Predicting of a person's position in trajectory tracking from a continuous video stream

Oleg Amosov*, and *Svetlana* Amosova

V. A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

Abstract. The paper proposes a method for predicting when a person enters a forbidden zone during his trajectory following a video stream, considering individual body parts. The authors used the PP-TinyPose PaddleHub neural network model with its implementation based on two deep neural networks to detect key points of the human body. The paper considers an example of human position prediction from a continuous video stream in indoor trajectory tracking. The authors predicted each key point in the coordinate space of the video stream using a recurrent deep neural network algorithm.

1 Introduction

An intelligent system in production must warn people of potential risks to stop dangerous activities before they cause serious damage. The goal is broader than just recognizing dangerous activities or detecting damage caused by them. The detection and prediction task focuses on recognizing humans and their body parts with fewer video frames to improve performance. The urgency of this task stems from the need for human recognition, with the following factors in mind:
- changing environment;
- insufficient computing power;
- rapid learning and retraining of the recognition and prediction model;
- unpredictability of human behavior.

We are currently observing research in human pose recognition (determination of body position) from images and video streams. There are models to identify and classify key points of the human body (Human Pose Estimation, HPE), such as, traditional [1-3] and modern methods (State-of-the-art methods, SOTA) based on deep learning: ConvNet [4], DeepPose [5], DeepCut [6], AlphaPose (RMPE) [7, 8], OpenPose [9], HigherHRNet [10], BAPose [11], Paddle Paddle [12], etc.

The pictorial structure framework (PSF) is a traditional method introduced by Fischler and Elschlager [13] to determine the human body position based on body parts. This method works when body parts are shown clearly and not hidden by image objects. This method has developed with contour detection algorithms such as the Histogram Oriented Gaussian (HOG), but it still has low accuracy.

With the increased interest in deep learning, scientists have solved the problem of determining the position of the human body using deep neural networks (DNNs).

* Corresponding author: osa18@yandex.ru

Convolutional network has been successful in machine vision tasks with a sufficient training dataset.

Since 2014, works on single-person pose estimation in images using deep neural networks have appeared. Work [4] proposes a new hybrid architecture comprising a convolutional network (ConvNet) and Markov Random Field (MRF) to get an articulated estimation of human pose on monocular images.

The idea of simultaneous recognition and determination of people's body position in images using GNS was developed in 2016. The work [6] proposed the DeepCut method, which performed bottom-up body position detection with a gradual recognition of all body parts in the image and their marking (head, arms, legs, and so on).

Work [7] proposes a method of Regional Multi-Person Pose Estimation (RMPE or AlphaPose), which is implemented "Top-Down Method" to eliminate inaccuracies in identifying people by key points of the human body. The authors based their solution to the problem of recognition of people on a two-step approach, based on two deep neural networks: SSTN and SPPE. Such methods are based on a separate human detector and must evaluate each person's pose individually. They require large computational resources and are not truly end-to-end systems.

Zhe Cao et al. [9] introduced the OpenPose method in 2019. The method features a bottom-up technique in which the VGG-19 convolutional deep neural network first determines the 2D poses of several individuals, including an extended number of key points for face, body, arm, hips, and feet, and then maps the key points to form pairs of related body points.

The High-Resolution Network (HRNet) has shown excellent results in both top-down [14, 15] and bottom-up [10] approaches. HRNet applied this deep neural network in BAPose, PaddleDetection models as a basis for feature extraction and comparison for many subsequent models for human body position determination in the video stream in top-down and bottom-up approaches. HRNet deep neural network is used in BAPose, PaddleDetection models for initial feature extraction, heatmap processing, and also for comparison of model results.

All research has focused on image and video stream processing to achieve accuracy in identifying key points of the human body. There are some works [16, 19] on predicting the movements of individual human parts and hits in the forbidden restricted area [20, 21]. Most of the research works reflect the tendency to focus on achieving the accuracy of keypoint detection and omit the problem of predicting body position.

Despite the existence of many works on the estimation by deep neural networks of the human body pose or joint localization of human joints on images and video stream, the following problems remain:

- consideration of joint interdependence to build a predictive trajectory of the body position;
- the consideration of the shape of the body, volumetric clothing;
- the influence of illumination and viewing angles on the recognition accuracy;
- overlapping of individual body parts by objects.

The article aims to develop and investigate an algorithm for predicting human behavior in the workspace for a time interval, considering the previous state and connection of key points of the human body.

2 Setting the problem of predicting a person's position to prevent him from invading a forbidden space

Considering the workspace W, in which a person carries out his activities. Access to a part of the space (forbidden) is restricted to a person. When predicting human dynamics to predict the entry of parts of the body into the forbidden zone, we will use the model of key points of

the human body. Let us introduce a vector of classes for this purpose $\mathbf{b}=[\beta_1,\ldots,\beta_\eta,\ldots,\beta_N]$, containing N classes of key points. Its elements are classes of key points for body parts like the nose, left and right eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles.

Let there be: a set of images of a person $\omega\in\mathbf{\Omega}$, set by the attributes X_u, $u=\overline{1,U}$, whose totality for the image $\omega\in\mathbf{\Omega}$ is represented by vector descriptions $\mathbf{\Phi}(\omega)=(X_1(\omega), X_2(\omega), ..., X_U(\omega))=\mathbf{X}$; the set of classes $\mathbf{B}=\{\beta_1,\ldots,\beta_\eta,\ldots,\beta_N\}$. The learning set (dataset) represents a priori information $\boldsymbol{D}=\{(\mathbf{X}^j,\mathbf{b}^j)\}$, $j=\overline{1,L}$, which is given by a table, each row j of which contains a matrix description of the image (2D or 3D image of a person) $\omega\in\mathbf{\Omega}$ and class vector $\mathbf{b}=[\beta_1,\ldots,\beta_\eta,\ldots,\beta_N]$. We note that the training set characterizes the unknown mapping $^*\mathbf{F}:\mathbf{\Omega}\to\mathbf{B}$ [22].

It is required by the frames $\mathbf{I}_t$ of the continuous video stream $\mathbf{V}=[\mathbf{I}_1,...,\mathbf{I}_t,...,\mathbf{I}_\tau]$ and a priori information given by the training set for deep training of neural networks with a teacher, to solve the problem of detection, recognition, and prediction of the behavior of images:

1) detect images – individual parts of the human body on the frames as an assessment of features $\tilde{\mathbf{X}}$ with the help of deep neural networks that implement mapping $\mathbf{F}_1:\mathbf{I}_t\to\tilde{\mathbf{X}}$ [22, 23], and classify them using the mapping $\mathbf{F}_2:\tilde{\mathbf{X}}\to\mathbf{b}$, according to the criterion $\mathrm{P}(\tilde{\mathbf{X}})$, minimizing the probability of classification error;

2) make a prediction of the coordinates of individual body parts using time sequences of coordinate measurements got from frames of a continuous video stream;

3) check the occurrence of predicted key points in the forbidden zones;

4) change human behavior scenarios to prevent undesirable actions.

3 Solving the problem of predicting the position of a person to prevent his invasion of the forbidden space with a separate procedure for selecting key points

The proposed solution method contains the following steps:

1. Separating from a continuous video stream $\mathbf{V}=[\mathbf{I}_1,...,\mathbf{I}_t,...,\mathbf{I}_\tau]$ frame $\mathbf{I}_t$ size $w^{\mathbf{I}_t}\times h^{\mathbf{I}_t}$ where t – current frame number.

2. Searching for images of a person on the frame $\mathbf{f}_1:\mathbf{I}_t\to\mathbf{G}_t$, where $\mathbf{G}_t$ – an array of elements containing parameters n objects in the frame of the video stream. Deep neural networks PP-PicoDet, VGG-19, HRNet, YOLO can detect the human body with success [14, 22, 23, 24, 25, 26]. If the sought object ω is present, we perform the transition to the next step, otherwise, the next frame is taken.

3. The Region of Interest (ROI) of the first level $\mathbf{R}^{\langle 1\rangle}=crop(\mathbf{I}_t,x^0,y^0,w^0,h^0)$, where x^0,y^0 – center coordinates of the ω-th object, w^0,h^0 – its dimensions, *crop* – cutting operation from $\mathbf{I}_t$ the submatrix by coordinates $(x^0-w^0/2,y^0-h^0/2)$, $(x^0+w^0/2,y^0+h^0/2)$.

4. Refinement of the area of interest for detailed information about the images of individual parts of a person $\mathbf{f}_2:(\mathbf{R}^{\langle 1\rangle},t)\to\mathbf{R}^{\langle 2\rangle}$.

5. Performing preprocessing of the region of interest $(\mathbf{R}^{\langle 2^* \rangle} = \mathbf{f}_3(\mathbf{R}^{\langle 2 \rangle}, \mathbf{M}, \mathbf{g})$, where $\mathbf{M}$ – matrix of geometric linear and affine transformations $\mathbf{R}^{\langle 2 \rangle}$, $\mathbf{g}$ – a set of matrix functions and their parameters for brightness and contrast transformations $\mathbf{R}^{\langle 2 \rangle}$. Each image undergoes a series of data magnification operations, including random rotation ([−30°, 30°]), random scale ([0.75, 1.25]), and random rollover to improve recognition efficiency.

6. Allocation of informative features $\mathbf{R}^{\langle 2^* \rangle}$ by extracting features from a pre-trained deep neural network (for example, Lite-HRNet [27]): $\mathbf{f}_4 : \mathbf{R}^{\langle 2^* \rangle} \to \widetilde{\mathbf{X}}$, where $\widetilde{\mathbf{X}}$ – region of interest translated into feature space (feature map evaluation). DNNs architectures pre-trained on both the COCO dataset [28] and on their own prepared datasets $\mathbf{D}$.

7. Attributing a feature vector to one class $\mathbf{f}_5 : \widetilde{\mathbf{X}} \to \mathbf{p}_{\widetilde{\mathbf{X}}}$, where $\mathbf{p}_{\widetilde{\mathbf{X}}}$ – is a vector of size $N \times 1$, containing classification probabilities, N – number of classes. We define the classification criterion as $J(\mathbf{f}_5) = \max \mathbf{p}_{\widetilde{\mathbf{X}}}$. If $J(\mathbf{f}_5) \geq \varepsilon$, where ε – is a threshold, then $\beta_\eta = \arg\max_{\eta \in 1..N} (\mathbf{p}_{\widetilde{\mathbf{X}}_\eta})$, otherwise the classification is considered wrong. The researchers summarized the accuracy characteristic of the method from the characteristics of the used algorithms and calculated it using traditional metrics [29].

The results of the stage are time sequences of measuring the position N of key points of the human body in the plane Oxy for 2D or in space $Oxyz$ for 3D images of a person.

8. Prediction of the human body position. Let us plan a prediction problem and consider its solution using time series for human key points.

Let the vector of the state of the human body, considering its key points, be a random sequence $\boldsymbol{\lambda}_i = [(\boldsymbol{\lambda}_i^1)^T, \ ...,(\boldsymbol{\lambda}_i^\eta)^T..., \ (\boldsymbol{\lambda}_i^N)^T]^T, \ i = 0, 1, ...$. The state vector of each key point is a random sequence $\boldsymbol{\lambda}_i^\eta = [\lambda_{1i}^\eta, \ ...,\lambda_{\mu i}^\eta..., \ \lambda_{Mi}^\eta]^T, \ \eta = \overline{1.N}$.

Required, using a random sequence of measurements of the human body position η at key points $\boldsymbol{\xi}_k = [(\boldsymbol{\xi}_k^1)^T, \ ...,(\boldsymbol{\xi}_k^\eta)^T..., \ (\boldsymbol{\xi}_k^N)^T]^T, \ \boldsymbol{\xi}_k^\eta = [\xi_{1k}^\eta, \ ...,\xi_{qk}^\eta..., \ \xi_{Qk}^\eta]^T, \eta = \overline{1.N}$, to find the best estimate $\widetilde{\boldsymbol{\lambda}}_{i/k}$, which minimizes the RMS criterion $J_{i/k} = E[(\boldsymbol{\lambda}_i - \widetilde{\boldsymbol{\lambda}}_{i/k})^T (\boldsymbol{\lambda}_i - \widetilde{\boldsymbol{\lambda}}_{i/k})]$. Here E – a symbol of the mathematical expectation. Note that the case $k > i$ corresponds to assessing the state forecast.

Optimal evaluation $\widetilde{\boldsymbol{\lambda}}_{i/k}$ let's define it as $\widetilde{\boldsymbol{\lambda}}_{i/k} = \varphi_i(\boldsymbol{\xi}_k)$. Thus, to solve the problem of estimating the prediction of human body position, it is necessary to find a nonlinear function $\varphi_i(\bullet)$ in the general case in some rational way [30]. We can choose the current coordinates, their velocities, and accelerations as elements of the state vector of the key point when finding a prediction estimate in the Cartesian coordinate system. For example, the coordinate x for the time moment t_i would be the current coordinate, its velocity, and acceleration x_i, v_i^x, a_i^x . The selection is similar for the coordinates y and z .

The Bayesian and non-Bayesian approaches, the least-squares method can be chosen to solve the problem of forecast estimation $\widetilde{\boldsymbol{\lambda}}_{i/k} = \varphi_i(\boldsymbol{\xi}_k)$ [31]. Within each method, the prediction estimate $\widetilde{\boldsymbol{\lambda}}_{i/k} = \varphi_i(\boldsymbol{\xi}_k)$ can be represented as $\widetilde{\boldsymbol{\lambda}}_{i/k} = \varphi_i(\boldsymbol{\xi}_k, \mathbf{w})$, where the nonlinear mapping $\varphi_i(\bullet)$ is parameterized by the vector $\mathbf{w}$. Machine learning using neural networks, fuzzy systems, wavelets, and their combinations can find $\mathbf{w}$ [32, 33].

9. The entry of a person into the forbidden zone is determined by checking the inequalities $\tilde{x}_k^{\eta} < x^f$, $\tilde{y}_k^{\eta} < y^f$, $\tilde{z}_k^{\eta} < z^f$, where $\tilde{x}_k^{\eta}, \tilde{y}_k^{\eta}, \tilde{z}_k^{\eta}$, $\eta = \overline{1.N}$ – the predicted coordinates of the key points of the person at the time t_k, a x^f, y^f, z^f – the boundary coordinates of the forbidden zone.

4 Example

The prediction of coordinates for a time interval, considering the previous state and connection of key points of the human body, we can estimate using deep neural networks of Long Short-Term Memory (LSTM) [34] by time series.

Dataset. We will use the pre-trained PP-TinyPose PaddleHub model on the COCO dataset to determine the coordinates of the key points. The COCO dataset [28] contains over 200 thousand 2D images and instances of 250 thousand people labeled with 17 key points (most people in COCO at medium to large scales). We trained the PP-TinyPose PaddleHub model on the COCO dataset, with the train2017 dataset (includes 57k images and 150k instances of people) and tested on val2017 (includes 5k images) and test-dev2017 (includes 20k images). Each key point is in the format $[x_1, y_1, v_1, ..., x_i, y_i, v_i]$ i.e., it has an indexed location x_i, y_i and a visibility feature (reliability) v_i.

Model. The PaddleDetection model supports keypoint detection with two top-down and bottom-up solutions. The top-down approach first detects the main body and then the local key points. This solution has higher accuracy but takes longer as the number of detected objects increases. The bottom-up approach first detects the key points and then combines them with the corresponding parts. It is fast and its speed does not depend on the number of detected objects.

Let us consider an example of implementing the proposed method of predicting human actions in the working and forbidden space for a time interval, considering the previous state and connection of key points of the human body. We processed video clips of duration 3 with a resolution of 1080x1920, mp4 format, and at 30 fps to show the work of the method.

Stage 1. Applies the PP-TinyPose PaddleHub model with a real-time top-down approach.

Stage 2. An "ultralight" PP-PicoDet neural network is used for human body detection [24].

Stage 3. We normalized all detected people to about the same scale by cropping and resizing the bounding rectangles.

Stage 4. Refine the area of interest in detailing the image information. Localization and cropping from frames of 60×60 px images of individual body parts takes place: face, ears, shoulders, elbows, wrists, hips, knees and ankles (Fig. 1).

a) b) c)

Fig. 1. Examples of splitting a frame: (a) ear; (b) wrist; (c) shoulder.

Stage 5. Each image undergoes a series of data magnification operations, including random rotation ([-30°, 30°]), random scale ([0.75, 1.25]), and random flip.

Stage 6. An efficient high-resolution Lite-HRNet neural network [27] is used for two-dimensional pose estimation and localization of key human anatomical points (e.g., elbow, wrist, hip, knee, etc.) or parts. Lite-HRNet is pre-trained on 8 NVIDIA V100 GPUs with a mini-package size of 32 GPUs, using an Adam optimizer with an initial learning rate of $2e^{-3}$. The model provides a high accuracy of keypoint detection when expanding the dataset.

We evaluate the heat maps of individual body parts got by Lite-HRNet NS to show the validity v the location of the key point. We evaluate the Heatmaps TopDown_DARK using a post-Gaussian filter, and average the predicted heatmaps of the original and inverted image. The coordinates of the key points are 0 "Nose", 1 "Left Eye", 2 "Right Eye", 3 "Left Ear", 4 "Right Ear", 5 "Left Shoulder", 6 "Right Shoulder", 7 "Left Elbow", 8 "Right Elbow", 9 "Left Wrist", 10 "Right Wrist", 11 "Left Thigh", 12 "Right Thigh", 13 "Left Knee", 14 "Right Knee", 15 "Left Ankle", 16 "Right Ankle" (Fig. 2).

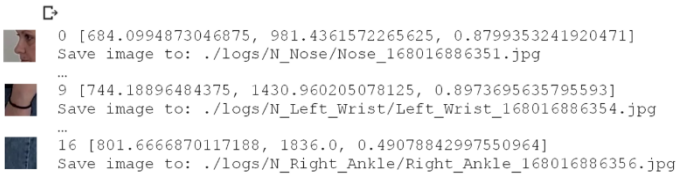

Fig. 2. An example of dividing the frame into fragments with the coordinates of the key points.

Stage 7. The standard means Average Precision (mAP) and mean Average Recall (mAR) help in evaluating the efficiency of the classification algorithm.

The accuracy specifications of the model for determining key points on the video clips with a duration of 3 seconds showed average accuracy values of mAR=89% and completeness of mAR=83%.

We emphasize the errors are because the overlap of the right parts of the body occurs in the studied video clip and the coordinates are relative to the location of the left parts, considering the physiological features of the human skeleton. For error-free classification, it is recommended that a person use a stricter form of clothing with distinctive attributes of the left and right parts of the body.

Stage 8. The input video signal is used for prediction by splitting the video into frames. The advantage of LSTM-based recurrent networks is the support for multiple parallel input sequences as multidimensional input data, which distinguishes them from other models.

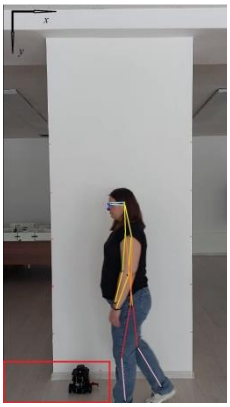

Fig. 3. Construction of the skeleton by the key points of the human body.

Temporal coordinate sequences for 17 key points of the human body (as in Fig. 3) in the plane Oxy are fed into the recurrence network with the LSTM layer.

To predict the values of future time-steps of a sequence for a time interval given the previous state, it is necessary to pre-train deep LSTM neural networks on sequences where the responses are training sequences with values shifted by one or more time-steps on COCO, Sub-JHMDB, and/or PennAction datasets. The LSTM network learns to predict the value for the next or more subsequent steps at each time step of the input sequence.

We used a recurrent NS with multiple layers in our simulations: a hidden LSTM layer, with the number of neurons varying from 200 or more; the LSTM layer is followed by a fully connected layer with 200 neurons to interpret the detected features; the dimensionality of the output layer is determined by the number of prediction items used, one for each prediction time interval.

Fig. 4 shows the comparison of the predicted coordinate values at 5-time steps with the test values. We found that the root-mean-square error of the predicted position of the selected key point, the left ankle, with a prediction for a time interval of 1 s can be reduced to 1 cm (up to 37 pixels) in real physical space (plane) for the considered experiment.

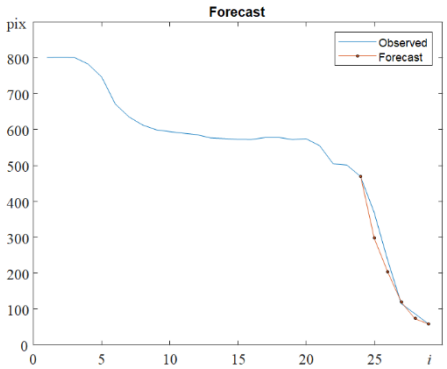

Fig. 4. Forecast by coordinate x .

5 Conclusion

The paper poses the problem of predicting the position of a person to prevent his invasion of forbidden space while trajectory following him, considering individual body parts. We have developed a method for recognizing human body parts in a continuous video stream using PP-PicoDet deep neural networks, Lite-HRNet. The human position prediction uses a separate procedure to extract key points and then transfers the coordinates into temporal sequences and processes them using LSTM deep neural networks.

The paper considered an example of human position prediction from a continuous video stream in indoor trajectory tracking. The resulting trajectories for a set of experimental videos were computed and evaluated using metrics that showed a positive effect on optical flow rate estimates. We can use our algorithm as a preliminary step for further joint processing of trajectories, such as object action recognition. We have implemented all neural network algorithms as software modules in the Python language.

We have proved, through field experiments, the effectiveness and applicability of the proposed method for solving the problem of human position prediction in various subject areas in real-time.

References

1. P. F. Felzenszwalb, D. P. Huttenlocher, Int. J. of Computer Vision **61**, 55–79 (2005)
2. A. Schick, R. Stiefelhagen, “3D Pictorial Structures for Human Pose Estimation with Supervoxels”, in *IEEE Winter Conference on Applications of Computer Vision* (2015)
3. B. Sapp, B. Taskar, “Modec: Multimodal decomposable models for human pose estimation”, in *CVPR* (2013)

4. J. Tompson, A. Jain, Y. LeCun, Ch. Bregler, “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”, in *NIPS* (2014)
5. A. Toshev, Ch. Szegedy, “DeepPose: Human Pose Estimation via Deep Neural Networks”, in *IEEE Conference on Computer Vision and Pattern Recognition* (2014)
6. L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, B. Schiele, “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation”, in *IEEE Conference on Computer Vision and Pattern Recognition* (2016)
7. H.-S. Fang, S. Xie, Y.-W. Tai, C. Lu, *RMPE: Regional Multi-person Pose Estimation* (2018)
8. H.-S. Fang, J. Li, H. Tang, Ch. Xu, H. Zhu, Yu. Xiu, Y.-L. Li, C. Lu, *AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time* (2022)
9. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”, in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019)
10. B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation”, in *CVPR* (2020)
11. B. Artacho, A. Savakis, “BAPose: Bottom-Up Pose Estimation with Disentangled Waterfall Representations”, in *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops* (2023)
12. PaddleDetection, Object detection and instance segmentation toolkit based on PaddlePaddle, https://github.com/PaddlePaddle/PaddleDetection (2019)
13. M. A. Fischler, R. A. Elschlager, IEEE Transactions on Computer **22(1)**, 67–92 (1973)
14. K. Sun, B. Xiao, D. Liu, J. Wang, “Deep high-resolution representation learning for human pose estimation”, in *CVPR* (2019)
15. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, “Deep high-resolution representation learning for visual recognition”, in *CoRR* (2019)
16. D. Sur´ıs, R. Liu, C. Vondrick, “Learning the Predictability of the Future”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021)
17. M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, B. Schiele, “PoseTrack: A Benchmark for Human Pose Estimation and Tracking”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018)
18. A. Khelvas, A. Gilya-Zetinov, E. Konyagin, D. Demyanova, P. Sorokin, R. Khafizov, Advances in Intelligent Systems and Computing **1251**, 10–22 (2021)
19. Y. Xiu, J. Li, H. Wang, Y. Fang, C. Lu, “Pose flow: efficient online pose tracking”, in *British Machine Vision Conference* (2018)
20. A. Antonucci, V. Magnago, L. Palopoli, D. Fontanelli, “Performance Assessment of a People Tracker for Social Robots”, in *IEEE Instrumentation and Measurement Society* (2019)
21. J. Docekal, J. Rozlivek, J. Matas, M. Hoffmann, “Human Keypoint Detection for Close Proximity Human-Robot Interaction”, in *IEEE-RAS 21st International Conference on Humanoid Robots* (2022)
22. O.S. Amosov, S.G. Amosova, S.V. Zhiganov, Yu.S. Ivanov, F.F. Pashchenko, J. Of Comp. And Systems Sciences Int. **59(5)**, 712–727 (2020)

23. O.S. Amosov, S.V. Zhiganov, Yu.S. Ivanov, Information Technology in Industry **6(1)**, 14–19 (2018)
24. G. Yu, Q. Chang, W. Lv, Ch. Xu, Ch. Cui, W. Ji, Q. Dang, K. Deng, G. Wang, Y. Du, B. Lai, Q. Liu, X. Hu, D. Yu, Y. Ma, "PP-PicoDet: A Better Real-Time Object Detector on Mobile Devices", in *Computer Vision and Pattern Recognition* (2021)
25. K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", in *ICLR* (2015)
26. YOLOv5 in PyTorch. Available online: https://github.com/ultralytics/yolov5
27. Ch. Yu, B. Xiao, Ch. Gao, L. Yuan, L. Zhang, N. Sang, J. Wang, *Lite-HRNet: A Lightweight High-Resolution Network, Computer Vision and Pattern Recognition* (2021)
28. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, "Microsoft COCO: Common objects in context", in *European Conference on Computer Vision* (2014)
29. O.S. Amosov, S.G. Amosova, Y.S. Ivanov, S.V. Zhiganov, "Using the deep neural networks for normal and abnormal situation recognition in the automatic access monitoring and control system of vehicles", in *Neural Computing & Applications* (2020)
30. O.A. Stepanov, O.S. Amosov, IFAC Proceedings Volumes (IFAC-PapersOnline) **37(12)**, 213-218 (2004)
31. O.S. Amosov, Proceedings of the 2nd International Conference on Intelligent Control and Information Processing (ICICIP, 2011) **6008233**, 208-213 (2011)
32. O.S. Amosov, S.G. Baena, IEEE International Conference on Control and Automation **8003045**, 118–123 (2017)
33. O.S. Amosov, S.G. Amosova, Y.S. Ivanov, S.V. Zhiganov, Procedia Computer Science **150**, 532–539 (2019)
34. O.S. Amosov, S.G. Amosova, I.O. Iochkov, "Deep Neural Network Recognition of Rivet Joint Defects in Aircraft Products", in *Sensors* **22(9)**, 3417 (2022)