The basic assembly of skeletal models in the fall detection problem

O.S. Seredin [1], A.V. Kopylov [1], E.E. Surkov [1], S.-C. Huang [2]
[1] Tula State University, 300012, Tula, Russia, Lenin Ave. 92;
[2] National Taipei University of Technology, Taipei 106, Taiwan

Abstract

The paper considers the appliance of the featureless approach to the human activity recognition problem, which exclude the direct anthropomorphic and visual characteristics of human figure from further analysis and thus increase the privacy of the monitoring system. A generalized pairwise comparison function of two human skeletal models, invariant to the sensor type, is used to project the object of interest to the secondary feature space, formed by the basic assembly of skeletons. A sequence of such projections in time forms an activity map, which allows an application of deep learning methods based on convolution neural networks for activity recognition. The proper ordering of skeletal models in a basic assembly plays an important role in secondary space design. The study of ordering of the basic assembly by the shortest unclosed path algorithm and correspondent activity maps for video streams from the TST Fall Detection v2 database are presented.

Keywords: skeletal model of human figure, pairwise similarity, activity map, featureless pattern recognition, basic assembly, convolutional neural networks.

Citation: Seredin OS, Kopylov AV, Surkov EE, Huang SC. The basic assembly of skeletal models in the fall detection problem. Computer Optics 2023; 47(2): 323-334. DOI: 10.18287/2412-6179-CO-1158.

Acknowledgments: The work was funded by the Ministry of Science and Higher Education of RF within the framework of the state task FEWG-2021-0012.

Introduction

Design of intelligent systems for daily human activities monitoring to maintain a healthy lifestyle becomes an increasingly actual problem due to the raising of elderly people amount on the planet. It causes a heavy load on welfare and healthcare systems. Most of the monitoring methods described in the literature are focused on activity recognition and identification systems using video surveillance and depth cameras [1, 2] or wearable devices [3, 5]. However, the implementation of systems that analyze information obtained from GPS receivers, accelerometers and other wearable devices is complicated by their rejection by elderly people. This is due to the necessity of system maintenance, learn interaction skills, fear of interference into personal life and changing its usual way [6]. It follows that the most preferable solution is non-invasive systems utilizing binary sensors, infrared cameras, depth sensors which provide the information that does not disturb a privacy. Thus, studies [7, 8] show that privacy-preserving representations of data, such as silhouettes [8] or skeletal representations [9 – 13], can reduce the anxiety of older people about video surveillance systems.

In this work, a skeletal model underlies the representation of the human figure. In general, a skeletal model (skeletal representation, skeleton) of the human figure is a graph formed by the spatial coordinates of vertices (points) reflecting the position of joints and edges that connects them [9 – 13].

Skeleton-based methods for figure, posture and active actions representation of the person could be divided into four main groups.

The first group of methods utilizes the position of the skeleton vertices, which approximately correspond to the position of the joints, in a 3D space. Pairwise relative positions [13] or covariance matrices of these positions are used to describe the human pose [11]. However, as shown in [14], the relative vertices positions are not sufficient for accurate human activity detection, specifically fall detection, and additional spatiotemporal features should be applied.

The second group of methods covers the general geometric characteristics of the skeleton such as bounding rectangle, geometric moments and its invariants, positions or distances from the specific point of the skeleton, i.e. the point corresponding to the head or center of mass, from the floor, etc. These methods are less sensitive to the skeleton estimation defects but do not have enough flexibility to operate well in the complex or changing environment. Thus, method [15] uses a bounding rectangle, the first derivative in height and the first derivative of the width-depth composition. The method involves the Kalman filter for a more accurate estimation of the rate of the height change and components of width and depth. However, the parameters described above are exposed to noise because of the low sensor accuracy. At works [16, 17] a combination of nine relative geometric features is proposed to represent a human pose. These features include eight static characteristics (e.g. the distance between skeleton points, the distance between point and a straight line between the two other points of the skeleton) and single characteristic that takes into account changes in time (changes in the position of points and the speed of their change). The combination of these features makes it possible to assess the position and movement of the human body.

The third group of methods utilize the correspondence between the skeleton and human body parts [18]. These methods consider the human body moves, in particular the fact that human body moves in accordance with shapes, lengths and location of bones, which are more convenient and steadier to observe than joints [19]. The work [20] also considers the relations of neighboring parts of a human body (two arms, two legs and torso).

The fourth group of methods operates in the framework of the featureless approach to pattern recognition, where objects are represented by appropriate pairwise similarity measure or difference. This approach was introduced in [9, 21 – 23] as an alternative to the feature-based methods. Distance metric learning represent a further development of such approach [24, 25].

Because of acquisition artifacts like missing or extra parts of skeletons, geometric and topological noise, etc., the immediate usage of corresponding skeleton representation often becomes intractable. Therefore, the popularity of methods based on pairwise comparison of human figures is increasing in the activity recognition field. The work [26] propose the similarity measure utilizing shape and gradient descriptors. In [16], a semi-supervised distance metric learning algorithm called Regularized Distance Metric Learning with Sparse Representation (RDSR) was introduced using Geometric Pose Descriptor.

Following the featureless pattern recognition idea, this paper proposes to work with pairs of skeletons at once, namely, with the measure of their dissimilarity, instead of parametric description the skeletal models.

The pairwise similarity function provides an opportunity to hide the coordinate representation of the skeleton joints from external observation. Therefore, it allows to better preserve the confidential information about clients and will reduce people's concern about personal privacy. Each skeletal model is represented by a vector of real values reflecting the dissimilarity measure of this model with respect to a fixed set of skeletons. The work [27] defines such a fixed data set as a "basic assembly". After that, for each skeletal model in three-dimensional space received from an RGB-D sensor a set of distances (column vector) to each object of the basic assembly is determined. We propose to call the sequence of such vectors as an activity map, which will be the subject of analysis. If we assign the pixel intensity level to the pairwise dissimilarity value from the activity map, then a greyscale image will be obtained. This image allows to provide a visual analysis of the activity map. The advantage of such an analysis is that it is not possible to recover any sensitive information, which increases the system's privacy.

Methods on the basis of pairwise similarity function allow to apply the featureless pattern recognition principles in deep learning approaches with convolutional neural networks (CNN) which recently have been successfully applied to visual pattern classification. The multilayer architecture of the convolutional neural network allows to move from the specific image features to more abstract concepts. CNN are self-configuring developing the necessary hierarchy of features, filtering the insignificant details of the bitmap image, and highlighting important features. As experiments have shown, the activity map could be a suitable object for deep neural networks applying to the human activity classification. Figure 1 demonstrates the major idea of the proposed approach to the human activity classification problem based on the activity map analysis.

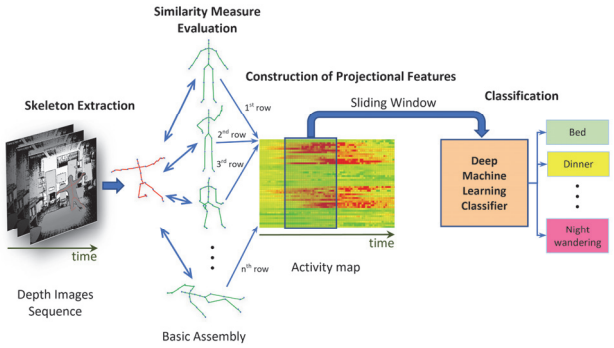


Fig. 1. Skeletal models representation for human activity classification by convolution neural networks based on activity map analysis

According to Fig. 1, it is necessary to prepare a representative set of skeletal models that forms a basic assembly. Then, a pairwise dissimilarity function should be constructed to compare skeletons from a video stream with the skeletons of the basic assembly. The activity map, obtained as a result of comparison, forms an input data for the neural-network-based classifier.

The concept of the activity map, the methods of its evaluation and analysis are the main contributions of this paper.

Skeletal models of the human figure obtained by RGB-D sensors

Nowadays, along with the growing interest in research of human behavior and activity, the problem of representing a human figure with a skeleton obtained from RGB-D sensors is also becoming relevant. Accordingly, there are a large number of approaches to the construction of 2D and 3D skeletal models [9, 10, 35, 36]. In previous study of the fall detection problem [10] a skeletal model (Fig. 2) was obtained by Microsoft Kinect v2 sensor [37] and related software. Fig. 2 shows the skeleton provided by this sensor. Points of the skeletal model that corresponds the fingers and feet coordinates in the space were excluded because of high moveability and absence of useful information for following fall activity analysis [10].

However, after the latest release of the Kinect sensors and the cessation of their production, more advanced devices were released, such as the Intel RealSense D400 [38]. The characteristics of Microsoft Kinect v2, Intel RealSense 435i and Orbecc Astra Pro cameras are considered in Tab. 1.

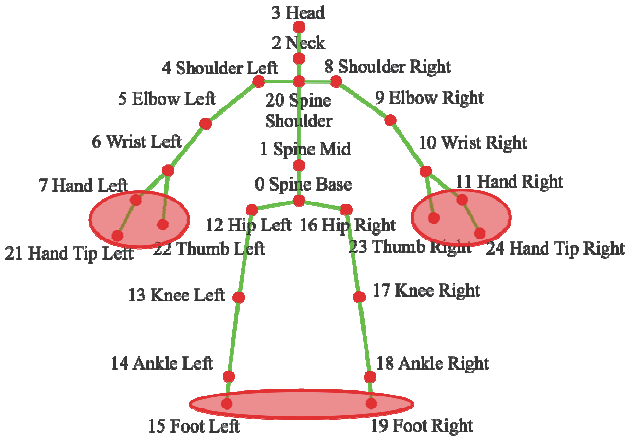


Fig. 2. Human figure skeletal representation obtained by Microsoft Kinect v2. Unused joints are highlighted by red ellipses

The most important technical characteristics of 3D sensors are maximum working distance, horizontal field of view and depth camera resolution. Analysis of presented characteristics shows that Intel RealSense 435i is a superior option. However, the skeletal models obtained from these cameras are different, that makes it difficult to transition between devices, since the characteristic description of a person in the work [10] is based on the position of the skeletal model points (Fig. 2) in space. Consequently, the problem of model's compatibility is occurred.

Tab. 1. RGB-D sensors characteristics

Device name	Microsoft Kinect v2	Intel RealSense 435i	Orbbec Astra Pro
Camera resolution	1920 × 1080 @30 fps	**1920 × 1080 @30 fps**	1280 × 720 @30 fps
Depth resolution	512 × 424 @30 fps	**1280 × 720 @60 fps**	640 × 480 @30 fps
Max distance	4.5 m	**10 m**	8 m
Horizontal view	70 deg	**87 deg**	60 deg
Vertical view	60 deg	**58 deg**	49.5 deg
Skeleton joints	25	**18**	19

Tab. 2 shows a comparison of the software to obtain skeletal models from the sensors presented in Tab. 1.

Tab. 2. Skeletal models SDK's characteristics comparation

	Microsoft Kinect	Nuitrack	Cubemos
Supported programming language			
C++	+	+	+
C#	+	+	+
Java	–	–	+
Python	–	beta-version	+
Supported devices			
RGB-D sensors	Microsoft Kinect	Microsoft Kinect, Intel RealSense, Azure Kinect, Orbbec Astra, Asus Xtion	Any
RGB cameras	–	–	Any
Skeletal model			
Skeletal joints	25	19	18
Dimension of skeletal mode	3D	3D	2D + distance to the skeletal points estimation based on the depth frame

Depth image allows to calculate more accurate skeletal model. It is an obvious advantage in favor of Nuitrack and Microsoft Kinect SDK's. However, the ability to construct a skeletal model from a 2D image provided by the Cubemos SDK is a useful tool while working with RGB cameras.

Fig. 3 shows examples of skeletal models that could be obtained by the Microsoft Kinect, Nuitrack and Cubemos software.

From the Fig. 3 analysis, it could be concluded that models 3a and 3b are practically identical and no additional transformations are required to bring one model to another. Nevertheless, it is not easy to get a skeletal model 3a from a skeletal model 3c. Table 3 shows expressions utilizing human anatomical features that allow to approximately transform model 3c to model 3a. The designations in Tab. 3 (Head, Neck, etc.) means the vector of coordinates of the corresponding skeleton points (Fig. 3).

Thus, the skeletal models shown in Fig. 3*b*, 3*c* could be converted to the model in Fig. 2. The conversion possibility of various skeletal models to the one unified form allows the pairwise similarity function to be invariant to the sensor and software type. Further, we will assume that all skeletons are transformed to the form in Fig. 3*b* using equation in Tab. 3.

The dissimilarity measure between two skeletal models

This paper proposes the representation of a human figure as a vector of real values reflecting the pairwise similarities of its skeletal model with other skeletons [6]. Such representation, instead of descriptive skeleton features proposed in [10], allows us to increase system robustness and privacy. The idea is to select and fix a subset of general assembly objects available to the observer (basic assembly). Further we define a set of secondary projection features for an arbitrary object as a set of its pairwise comparisons with objects of basic assembly [23, 28]. The sequence of projection feature vectors of skeletal models (Fig. 1) obtained from the correspondent sequence of experimental video frames let us to form the activity map.

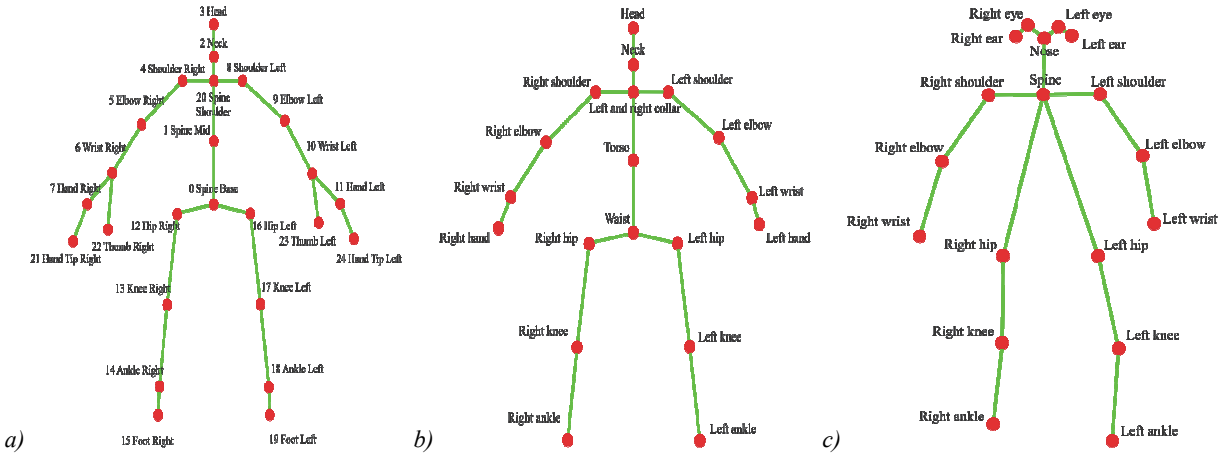


Fig. 3. Skeletal models obtained by a) Microsoft Kinect SDK, b) Nuitrack SDK, c) Cubemos SDK

Tab. 3. Transformations for conversion skeletal model 3c to model 3a

Matching skeletal model points	Skeletal model points transformation	Unused skeletal model points
$Spine\ Shoulder = Spine$; $Hip\ Left = Left\ hip$; $Knee\ Left = Left\ knee$; $Ankle\ Left = Left\ ankle$; $Hip\ Right = Right\ hip$; $Knee\ Right = Right\ knee$; $Ankle\ Right = Right\ ankle$; $Shoulder\ Left = Left\ shoulder$; $Elbow\ Left = Left\ elbow$; $Wrist\ Left = Left\ wrist$; $Shoulder\ Right = Right\ shoulder$; $Elbow\ Right = Right\ elbow$; $Wrist\ Right = Right\ wrist$.	$Head = \frac{Right\ eye + Left\ eye + Right\ ear}{5} + \frac{Left\ ear + Nose}{5}$; $Neck = \frac{Head + Spine\ Shoulder}{2}$; $Spine\ Base = \frac{Right\ hip + left\ hip}{2}$; $Spine\ Mid = \frac{Spine\ Base + Spine\ Shoulder}{2}$.	**Microsoft Kinect:** *Hand Left*; *Hand Right*; *Hand Tip Left*; *Hand Tip Right*; *Thumb Left*; *Thumb Right*; *Foot Left*; *Foot Right*. **Cubemos SDK:** *Right eye*; *Left eye*; *Right ear*; *Left eye*; *Nose*. **Nuitrack SDK:** *Right hand*; *Left hand*.

To calculate the dissimilarity measure (distance) between the objects from the basic assembly and the skeletons from the experimental sequence it is necessary to determine a non-negative continuous real function for comparing two skeletal models. Since determining the distance between two skeletal models is based on the human figure shape, the following aspects should be taken into account:

– various people height. For a correct comparison of two skeletal models to analyze activity by any dissimilarity measure, it is necessary to exclude the influence of human anthropometric characteristics on the length of the skeleton segments;

– when a person moves in the scene, the skeletal model shifts with respect to the camera position (origin of coordinates) and have coordinates relative to the 3D space of the entire room. However, only the relative coordinates of the two compared models are of interest. It follows that it is necessary to neglect the shifting relative to the camera position while comparing skeletal models.

Human height estimation

We assume here that the human height has little effect on how he or she performs a particular action. Therefore, it is necessary to make a dissimilarity measure invariant to the human height and normalize the skeletal model by the corresponding height value.

Let S – is a set of skeletal points and S_i – i-th skeletal point provided by the sensor. Since every i-th point is represented by three coordinates S_i^x, S_i^y, S_i^z, we scale all of them according to the person height by the following expressions:

$$\left\|s_i^x\right\| = \frac{S_i^x}{h}, \quad \left\|s_i^y\right\| = \frac{S_i^y}{h}, \quad \left\|s_i^z\right\| = \frac{S_i^z}{h}, \tag{1}$$

where h – person height, $i = 0,\ldots,16$ – number of used skeletal model points (fig. 2).

Two methods of person height estimation have been proposed.

Method 1 assumes that a person height could be evaluated as an average of Euclidean distance from point 3 of the skeletal model (head) to point 14 (ankle left) and the Euclidean distance from point 3 (head) to point 18 (ankle right) (Fig. 4).

The Euclidean distance between two points of the skeletal model has the following form:

$$d(S_i, S_j) = \sqrt{\sum_{m\in\{x,y,z\}} (S_i^m - S_j^m)^2}, \quad (2)$$

where S – a point coordinate, i,j – point number in a skeletal point set (Fig. 2).

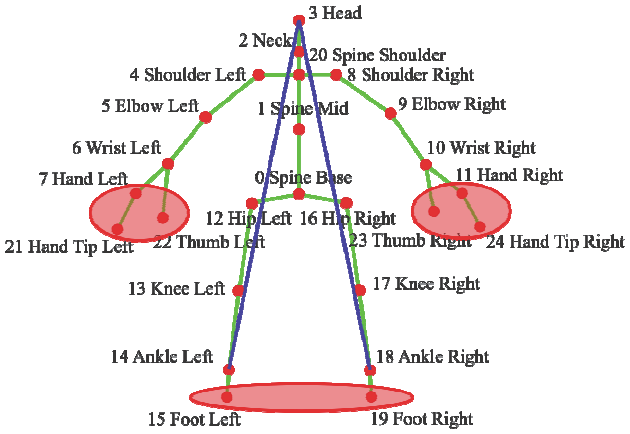


Fig. 4. Skeletal representation of human figure. Auxiliary distances for define height estimate by method 1 are highlighted by blue lines

Height estimation is done by the expression:

$$h = \frac{d(S_3, S_{14}) + d(S_3, S_{18})}{2}, \quad (3)$$

where d is calculated by (2).

A more reliable and accurate estimate of height is the average value calculated from the ten highest values obtained after a certain time of a person's stay in the field of view of the sensor. However, the height estimation determined by method 1 significantly depends on a human pose (e.g., sit or stay). Hence, it is proposed to consider another method.

Method 2 assumes the human height estimation as the geodesic distance between points 3 and 15 and between points 3 and 19 (Fig. 5). We propose to estimate human height by following expression (4).

$$h = d(S_3, S_2) + d(S_2, S_{20}) + d(S_{20}, S_1) + d(S_1, S_0) + \\ + \frac{d(S_0, S_{12}) + d(S_{12}, S_{13}) + d(S_{13}, S_{14})}{2} + \\ + \frac{d(S_0, S_{16}) + d(S_{16}, S_{17}) + d(S_{17}, S_{18})}{2}, \quad (4)$$

where d – is calculated in accordance with (2).

As well as in the first method, the final heigh estimation is the average of the ten highest values obtained after a certain time of a person's stay in the sensor's field of view.

Both methods were evaluated on a video stream in which the actor stands for the first 35 frames and then sits down on a chair. Fig. 6 and 7 show a plot of frame-by-frame height estimations calculated by the first (Fig. 6) and the second methods (Fig. 7).

The diagram analysis of height estimation shows that second method allows to calculate the steadier values than the second method. Bar charts of height estimation were also constructed (Fig. 8 and 9).

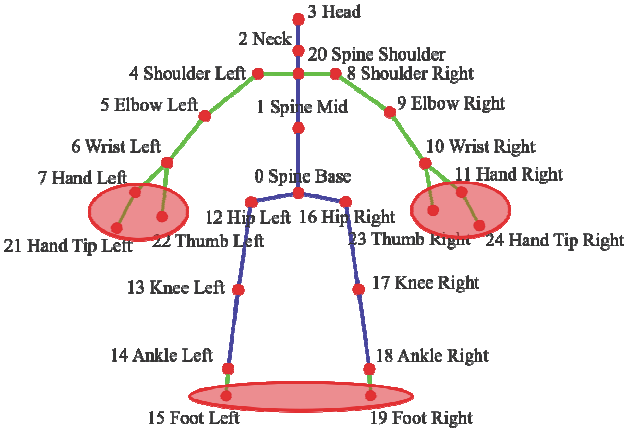


Fig. 5. Skeletal representation of human figure. Auxiliary distances for the height estimation by method 2 are highlighted by blue lines

Bar charts analysis show that geodesic distances between points lead to an asymmetric unimodal distribution on the histogram. The actor's height in the video stream could be clearly recorded as 170 cm. Note, that the first method finds here at least two maximum values.

The fact that the actor in the video stream has two states (standing and sitting) is the reason for the presence of two quasi-constant values in human's height calculation utilizing the Euclidean measure. Such a defect is eliminated by calculating geodetic distances between points 3, 15 and 3, 19 (see Fig. 5).

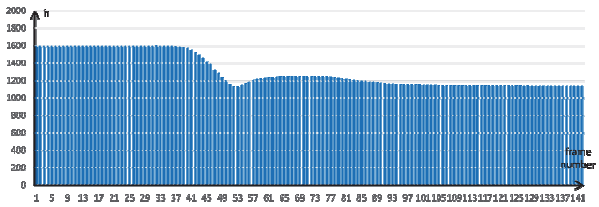


Fig. 6. Frame-by-frame estimation of actor height by method 1

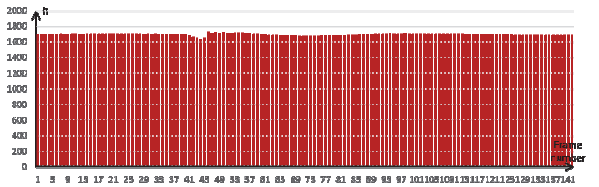


Fig. 7. Frame-by-frame estimation of actor's height by method 2

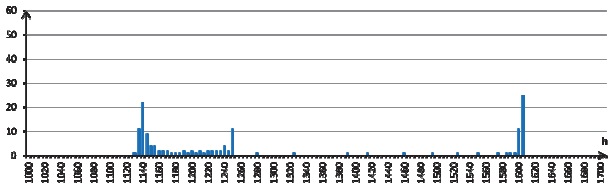


Fig. 8. Bar chart of frame-by-frame height estimation for an actor in a video sequence according to method 1

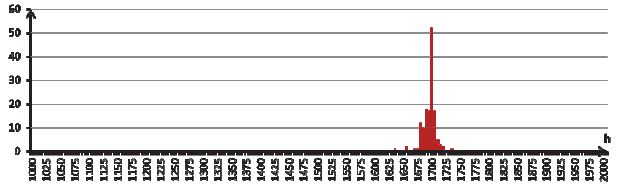


Fig. 9. Bar chart of frame-by-frame height estimation for an actor in a video sequence according to method 2

Distance between two skeletal models

The following problems arise when comparing skeletons in three-dimensional space. All skeletal models move around the scene, even within the same video sequence. Thereby, one skeleton can be farther from the camera or higher than the other. It follows that a correct comparison of skeletal model requires elimination of the bias between them. Considering that the height of the skeleton position in space depends on the person pose, it is proposed to eliminate the vertical bias of each skeletal model in the following way:

– determine the point of the skeletal model with the smallest ordinate (fig. 10*a*)

– subtract the smallest ordinate from each point of the skeletal model. (fig. 10*b*).

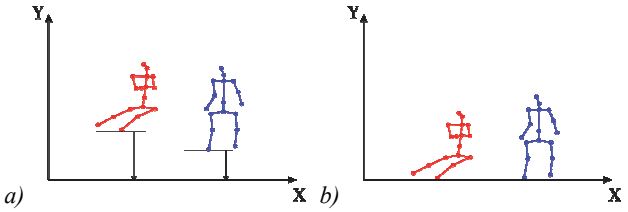


Fig. 10. The elimination of skeletal model vertical bias a) location of the skeletal models before bias elimination, b) location of the skeletal models after bias elimination

The elimination of vertical bias in camera space is performed by the following expression:

$$s_i^y = s_i^y - \min s^y, \quad (5)$$

where $i = 0,\ldots,16$ – index of skeletal model point (Fig. 2).

After that, it is necessary to move the skeletal models to the origin of X and Z axes and combine them at the zero point of the skeletal model (Spine-Base) (Fig. 11).

The combination is performed by the following expression:

$$s_i^x = s_i^x - s_0^x, \quad s_i^z = s_i^z - s_0^z, \quad (6)$$

where $i = 0,\ldots,16$ – index of skeletal model point (fig. 2).

After bias elimination, the measure of dissimilarity between skeletons is calculated. This work considers the average Euclidean distance between the corresponding points of two skeletons as a dissimilarity measure (Fig. 12).

The distance between pairs of skeletons P and Q could be determined by the following expression:

$$R(P,Q) = \frac{1}{N}\sum_{k=0}^{N-1}\sqrt{\sum_{m\in\{x,y,z\}}\left(p_k^m - q_k^m\right)^2}. \quad (7)$$

where N – number of used points of skeletons, p_k – k-th point of skeletal model P, q_k – k-th point of skeletal model Q.

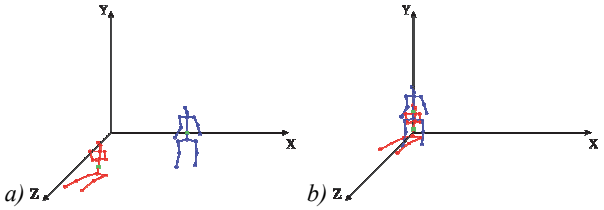


Fig. 11. Combining two skeletal models along X and Z axes at the point 0 (SpineBase) a) location of the skeletal models before combining, b) location of the skeletal models after combining

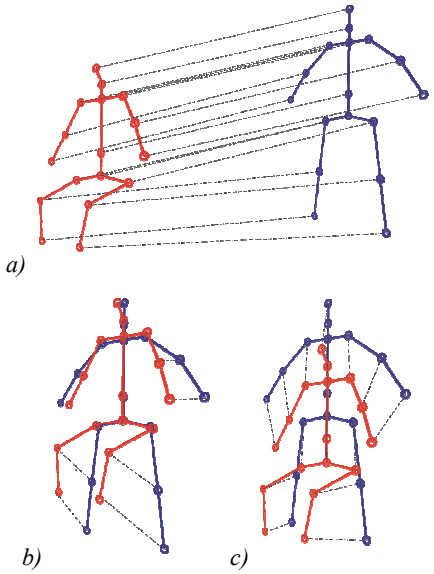


Fig. 12. Combining of two skeletal models a) skeletal models without combining, b) combining of skeletal models at point 0, c) combining of skeletal models along X, Z axes and along common line

Overview of databases with falls recorded by Microsoft Kinect v2

Experimental studies require a database containing human activity monitoring data, including falls. A comprehensive overview of databases could be found in [29]. In the works [9, 10] dedicated to the human fall detection problem, the TEST Fall Detection v2 database [30] was used for experimental research of fall detection algorithms. It contains depth maps and 3D skeletons collected by the Microsoft Kinect v2 sensor and presented as records of various durations. The dataset consists of daily activity records and falls, modeled by 11 actors. It includes Daily Living Activities (ADL) in the following categories: sitting, bending over and lifting, walking, lying down, as well as activities in the FALL category: falling forward, falling backward, falling sideways, falling back, and staying in a sitting position.

The total number of entries in the TST Fall Detection v2 dataset is 264, and the total number of frames is 46,418. The frame rate of records is 30 frames per second. The shortest record duration is 2.5 s (75 frames), and the longest record duration is 15.4 s (463 frames).

Another considered database NTU RGB+D 120 [31] is also recorded by the Microsoft Kinect v2 sensor. It contains 120 classes of activities presented in 114 480 videos. Each instance in the database is accompanied by the following data: RGB video, depth maps, skeletal models and infrared data. Activities are divided into 3 categories:

– daily actions, such as "drinking water", "eating food", "brushing teeth", "jumping", "putting on / taking off clothes" and so on;

– medical conditions such as "cough", "back pain", "neck pain", "fall", "yawn";

– mutual actions or two-person interactions.

The activities were recorded by three cameras with three different horizontal views at different heights and distances to the object. For each activity, there are also several videos with different actors in this database. 106 different actors from different countries were invited. Actors of various ages from 10 to 57 years old, the height of actors from 1.3 to 1.9 m [31].

The TST Fall Detection database has fewer different activities than the NTU dataset. However, each video in TST database contains not only a record of the activity itself, but the actions before and after this activity. This allows us to trace the transition processes between human activities on the distance matrix visualization.

TST Fall Detection v2 is one of the most modern datasets that have a quite large number of video streams with various contents. Therefore, this particular data set was chosen for experiments.

Activity map construction based on the distances between skeletal models

After determining the pairwise dissimilarity function of two skeletal models, it becomes possible to use the principle of featureless pattern recognition based on the basic assembly idea [23]. A representative set of skeletal models recorded during the laboratory research is used as a basic assembly. The objects are chosen empirically, so that the skeletal models of the basic assembly correspond to different positions of an actor:

– standing (32 frames);

– sitting on the char (28 frames);

– sitting on the floor (72 frames);

– lying (4 frames).

For a clear demonstration, it is proposed to construct a distance matrix between objects of the basic assembly and also to provide its visualization.

The values in the distance matrix could be replaced by grayscale values, where black is the zero distance between objects, and white is the maximum distance. The distance between the skeletal models from all videos in TST Fall Detection v2 database and skeletons from basic assembly was accepted as a maximum distance value. This value is approximately equal to 0.86.

A matrix of pairwise distances between the basic skeletons $K \times K$ is constructed following the dissimilarity measure in the form of (7). The visualization of such a matrix is shown in Fig. 13. The skeletons in the basic assembly are not ordered in any way. They are only divided according to the classes of activities described above.

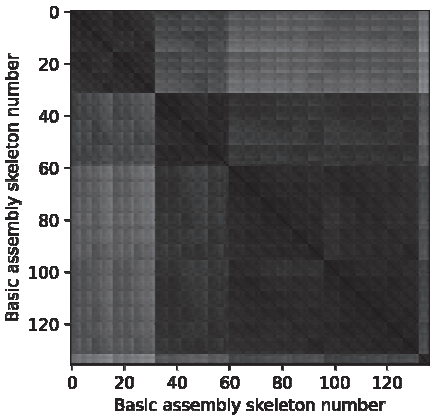


Fig. 13. The grayscale representation of the distance matrix between skeletal models of the basic assembly. The axes are numbered according to the order number of a skeletal model in the basic assembly

The fixed basic assembly allows us to obtain a distance matrix between its elements and frames of the video sequence. Such a distance matrix is called an activity map. For each skeleton from the video sequence a vector of secondary projective features is calculated which we will interpret as one column of the activity map. A sequence of such dissimilarity measure values, with respect to the certain object in the basic assembly in time, forms one row of the activity map. The greyscale visualization of the activity map clearly shows how far the objects are from each other with respect to the proposed dissimilarity measure.

Several video sequences from the TST Fall Detection v2 database were randomly selected to provide experiments and obtain activity maps. The first video sequence (Data1/Fall/EndUpSit/1) has 180 frames with the fall activity. The fall start was noted by experts at frame 55, and the end of the fall at 146. After the fall, the actor remains in a sitting position. The visualized activity map for video sequence 1 is shown in Fig. 14.

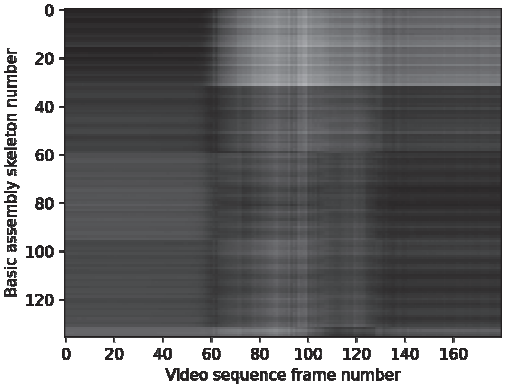


Fig. 14. Activity map visualization for video sequence 1 (action: falling)

The second experimental video sequence (Data 2/ADL/sit/1) contains 155 frames. It has only two activities. The first activity (0 – 100 frames) – the actor is standing, the second activity (100 – 155 frames) - the actor is sitting. Such a video sequence allows us to visually estimate the presence of an explicit brightness boundary between activities on the activity map. The visualization of the activity map for video sequence 2 is shown in Fig. 14.

From the Fig. 14 analysis it could be concluded that the visualization of the activity map reflects obvious transitions between types of activities but does not demonstrate its clearly enough.

Ordering of the basic assembly of skeletal models

The quality of the activity map directly depends on the rows order. Initially the rows order in the distance matrix of the basic skeletal representations is not strictly defined.

But such an order should be obviously determined by the structure of the basic assembly. This structure, in turn, should reflect the similarity between the objects of the basic assembly itself. Since the row for the particular basic skeleton represents the distance to other objects in basic assembly, the following hypothesis could be set up. If each element of the basic assembly is an element of the metric space, then the shortest path will arrange the elements in such an order that:

– the transition between the boundaries will be smoother;
– the boundaries between the individual activities themselves will be more explicit and recognizable;
– the activity map will be smoother and more representative.

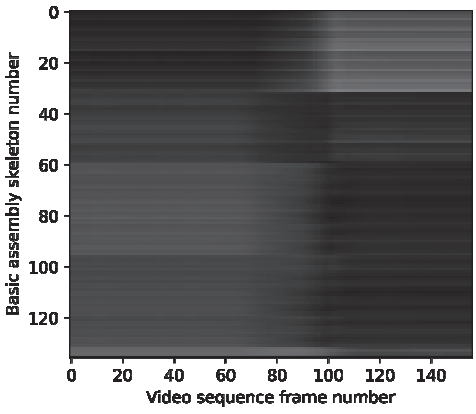


Fig. 15. Activity map visualization for video sequence 2 (action: sitting)

The use of the principle of the decision rule smoothness promotes to reduce the Curse of Dimensionality when recognizing images in the work [32]. The experiments on face recognition showed that the proposed modification significantly increases the predictive ability of the Support Vector Machine learning method initially used in several previous works. In [32] it is assumed that ignoring individual insignificant details of the face by smoothing the training set images will help to significantly improve the quality of recognition. Therefore, when working with images as training objects, it is useful to consider the proximity between objects and their order. The introduction of constraints on the ordering of objects in the training set is called the decision rules regularization [33]. Works [33] also shows that the application of the principle of decision rule smoothness will lead to an improvement in the recognition quality.

To find the shortest unclosed path (SUP) between all the objects of the basic assembly a SUP search algorithm from [34] could be applied. The article [34] proposes 8 algorithms for searching the quasi-shortest unclosed path and provides a comparative analysis on several datasets. Recursive modification of the A4 algorithm makes it possible to find the best result on the data of the basic assembly in a reasonable time [34]. The "ant" algorithm was proposed to improve the solution obtained by modifications of the greedy algorithm. This algorithm search for the SUP between two terminal points [39]. Such points are assumed to be the terminal points of the SUP found by the A4R algorithm from [34].

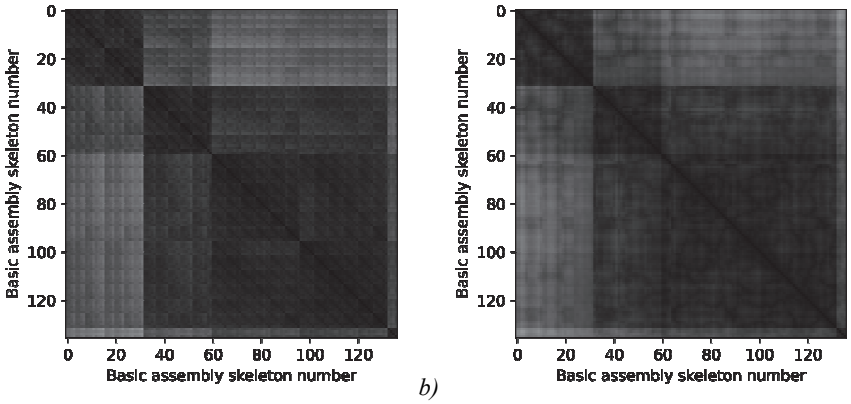
The "ant" algorithm simulates the process of ant's natural behavior in nature. It is based on the exploring of the territory adjacent to the anthill (starting point) for the presence of food sources (end point) and marking more successful paths from the ant colony to the source with a large amount of pheromone [6]. The principle of the algorithm is as follows: a virtual "ant" is placed at the initial node. Further, the probability of moving to the next available points is determined by the edge length (the distance between points) and the amount of pheromone lay on it by previous "ants" generations. The initial amount of pheromone on all edges is initialized with a nonzero starting value. It should be noted that the initial idea of the algorithm is to find the shortest path from the starting point to the final point, without considering the obligatory passage through all points. Thus, to solve the problem the shortest path searching through all points, it is necessary to enter a queue that contains unvisited nodes except for the final one. By the moment when all nodes from the queue have been visited, the accessibility of the transition from the last point in the path to the terminal one is checked. If the transition is available, then the path is considered successful.

Eventually, the result obtained by the basic algorithms was improved by applying the "ant" algorithm [38]. The length of the SUP found by this algorithm decreased from 5.95 to 5.897. During the SUP length decreasing the objects sequence of the first and second cluster in basic assembly was greatly reordered. The third cluster was recorded in reverse order, the fourth cluster remained unchanged.

Fig. 16 shows a comparison of visualization of distance matrices between skeletal models of the basic assembly placed in the arbitrary order and in the strict order, which is determined by the solution of the SUP

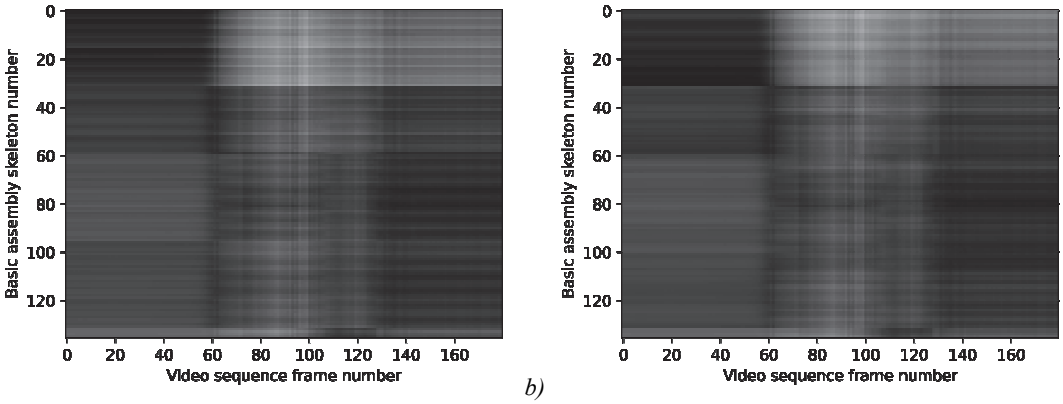
search problem. By Fig. 16 analysis, it could be noted that Fig. 16*b* has smoother color transitions between skeletal models.

A comparison of activity maps with an unordered and ordered basic assembly obtained for video sequences 1 and 2 is presented in Fig. 17 and 18.



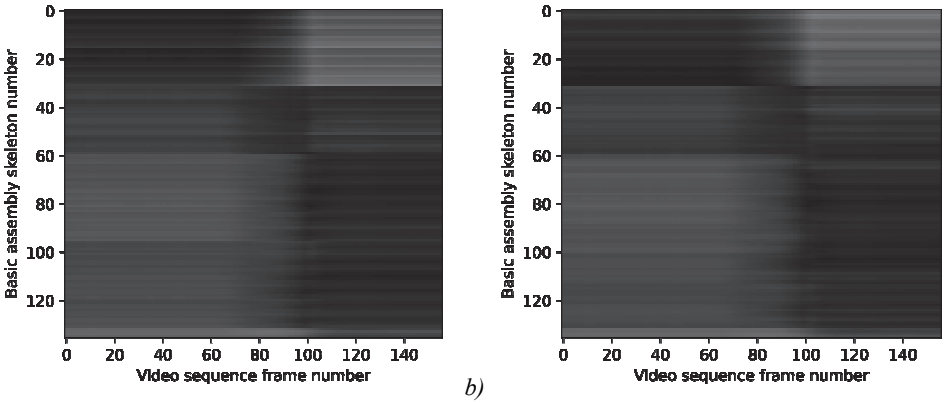
a) *b)*

Fig. 16. Visualization of distance matrix between basic skeletons a) any order, b) strict order



a) *b)*

Fig. 17. Activity map comparation for video sequence 1 (action: falling) a) Unordered basic assembly b) Ordered basic assembly according to solution obtained by the SUP search



a) *b)*

Fig. 18. Activity map comparation for video sequence 2 (action: sitting) a) Unordered basic assembly b) Ordered basic assembly according to solution obtained by the SUP search

Comparison of the visualizations in Fig. 17 and 18 shows that the ordering of basic assembly allows us to obtain more explicit and contrast transitions between the activities on the distance matrix visualizations. It allows us to assume that the regularization of the decision rules [33] will entail a more effective recognition result.

Experimental results

The experimental part compares results of fall detection, obtained by the proposal method, with results published earlier. The evaluation was provided on the TST Fall Detection v2. In this paper the FALL and ADL activities represented by activity maps was recognized by residual convolutional neural network ResNET50 without pretrained weights. Resnet50 decision is reinforced by CUSUM procedure to adjust classifier decisions on successive frames as proposed in [10].

We solve two-class problem using the binary cross-entropy loss function. The neural network was trained on the activity maps with 136×32 shape because the size of basic

assembly is 136 and the minimal first layer input shape is 32. The training dataset was prepared as follows (Tab. 4).

The pairwise distances (7) between the skeleton from the video frame and basic assembly were evaluated on each frame. The first activity map we obtained after the first 32 frames of video sequence. Then we accumulated activity maps for whole video stream by the one-step sliding window (fig. 1).

Tab. 4. The accuracy of fall detection algorithms on TST Fall Detection Dataset v2

Method	Source data	Classifier	Evaluation Scheme	Accuracy
Gasparrini et al., 2016 [40]	Skeleton joint position; accelerometer data	Empirical thresholding rule	Not described	0.990
Fakhrulddin et al., 2018 [41]	Two accelerometers time series data	CNN (self-structured)	Separate data on 90 % and 10 %, then averaging	0.923
Hwang et al., 2017 [42]	Depth map	3D-CNN (self-structured)	5 random trials from 240 and 24 records splitting and averaging	0.942
Min et al., 2018 [43]	Skeleton joints information	SVM	Two-third for training and one-third for tests	0.920
Seredin et al., 2021 [9]	Reduction skeleton joints information	SVM + One-class classifier + CUSUM	Leave-One-Person-Out	0.936
Proposed	Skeleton based activity map	CNN (ResNET50) + CUSUM	Leave-One-Person-Out	0.947

This pipeline is applied for all video sequences in TST V2. The activity map categories “FALL” and “ADL” were defined by the binary code. If the activity map contains entire fall activity then we marked it as “FALL” and code 10, otherwise the activity map had “ADL” label and code 01. It’s important that activity maps which contain both “ADL” and “FALL” labeled vectors were excluded from the train dataset.

Next, we splitted the dataset into training and test parts, with the condition that the training set includes ten of the eleven actors in the database, and the remaining actor record was used exclusively for tests. Such a test procedure was applied to each actor (Leave-one-person-out) [9, 10].

The experiments show that the application of the proposed approach to the fall activity recognition increases the accuracy to 0.947. It follows that we outperform our previous results as well as others except of Gasparrini et al., 2016 (tab. 4). Notice that our method excludes any wearable devices as opposed to Gasparrini’s method that is important for non-invasive systems.

Conclusion

The paper describes an approach to representing a skeletal model by projecting the object of interest to the secondary feature space, formed by the basic assembly of skeletons. A series of experiments was provided to determine the best method of human height estimation to eliminate difference in the height of skeletal models. Also, we propose a measure of dissimilarity between skeletal models and the activity map concept. As a result, activity maps were obtained for video sequences from the TST Fall Detection v2 database.

The proposed method for representing human activity by activity maps extends the principles of featureless pattern recognition to deep learning methods based on convolutional neural networks for solving the activity classification problem. Also, the experiments based on Leave-one-person-out procedure showed that this approach increases the accuracy from 0.936 to 0.947.

References

[1] Chen D, Bharucha AJ, Wactlar HD. Intelligent video monitoring to improve safety of older persons. Proc Annual Int Conf of the IEEE Engineering in Medicine and Biology 2007: 3814-3817.

[2] Di Huang C, Wang CY, Wang JC. Human action recognition system for elderly and children care using three stream ConvNet. Proc 2015 Int Conf on Orange Technologies (ICOT 2015) 2016; 4: 5-9.

[3] Abbate S, Avvenuti M, Light J. MIMS: A minimally invasive monitoring sensor platform. IEEE Sens J 2012; 12(3): 677-684.

[4] Vuong NK, Chan S, Lau CT, Chan SYW, Yap PLK, Chen ASH. Preliminary results of using inertial sensors to detect demential related wandering patterns. Proc Annual Int Conf of the IEEE Engineering in Medicine and Biology Society (EMBS) 2015; 2015: 3703-3706.

[5] Kumar A, Lau CT, Chan S, Ma M, Kearns WD. A unified grid-based wandering pattern detection algorithm. Proc Annual Int Conf of the IEEE Engineering in Medicine and Biology Society (EMBS) 2016; 2016: 5401-5404.

[6] Wearables have a dirty little secret: 50% of users lose interest – TechRepublic. Source: ⟨http://www.techrepublic.com/article/wearables-have-a-dirty-little-secret-most-people-lose-interest/⟩.

[7] Wild K, Boise L, Lundell J, Foucek A. Unobtrusive in-home monitoring of cognitive and physical health: reactions and perceptions of older adults. J Appl Gerontol 2008; 27: 181-200.

[8] Demiris G, Oliver DP, Giger J, Skubic M, Rantz M. Older adults’ privacy considerations for vision-based recognition methods of eldercare applications. Technol Heal Care 2009; 17(1): 41-48.

[9] Seredin OS, Kopylov AV, Surkov EE. The study of skeleton description reduction in the human fall-detection task. Computer Optics 2020; 44(6): 951-958. DOI: 10.18287/2412-6179-CO-753.

[10] Seredin OS, Kopylov AV, Huang SC, Rodionov DS. A skeleton features-based fall detection using Microsoft Kinect v2

with one class-classifier outlier removal. ISPRS –International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 2019; 4212: 189-195.
[11] Hussein ME, Torki M, Gowayyed MA, El-Saban M. Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations. Proc Twenty-Third Int Joint Conf on Artificial Intelligence (IJCAI '13) 2013: 2466-2472.
[12] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition 2014: 588-595.
[13] Wang J, Liu Z, Wu Y, Yuan J. Mining actionlet ensemble for action recognition with depth cameras. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition 2012: 1290-1297.
[14] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton based action recognition. arXiv Preprint. 2018. Source: ⟨https://arxiv.org/abs/1801.07455⟩.
[15] Bevilacqua V, et al. Fall detection in indoor environment with kinect sensor. 2014 IEEE Int Symposium on Innovations in Intelligent Systems and Applications (INISTA) 2014: 319-324.
[16] Chen C, Zhuang Y, Nie F, Yang Y, Wu F, Xiao J. Learning a 3D human pose distance metric from geometric pose descriptor. IEEE Trans Vis Comput Graph 2011; 17(11): 1676-1689.
[17] Zhang S, Liu X, Xiao J. On geometric features for skeleton based action recognition using multilayer LSTM networks. Proc 2017 IEEE Winter Conf on Applications of Computer Vision (WACV 2017) 2017: 148-157.
[18] Bian P, Hou J, Chau P, Magnenat Thalmann N. Fall detection based on body part tracking using a depth camera. IEEE J Biomed Heal Informatics 2015; 19(2): 430-439.
[19] Zhang X, Xu C, Tao D. Graph edge convolutional neural networks for skeleton based action recognition. arXiv Preprint. 2018. Source: ⟨https://arxiv.org/abs/1805.06184⟩.
[20] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition. 2015 IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2015: 1110-1118.
[21] Mottl V, Seredin O, Dvoenko S, Kulikowski C, Muchnik I. Featureless pattern recognition in an imaginary Hilbert space. 2002 Int Conf on Pattern Recognition 2002; 2: 88-91.
[22] Duin RPW, Pękalska E, De Ridder D. Relational discriminant analysis. Pattern Recognit Lett 1999; 20(11-13): 1175-1181.
[23] Seredin OS. Methods and algorithms for features pattern recognition. The thesis for the Candidate's degree in Physical and Mathematical Sciences. Tula, 2001.
[24] Yang L, Jin R. Distance metric learning: A comprehensive survey. Michigan State Universiy; 2006: 1-51.
[25] Kaya M, Bilge HS. Deep metric learning: A survey. Symmetry 2019; 11(9): 26.
[26] Wang T, Wang S, Ding X. Learning a similarity metric discriminatively for pose exemplar-based action recognition. 2011 4th Int Congress on Image and Signal Processing 2011; 1: 404-408.
[27] Mottl V, Seredin O, Krasotkina O. Compactness hypothesis, potential functions, and rectifying linear space in machine learning. In Book: Braverman readings in machine learning. Key ideas from inception to current state. Cham: Springer; 2018: 52-102.
[28] Pekalska E, Duin RPW, Paclik P. Prototype selection for dissimilarity-based classifiers. Pattern Recognit 2006; 39: 189-208.
[29] Cai Z, Han J, Liu L, Shao L. RGB-D datasets using Microsoft Kinect or similar sensors: a survey. Multimed Tools Appl 2017; 76: 4313-4355. DOI: 10.1007/s11042-016-3374-6.
[30] IEEE DataPort TST Fall Detection Dataset v2. Source: ⟨https://ieee-dataport.org/documents/tst-fall-detection-dataset-v2⟩.
[31] Liu J, Shahroudy A, Perez M, Wang G, Duan L-Y, Kot AC. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Trans Pattern Anal Machine Intell 2020; 42(10): 2684-2701. DOI: 10.1109/TPAMI.2019.2916873.
[32] Seredin O, Mottl V. Regularization in image recognition: The principle of decision rule smoothing. Proc Ninth Int Conf Pattern Recognition and Information Processing 2007; II: 151-155.
[33] Seredin O, Kopylov A, Mottl V. Selection of subsets of ordered features in machine learning. machine learning and data mining in pattern recognition. In Book: Perner P, ed. Machine learning and data mining in pattern recognition. Springer; 2009: 16-28.
[34] Seredin O, Surkov E, Kopylov A, Dvoenko S. Multidimensional data visualization based on the shortest unclosed path search. In Book: Dang NHT, Zhang YD, Tavares JMRS, Chen BH, eds. Artificial intelligence in data and big data processing. ICABDE 2021. Cham: Springer; 2022: 279-299.
[35] Efstratiou P. Skeleton tracking for sports using LiDAR depth camera. 2021. Source: ⟨http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-297536⟩.
[36] Ponmozhi K, Deepalakshmi P. A posture recognition system for assisted self-learning of Yoga. EAI Int Conf on Big Data Innovation for Sustainable Cognitive Computing (BDCC 2018) 2019: 231.
[37] Wasenmüller O, Stricker D. Comparison of Kinect v1 and v2 depth images in terms of accuracy and precision. Asian Conf on Computer Vision 2016: 34-45.
[38] Grunnet-Jepsen A, et al. Projectors for Intel® RealSense™ depth cameras D4xx. Santa Clara, CA: Intel Support, Intel Corporation; 2018.
[39] Kazharov AA, Kureichik VM. Ant colony optimization algorithms for solving transportation problems. Int J Comput Systems Sci 2010; 49(1): 30-43.
[40] Gasparrini S, et al. Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In Book: Loshkovska S, Koceski S, eds. ICT innovations 2015. Cham: Springer; 2016: 99-108.
[41] Fakhrulddin AH, Fei X, Li H. Convolutional neural networks (CNN) based human fall detection on Body Sensor Networks (BSN) sensor data. 2017 4th Int Conf on Systems and Informatics (ICSAI) 2017: 1461-1465.
[42] Hwang S, Ahn DH, Park H, Park T. Maximizing accuracy of fall detection and alert systems based on 3D convolutional neural network: Poster abstract. Proc Second Int Conf on Internet-of-Things Design and Implementation (IoTDI '17) 2017: 343-344.
[43] Min W, et al. Support vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle. IET Comput Vis 2018; 12: 1133-1140.

Authors' information

Oleg Sergeevich Seredin received the Ph.D. Degree in Theoretical Foundations of Informatics from Computing Center of the Russian Academy of Sciences, Moscow, Ph.D. Thesis: "Methods and algorithms of featureless pattern recognition" (2001). Now he is Associate Professor at the Institute of Applied Mathematics and Computer Science, Tula State University and Leading Researcher at Laboratory of Cognitive Technologies and Simulating Systems, Tula State University. His scientific interests are data mining, pattern recognition, machine learning, signal and image analysis, visual representation of multidimensional data, statistical methods of decision making. He is a member of program committee at several conferences (CloudCom, AIST, GraphiCon, VISAPP, PSBB, PRIB, MaDaIn) and Reviewer Board Member of several journals (Sensors, Computer Optics, SN Computer Science Journal, IEEE Signal Processing Letters, Applied Science, etc.). Prof. Seredin is principal investigator of several grants of the Russian Science Foundation and Russian Fund for Basic Research, including international. He worked as visiting scientist at Rutgers University and National Taipei University of Technology. He has published more than 100 scientific papers in refereed journals, handbooks, and conference proceedings in the areas of machine learning, pattern recognition and computer vision. Prof. Seredin is a member of The International Association for Pattern Recognition (IAPR). E-mail: *oseredin@yandex.ru* .

Andrei Valerievich Kopylov received the Ph.D. degree from the Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia, in 1997. In 1997, he joined of Automation and Remote Control department, Tula State University, as an Assistant Professor and became an Associate Professor in 2005. Currently, he is an Associate Professor with the Institute of Applied Mathematics and Computer Science, Tula State University. Since 2022, he is also a leading researcher in the Laboratory of Cognitive Technologies and Simulating Systems, Tula State University. He worked as visiting researcher at the Dorodnicyn Computing Centre of Russian Academy of Sciences and National Taipei University of Technology. His scientific interests are signal and image analysis, data mining, machine learning. Prof. Kopylov was the principal investigator of several grants of the Russian Fund for Basic Research, including international. He is a member of program committee at several conferences (CloudCom, PSBB, SoICT, AIST, GraphiCon, VISAPP, ICPR), reviewer of scientific journals Sensing and Imaging (SSTA), Computer Optics, Machine Learning and Data Analysis (JMLDA), IEEE Signal Processing Letters, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Access, etc. He has published more than 100 scientific papers in refereed journals, handbooks, and conference proceedings in the areas of machine learning, pattern recognition and computer vision. Prof. Kopylov is a member of The International Association for Pattern Recognition (IAPR). E-mail: *and.kopylov@gmail.com* .

Egor Eduardovich Surkov received the B.S. degree in Computer Science from Tula State University, Tula, Russia, in 2021. His research interest includes machine learning, pattern recognition, human activity detection, visualization of multidimensional data. He has participated in all-Russian and international conferences (ICABDE, MMPR, TVCS, Lomonosov). He has published about 10 scientific papers in refereed journals, conference proceedings in the areas of machine learning, pattern recognition and computer vision. Surkov is investigator of several grants of Russian Fund of Basic Research and the Innovation Assistance Fund. He is also winner of several nominal grants and awards in science and academic achievements. E-mail: *eg-su@mail.ru* .

Shih-Chia Huang is a Full Professor of Electronic Engineering department, National Taipei University of Technology, Taipei, and an International Adjunct Professor of Business and Information Technology faculty, University of Ontario Institute of Technology, Oshawa, ON, Canada. He received the B.S. degree from National Taiwan Normal University, Taipei, Taiwan, the M.S. degree from National Chiao Tung University, Hsinchu City, Taiwan, and the Doctorate degree in Electrical Engineering from National Taiwan University, Taipei, in 2009. Dr. Huang is currently the Chapter Chair of the IEEE Taipei Section Broadcast Technology Society and an Associate Editor of the IEEE Sensors Journal, IEEE Open Journal of the Computer Society, and Electronic Commerce Research and Applications, respectively. His research interests include intelligent multimedia systems, deep learning and artificial intelligence, image processing and video coding, intelligent video surveillance systems, cloud computing and big data analytics, and mobile applications and systems. He was the recipient of the Kwoh-Ting Li Young Researcher Award in 2011 by the Taipei Chapter ofthe Association for Computing Machinery, the 5th National Industrial Innovation Award in 2017 by the Ministry of Economic Affairs, Taiwan, as well as the Dr. Shechtman Young Researcher Award in 2012 by the National Taipei University of Technology. He was also the recipient of an Outstanding Research Award from the National Taipei University of Technology in 2014 and 2017, and the College of Electrical Engineering and Computer Science, National Taipei University of Technology in 2014–2016. E-mail: *schuang@ntut.edu.tw* .

Code of State Categories Scientific and Technical Information (in Russian – GRNTI)): 29.31.15, 29.33.43, 20.53.23.
Received May 18, 2022. The final version – September 1, 2022.