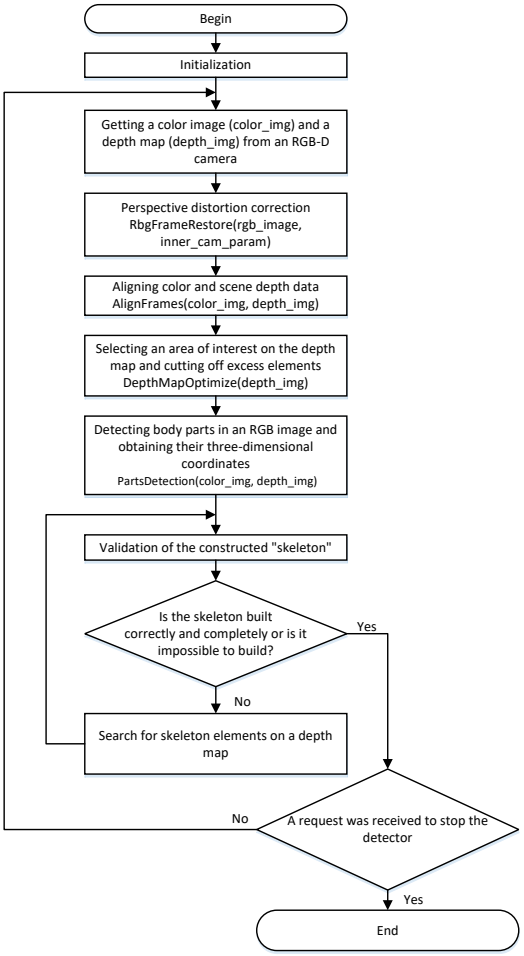

Fig.3. Human skeleton detection algorithm

Let us consider in more detail each of the stages of the method for detecting a human skeleton.

3.3 Acquiring RGB images and depth maps and distortion correction

The RGB-D camera provides a couple of frames: RGB images and depth maps. Frame construction is done by converting the byte stream from the RGB-D camera to the corresponding data structures.

The color image is in RGBA8888 format, the depth map is a 16-bit image in grayscale. The resolution of the RGB image is usually greater than the resolution of the depth map, so at one of the next steps it is necessary to display pixels between this pair of frames.

According to the internal parameters of the camera obtained at the preparatory stage and the calculated distortion coefficients, the perspective distortions of the RGB image are corrected.

3.4 RGB images and depth maps

Knowing the external parameters of a pair of cameras (color and rangefinder), we combine the pair of frames. For each pixel of the depth map, knowing (x, y, z), we calculate the coordinates and size of the region in the color image, the Z coordinate is the value of the depth map in the selected pixel, and it is equal to the distance from the camera plane to the point. The pixel of the depth map corresponds to the region in the color image, and not one dot due to the fact that the resolution of the color image is greater than the resolution of the depth map.

Since the distances to the points in the color image are not known, it is not possible to make directly correspondence of a pixel of the color image to the pixel of the depth map. Therefore, to establish such a correspondence, simply a “back link” is used. The principle of matching by “backlink” is shown in Figure 4.

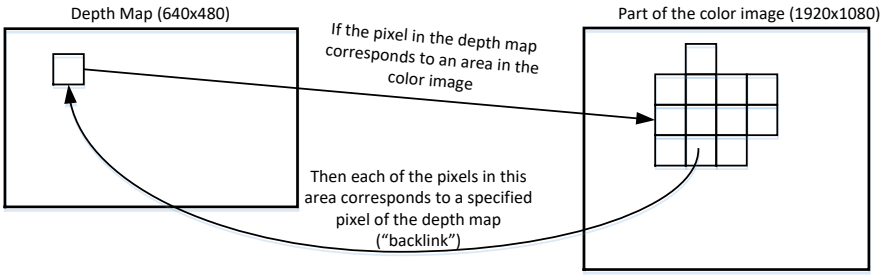

Fig.4. Отображение пикселей между цветным изображением и картой глубины

The correspondence of pixels is established by storing the indices of the corresponding points.

3.5 Selecting a region of interest on a depth map

With the known internal parameters of the depth camera, and data on the location of the control point, it is necessary to cut off all unnecessary details on the depth map, except for the surface related to the human body. The floor and objects located outside the control point can be cut off according to the specified threshold values Y and Z of the coordinates of the points in the scene.

Each pixel of the depth map is converted to a vertex in three-dimensional space using expression (1):

$$\begin{cases} w_x = (d_x - c_x) \cdot \frac{d_z}{f_x}, \\ w_y = (d_y - c_y) \cdot \frac{d_z}{f_y}, \\ w_z = d_z \end{cases} \quad (1)$$

where d_x, d_y, d_z – (x, y) pixel coordinates on the depth map, (z) distance to the corresponding point,

w_x, w_y, w_z – vertex coordinates in three-dimensional space,

c_x, c_y – optical axis coordinates,

f_x, f_y – focal length along the x and y axes

Depending on the position of the camera and the control point, threshold values are set for cutting off excess vertices and points on the depth map. For example, if the frame of the control point is located directly in front of the camera at a distance of 2 m, has a width of 80 cm, and the camera is at a height of 2 m from the floor, you can specify the following set of restrictions:

$$d_z = \begin{cases} 0, \text{при } w_x(i) < -400 \text{ или } w_x(i) > 400, \\ 0, \text{при } w_y(i) < -2000, \\ 0, \text{при } w_z(i) > 2400 \end{cases} \quad (2)$$

The restriction on z (2.4 meters) in (2) was made in order not to capture objects located behind a person passing the control point. Those elements of the depth map that were cut off using expression (2) will not be taken into account when constructing the human “skeleton”.

Performing such post-processing of the depth map is aimed at speeding up calculations and reducing the likelihood of detecting “nonexistent” erroneous parts of the body due to foreign objects in the frame. If the method should be applied in the system without the presence of a limited area of human location, the area of interest can be selected dynamically by searching for a person in an RGB image with further translation of the coordinates into the coordinate system of the depth map.

3.6 Detecting human body parts in an RGB image and obtaining their three-dimensional coordinates and validation of the received "skeleton"

Using a trained neural network, a color image searches for parts of the human body. The points found in the color image are transferred to the corresponding points on the depth map, as a result of which, for each found part of the body, real three-dimensional coordinates are calculated using expression (1).

When performing a detection, the main condition is the search for the head. If it was not found in the color image, then an attempt is made to search for it using the Haar cascade, since the person will most likely look towards the camera. If neither the neural network nor the Haar cascade could perform a search for a person's face, this frame is skipped. This is due to the fact that the head is the simplest element to search, and can act as a starting point for checking the correctness of the construction of the entire skeleton.

After receiving the three-dimensional coordinates of the center of each of the parts of the human body, the correctness of the construction of the skeleton is checked. The following set of criteria is used:

1. The length of the humerus of the left and right hand should differ by no more than 5%;
2. The length of the radius of the left and right hands should differ by no more than 5%;
3. The distance from the center of the neck to the center of the head cannot exceed 1.5 * the height of the head;
4. The center of the body cannot be higher than the neck or head (this test is connected with the assumption of "standard" postures of the person when passing the control point);
5. The lengths of the thighs of the left and right legs should not differ by more than 10% (if the thigh is not visible on the frame, this check is not taken into account);
6. The lengths of the tibia of the left and right leg should not differ by more than 10% (if the tibia is not visible on the frame, the check is not taken into account);
7. The distance from the neck to the shoulders should differ by no more than 15%;
8. The neck and shoulders should be located in a rectangular area responsible for the detection of the human body.

If any of the parts of the body does not satisfy the specified conditions, then it is marked with a marker, according to which at the next stage an attempt will be made to complete the specified part.

On the one hand, the difference in the measurement of the lengths of the left and right arms or legs of 0.5-1.5 centimeters is a rather large measurement error, however, for the tasks of assessing human postures corresponding to non-standard behavior, there is no need for an absolute measurement error of 1 mm. Even with an error of 1.5 cm, it is possible to evaluate the relative position of body parts relative to each other. If from frame to frame, the resulting lengths will range from -0.5 to +0.5 cm, such vibrations can be smoothed out by applying a filter. This can be either a smoothing filter or a Kalman recursive filter.

If violations are observed during the verification of the criteria described above, then an attempt is made to make corrections to the constructed skeleton; if this fails, then the frame is skipped. In case of successful processing of the next frame, the coordinates of the body parts on the previous “unsuccessful” frame are calculated using interpolation. If the number of iterations of this stage exceeds 5 times and the structure of the skeleton with the available parts does not pass verification, then the current frame is marked as incorrect and is skipped.

Figure 7 shows an example of the operation of the detector of body parts and the distance from the plane of the camera to each point:

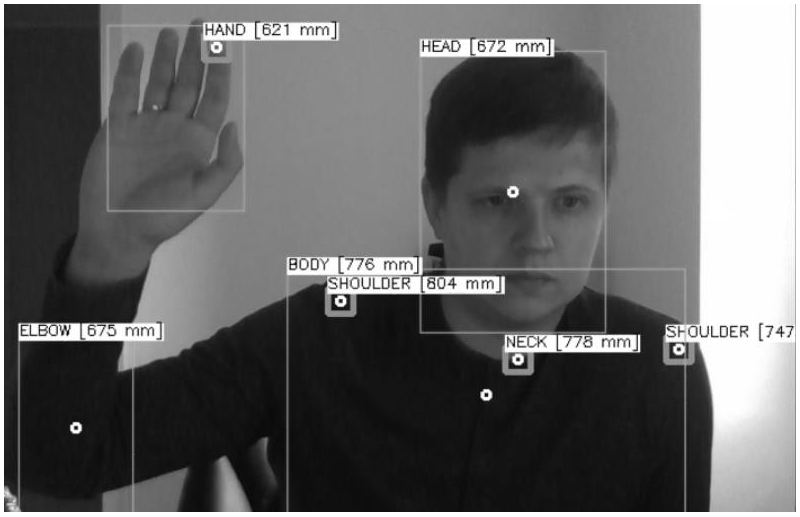

Fig.5. The result of using a body parts detector

3.7 Search for skeleton elements by depth map

Following a given set of rules, an attempt is made to complete the missing parts of the skeleton. Such parts are considered elements that were not found when detecting body parts in the RGB image and obtaining their three-dimensional coordinates, as well as elements marked as unreliable when checking the correctness of the skeleton obtained.

The set of rules is based on a priori information about the structure of the skeleton, the relationship of body parts with each other, and the proportions of the human body. Here are a few examples of the applicable rules.

If there is no neck, then the angle of inclination is determined from the known elements of the torso, and the main neck is searched from the center ready in the given direction - the point of intersection of the axis passing through the center of the head and body-axis and the axis passing through the centers of the left and right shoulder.

If there is no information about the coordinates of the elbows, a gradient descent along the visible surface of the arm is made from the shoulder to the side of the hand,

and a point is searched between them in accordance with the ratio of the length of the humerus and radius of the person.

A similar approach is used for knee detection, only the two edge points are the feet and lateral parts of the human pelvis.

When obtaining coordinates for each missing part of the skeleton, the skeleton structure is also checked for correctness when adding a new node. If the element is found correctly, then its coordinates on the depth map are converted into three-dimensional world coordinates using the expression (1).

4 Acknowledgment

The reported study was performed within the federal program of the Ministry of Science and Education; the agreement unique ID RFMEFI60819X0281.

5 Conclusions and Future Work

The obtained results allow real-time recognition of all visible on the image parts of the body of a person passing the control point.

In the future, it is planned to use the information obtained on the coordinates of the parts of the body and the change in these coordinates over time, to analyze the nature of the movements of the person passing the control point. An analysis of the nature of the movements will make it possible to determine some types of deviations in behavior.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Comput. Vis. image Underst. 108, 4–18 (2007).
2. Patrona, F., Chatzitofis, A., Zarpalas, D., Daras, P.: Motion analysis: Action detection, recognition and evaluation based on motion capture data. Pattern Recognit. 76, 612–622 (2018).
3. Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., Chen, D.-S.: A Comprehensive Survey of Vision-Based Human Action Recognition Methods. Sensors. 19, 1005 (2019).
4. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: A review. ACM Comput. Surv. 43, 16 (2011).
5. Ziaeefard, M., Bergevin, R.: Semantic human activity recognition: A literature review. Pattern Recognit. 48, 2329–2345 (2015).
6. Papadopoulos, G.T., Axenopoulos, A., Daras, P.: Real-time skeleton-tracking-based human action recognition using kinect data. In: International Conference on Multimedia Modeling. pp. 473–483 (2014).
7. Presti, L. Lo, La Cascia, M.: 3D skeleton-based human action classification: A survey. Pattern Recognit. 53, 130–147 (2016).

8. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.-P.: Optimization and filtering for human motion capture. Int. J. Comput. Vis. 87, 75 (2010).
9. Hofmann, M., Gavrila, D.M.: Multi-view 3d human pose estimation combining single-frame recovery, temporal integration and model adaptation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2214–2221 (2009).
10. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1381–1388 (2009).
11. Hen, Y.W., Paramesran, R.: Single camera 3d human pose estimation: A review of current techniques. In: 2009 International Conference for Technical Postgraduates (TECHPOS). pp. 1–8 (2009).
12. Hirai, M., Ukita, N., Kidode, M.: Real-time pose regression with fast volume descriptor computation. In: 2010 20th International Conference on Pattern Recognition. pp. 1852–1855 (2010).
13. Luo, X., Berendsen, B., Tan, R.T., Veltkamp, R.C.: Human pose estimation for multiple persons based on volume reconstruction. In: 2010 20th International Conference on Pattern Recognition. pp. 3591–3594 (2010).
14. Tran, C., Trivedi, M.M.: Human body modelling and tracking using volumetric representation: Selected recent studies and possibilities for extensions. In: 2008 Second ACM/IEEE International Conference on Distributed Smart Cameras. pp. 1–9 (2008).
15. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011. pp. 1297–1304 (2011).
16. Dawn, D. Das, Shaikh, S.H.: A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis. Comput. 32, 289–306 (2016).
17. Liu, A.-A., Xu, N., Nie, W.-Z., Su, Y.-T., Wong, Y., Kankanhalli, M.: Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Trans. Cybern. 47, 1781–1794 (2016).
18. Liu, A.-A., Su, Y.-T., Nie, W.-Z., Kankanhalli, M.: Hierarchical clustering multi-task learning for joint human action grouping and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 39, 102–114 (2016).
19. Gao, Z., Zhang, Y., Zhang, H., Xue, Y.B., Xu, G.P.: Multi-dimensional human action recognition model based on image set and group sparisty. Neurocomputing. 215, 138–149 (2016).
20. Fernando, B., Gavves, E., Oramas, J.M., Ghodrati, A., Tuytelaars, T.: Modeling video evolution for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5378–5387 (2015).
21. Yang, X., Tian, Y.: Super normal vector for activity recognition using depth sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 804–811 (2014).
22. Li, M., Leung, H., Shum, H.P.H.: Human action recognition via skeletal and

depth based feature fusion. In: Proceedings of the 9th International Conference on Motion in Games. pp. 123–132 (2016).
23. Chen, C., Liu, K., Kehtarnavaz, N.: Real-time human action recognition based on depth motion maps. J. real-time image Process. 12, 155–163 (2016).
24. Zhang, J., Li, W., Ogunbona, P.O., Wang, P., Tang, C.: RGB-D-based action recognition datasets: A survey. Pattern Recognit. 60, 86–105 (2016).
25. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European Conference on Computer Vision. pp. 816–833 (2016).
26. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 716–723 (2013).
27. Yang, X., Tian, Y.: Effective 3d action recognition using eigenjoints. J. Vis. Commun. Image Represent. 25, 2–11 (2014).
28. Korovin, I., Khisamutdinov, M., Ivanov, D.: Improvement of a Video Sequence Singular Image. In: Proceedings of the 2nd International Conference on Advances in Artificial Intelligence. pp. 12–15 (2018).
29. Korovin, I., Khisamutdinov, M., Ivanov, D.: A Basic Algorithm of a Target Environment Analyzer. In: Proceedings of the 2nd International Conference on Advances in Artificial Intelligence. pp. 7–11 (2018).
30. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., others: Tensorflow: A system for large-scale machine learning. In: 12th Symposium on Operating Systems Design and Implementation 2016. pp. 265–283 (2016).
31. Rampasek, L., Goldenberg, A.: Tensorflow: Biology's gateway to deep learning? Cell Syst. 2, 12–14 (2016).
32. Tsai, R.Y.: An efficient and accurate camera calibration technique for 3D machine vision. Proc. Comp. Vis. Patt. Recog. 364–374 (1986).
33. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. 22, (2000).