Estimating Position of Multiple People in Common 3D Space via City Surveillance Cameras

Igor Ryabchikov
ITMO University
St.Petersburg, Russia
i.a.ryabchikov@gmail.com

Nikolay Teslya
SPC RAS
St.Petersburg, Russia
teslya@iias.spb.su

***Abstract*—Automatic detection of dangerous situations to ensure the safety of residents is one of the areas of Smart Surveillance. Often dangerous situations are caused by deviant behavior of people (vandalism, brawl, robbery, etc.). An important technology for analyzing the actions and interactions of people in the context of detecting deviant behavior is 3d human poses estimation, but the estimation of the relative position of multiple people in 3d space, which is necessary to analyze the interaction of people, is a separate task that remains outside the scope of this technology. In this paper, we propose an approach for estimating the relative position of 3d poses of people based on the surface normals detection in plain RGB images. The approach was tested using a computer-graphic based dataset containing scenes of interacting people in a city. The results were compared with the existing approach for estimating the relative position of 3d poses of people, based on the assumption of the constant human skeleton length.**

I. Introduction

Together with the Smart City concept, the Smart Surveillance concept which implies the use of intelligent technologies for analyzing video surveillance systems data is attracting more and more attention. One of the main areas of the Smart Surveillance concept is the detection of dangerous situations to ensure the safety of residents. Examples of tasks solved for this purpose are the recognition of wanted criminals [1], the detection of orphan objects [2], the detection of weapon [3] and the detection of fire and smoke [4].

Often a dangerous situation is caused by actions of people (robbery, vandalism, assault, brawl, etc.). The prompt detection of such situations allows taking timely measures to eliminate them and help victims, but due to the complexity of the problem of understanding human interaction by a computer the problem of automatic detection of dangerous situations caused by deviant behavior of people remains unsolved.

A possible approach to solving this problem is the integration of modern computer vision and knowledge management technologies, proposed in [5]. It's refined diagram depicting the proposed integration is shown in Fig. 1. Technologies such as segmentation and classification of objects and people [6], tracking of people [7], 3d human skeleton detection in RGB image [8], as well as classification of short-term human actions of people [9] allow recognition of fine grained features of scenes that reflect the events taking place on the video and serving as atomic knowledge about the observed scene. While knowledge management technologies ensure the incorporation of this atomic knowledge with expert knowledge about the human deviant behavior scenes, allowing the computer to build chains of reasoning and to suggest whether a particular video segment depicts a dangerous situation or not.

To detect dangerous situations caused by deviant behavior of people, the most important features are 3d poses of people. The estimation of 3d poses of people in a common 3d space (for example, in the camera coordinate system) for each frame of the video allows the recognition of a wide range of actions and interactions of people [10], for instance, punching, pushing, falling, shaking hands or searching a person's pockets. However, to obtain coordinates in camera space we need to know the distance from the camera to each person. Determining the distance is a trivial task using special equipment such as a LiDAR, but usually cameras of a city's video surveillance systems provide only plain RGB images. At the same time, the existing technologies for estimating 3d poses from a 2d image (for example, [8]) measure the depth coordinate of human pose keypoints from the central point of the human body considering the detection of the absolute depth in camera space as a separate task.

There are various approaches for estimating the distance from a camera to a person from a plain RGB image, which differs in the use of different heuristics. The validity of these heuristics determines the error. In this work, we propose an approach for estimating the distance from the camera to a person, which is distinguished by the following features and assumptions:

- Intrinsic camera parameters (focal length, matrix size and lens distortion coefficient) are known.
- It is assumed that the camera and the interacting people are located on the same plane (flat surface).
- The height of the camera above the ground plane is known.
- At each moment in time, a person touches the surface with at least one keypoint of his estimated 3d pose.
- When estimating the camera-to-person distance, the rotation angles of the camera to the surface are used, calculated using technology for detecting surfaces and normal vectors in a plain RGB image.

The above assumptions apply to city's surveillance cameras, and the required parameters of cameras can be easily discov-

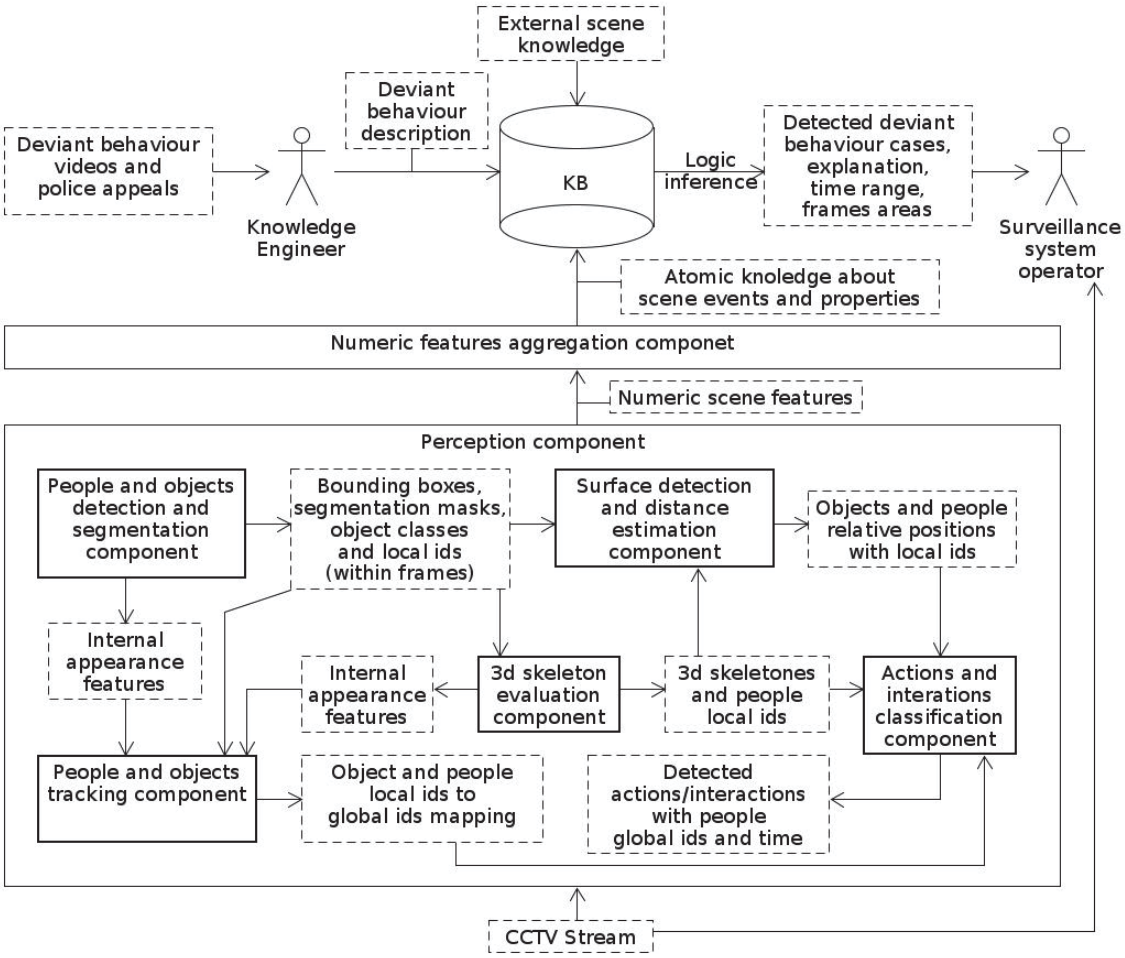

Fig. 1. Diagram of the approach for detecting deviant behavior of people via city surveillance cameras

ered.

In this work we also implemented the proposed approach, improving the system of tracking the interaction of people in 3d space, developed in [11]. Based on conducted experiments using the dataset developed within [11], we have shown that the proposed approach significantly reduces the error in detecting the relative position of people, in comparison with the previously used approach based on the assumption of the constant length of a person skeleton.

This work is a continuation of the work [5] and [11], aimed at developing the automatic deviant behaviour detection approach and integrating people detection, people tracking and pose estimation technologies. This work is focused only on the human position estimation in 3d space as a step towards automatic deviant behaviour detection, and the results obtained in this work will be used for the development of the Surface detection and distance estimation component shown in Fig. 1.

The rest of the paper is structured as follows. Section II provides an overview of technologies for estimating depth from a plain RGB image and technologies for estimating normal vectors of surfaces. Section III provides a detailed description of the proposed approach for estimating the coordinates of a human body keypoints in camera space. Section IV presents the results of testing the proposed approach.

II. Related Work

Detecting the distance from the camera to objects using plain RGB images is an urgent task in many fields, for example, robotics, augmented reality, or 3d scene reconstruction. The need to estimate depth from RGB images arises when special equipment (for example, LiDAR) cannot be used due to the high cost or complexity of deployment and use.

Transformation of a 2d view into 3d is ambiguous, so various heuristics are used to solve this problem. For example, [12] offers advanced driver assistance system, which estimates the distance from the camera to vehicles. For this, the authors carry out the detection and classification of cars to obtain the corresponding real world dimensions, and also estimate the attitude angle of cars. Then, by correlating the actual and observed size of the car, authors calculate the required distance.

The work [13] proposes an approach for estimating the camera-to-car distance by detecting road lanes. Knowing a priori the distance between lanes and assuming that lanes are drawn straight and parallel on the road, and that the road is a plane, the authors calculate the distance from the camera to the points of contact of cars with the road.

In many cases, estimating the distance to objects from a 2d image is not a difficult task for a person. To do this, the human brain applies its experience, takes into account the textures, shadows and known patterns of the observed scenes and objects. Many modern works follow the assumption that neural networks can be trained to these patterns. For example, [14] proposes a convolutional neural network to estimate the depth of each pixel in the original RGB image. Assuming that some computer vision tasks are related to the depth estimation task and applying additional constraints and training data to a neural network can improve its accuracy, a number of works propose the creation of multitask neural networks. For example, [15] proposes a neural network for joint depth estimation and semantic segmentation.

Another approach to depth estimation and 3d view reconstruction is a planar reconstruction, implying representation of space as a set of 3d planes. For such a reconstruction, it is necessary to segment the planes, estimate their normal vectors in the camera space and estimate the position of the camera relative to the detected planes. This view is more convenient than pixelwise depth maps in some problems of augmented reality and robotics [16]. As demonstrated in [17], the segmentation of 3d planes and estimation of normals generalizes well to new datasets and only requires analysis of local areas of the image. To estimate the position of the camera relative to each plane, [17] proposes to calculate the pixelwise depthmap for the whole image, which requires a global image analysis and is more sensitive to unseen datasets.

Among the most recent works devoted to plane detection using plain RGB images ([16], [17], [18], [19], [20], [21]) the most advanced results were obtained in [17]. The work [17] proposes the PlaneRCNN neural network based on Mask-RCNN [22] (Fig. 2). First, deep features are extracted from a plain RGB image using the Feature pyramid network (FPN). These features are then used in two separate branches to detect planes and evaluate a pixelwise depth map. To detect planes, by analogy with Mask-RCNN, a selection of regions containing individual planes (Region proposal phase) is performed, after which spatial features are separately extracted for each region and reduced to a universal size (Roi pool phase). Then the features of each region are fed to the input of a neural network composed of fully-connected layers to calculate normal vectors and a neural network composed of convolution layers to estimate planar masks. The difference from Mask-RCNN is the selection of one of a seven predefined anchor vectors instead of object classes and the estimation of residual vectors for each anchor vector. Sum of an anchor vector and corresponding residual vector produces a normal vector. The estimation of a pixelwise depth map happens in a separate branch of the neural network. For that purpose a decoder is applied to the deep features of the original image, the output of which is a depth map corresponding to the size of the image. Next, the depth map is used to calculate the distance from the camera to each plane and the final depth map in the camera coordinate system. Finally, based on all the results obtained, the planar masks are refined using a neural network based on the U-Net architecture.

III. Human position estimation

To estimate human poses in a common 3d space with the aim of tracking human interaction, this paper proposes an approach consisting of the following stages:

1) Detection of bounding boxes of people.
2) Detection of a 3d pose of a person in each bounding box.
3) Estimation of the ground plane normal.
4) Selection of the closest to the ground plane keypoint for each person.
5) Calculation of the distance from the camera to the selected keypoint of each person.
6) Calculation of coordinates of 3d human poses in camera space.

To apply this approach, the following assumptions must be met:

- Camera parameters are known: focal length, sensor size and lens distortion coefficient. They can be obtained from the documentation for the camera model.
- The camera and people are placed on a single flat surface. Since usually more or less flat ground prevails in the city, this assumption is appropriate. The presence of different height of the road and sidewalk can introduce an error, but its impact on tracking of human interaction should be small if the interacting people are simultaneously standing on the same height surface.
- The height of the camera relative to the ground plane is known. This value could be documented when deploying the camera or it can be measured once for each camera.
- People must touch the surface with at least one estimated keypoint. This assumption is incorrect if the person is jumping or is standing on a bench, car, or other object. In this case, it will be considered that the person is further away, and that the human body is larger than it actually is. To solve this problem, the distance can be corrected by introducing a limit on the size of a person's body, as well as by comparing it with the size of the person's body observed in previous frames.

Detection of bounding boxes of people. There are numerous detectors of people. In this work, the implementation from [6] has been used.

Detection of a 3d pose of a person in each bounding box. For this purpose, the neural network presented in [8] was used. As an input, the neural network takes an image patch containing a single person, and as an output it provides the three-dimensional coordinates of eighteen key points of the human body. In this case, the x and y coordinates are

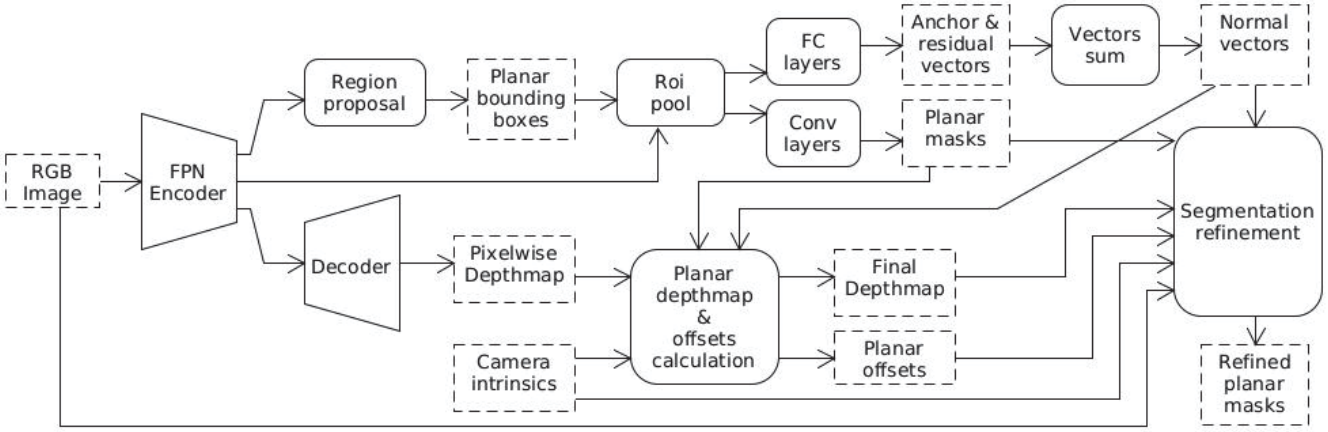

Fig. 2. A diagram of the PlaneRCNN neural network for detecting planes in images

measured from the upper left corner of the patch, and the z coordinate - from the center point of the human body. This coordinate system can be converted to the camera coordinate system by parallel translation. All coordinates are measured in pixels.

Estimation of the ground plane normal. To estimate the normal to the ground plane, we propose to use the PlaneRCNN neural network presented in [17]. The network accepts an RGB image as an input, and provides a segmentation mask and a normal vector in the camera space (Fig. 3) for each detected plane as an output. The shortcoming of the PlaneRCNN is that

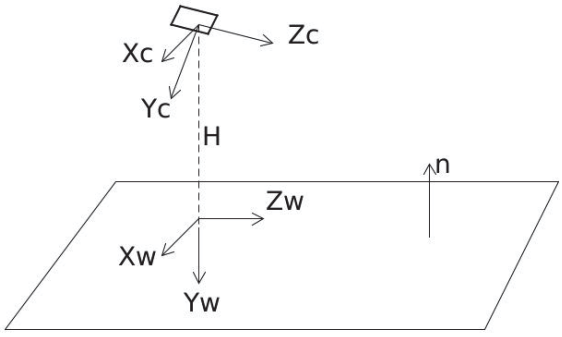

Fig. 3. An example of placing a surveillance camera relative to the ground plane

it does not distinguish the ground plane from the others. In addition, as our experiments have shown, the neural network may not segment the entire surface and perceive a person as a surface (Fig. 4). For automatic selection of the ground plane, we use several assumptions: the ground plane should occupy a large area; it must be close to a detected person; and it should be located at the bottom of the image. Based on these assumptions, we propose the formula 1, according to which we calculate the score for each plane. The plane with the highest score is selected as the ground plane. The first term in the formula 1 reflects the inherent position of the ground plane relative a person and does not exceed 1 for each person. The second term reflects the inherent absolute position of the ground plane in the image and also does not exceed 1.

$$S_k = \sum_{i,x,y} M_{k,x,y} * C_{i,y} * \frac{e^{-\frac{(x-X_i)^2+(y-Y_i)^2}{2L_i^2}}}{L_i^2 * 2\pi} + \sum_{x,y} M_{k,x,y} * \frac{2*y}{(H-1)*H*W} \quad (1)$$

where S_k - a score of the kth plane; i - a number of the detected person; x and y - coordinates of the pixel, measured from the upper left corner of the image starting at zero; $M_{k,x,y}$ - a belongingness coefficient of the (x, y) pixels to the kth plane, equals to 1 if the pixel belongs to the plane and 0 otherwise; X_i and Y_i - coordinates of the lowest relative to the image keypoint (with the biggest y coordinate) of the ith person; L_i - a leg's length of the ith person in pixels, calculated using the estimated pose; $C_{i,y}$ - coefficient equals to 0.5 if $y < Y_i$ and 1.5 if $y >= Y_i$; H and W - the height and width of the image in pixels.

Selection of the closest to the ground plane keypoint for each person. To select the keypoint closest to the ground plane, which is considered the point of contact with the surface, we rotate the coordinate system of the person's pose in accordance with the system of equations 2. As a result of rotation, the normal vector $-\vec{n}$ (Fig.4) becomes co-directional with the axis Oy of the resulting coordinate system. Thus, the keypoint with the largest value of the y coordinate becomes

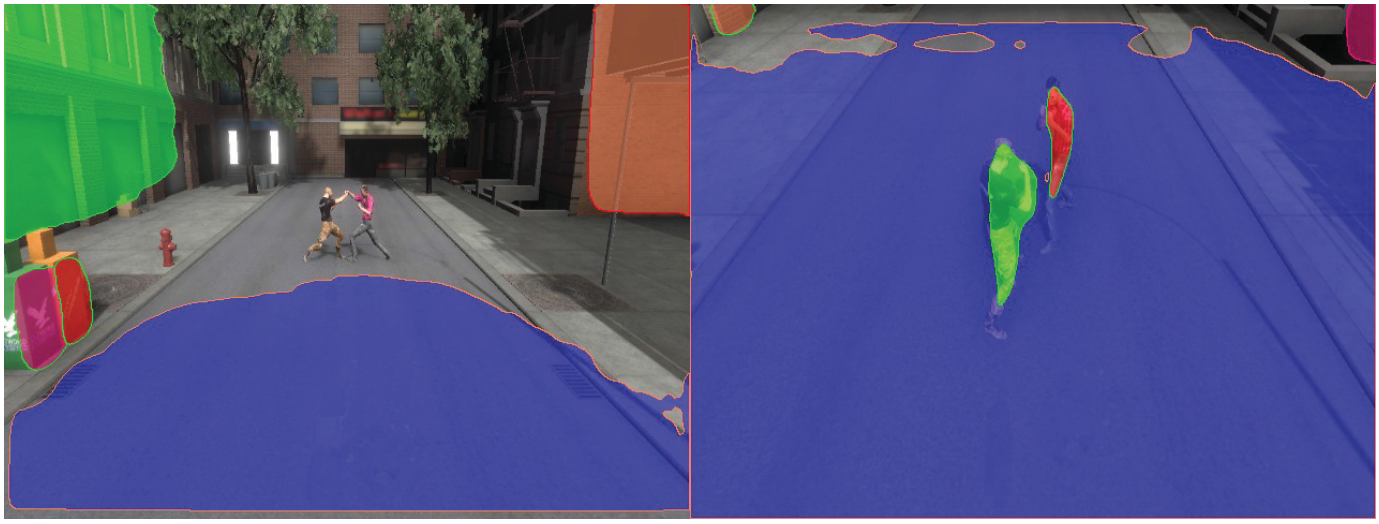

Fig. 4. Examples of planes detection using PlaneRCNN

the keypoint closest to the surface.

$$\begin{cases} \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} = (-\vec{n}) \times \vec{y_c} \\ c = (-\vec{n}) * \vec{y_c} \\ V = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix} \\ R = I + V + V^2 * \frac{1}{1+c} \\ \begin{pmatrix} X_{w'} \\ Y_{w'} \\ Z_{w'} \end{pmatrix} = R * \begin{pmatrix} X_{c'} \\ Y_{c'} \\ Z_{c'} \end{pmatrix} \end{cases} \tag{2}$$

where $\vec{n}$ - unit normal vector of the surface (Fig. 3); $\vec{y_c}$ - unit vector co-directional with axis Oy_c; I - identity matrix; $X_{c'}, Y_{c'}, Z_{c'}$ - estimated coordinates of a human keypoint; $X_{w'}, Y_{w'}, Z_{w'}$ - coordinates of the person's keypoint in the coordinate system in which the vector $-\vec{n}$ is co-directed with the Oy axis.

Calculation of the distance from the camera to the selected keypoint of each person. Since we assume that the selected keypoint lies on the surface, its y coordinate in the world coordinate system (Fig. 3) is zero. Applying the pinhole camera model [23] we derive the formula of the inverse projection of an image point into the world coordinate system. So we have a system of equations 3, from which all unknowns are uniquely found. The transformation of the obtained coordinates of the point into the camera coordinate system is carried out according to the formula 4.

$$\begin{cases} K = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ \begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} = \begin{pmatrix} 0 \\ -H \\ 0 \end{pmatrix} + \lambda * R * K^{-1} * \begin{pmatrix} (X_{c'} + X_p) * S_p \\ (Y_{c'} + Y_p) * S_p \\ 1 \end{pmatrix} \\ Y_w = 0 \end{cases} \tag{3}$$

where f - camera focal length in millimeters; X_w, Y_w, Z_w - coordinates of the person's keypoint lying on the surface in the world coordinate system, in millimeters; $X_{c'}$, $Y_{c'}$ - estimated coordinates of the person's keypoint lying on the surface, in pixels; X_p, Y_p - coordinates of the upper left corner of the image patch containing a person in the camera coordinate system, in pixels; S_p - pixel size in millimeters; R - rotation matrix from equation 2; λ - unknown coefficient; H - camera height above the ground plane in millimeters (Fig. 3).

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = R^{-1} * \begin{pmatrix} X_w \\ Y_w + H \\ Z_w \end{pmatrix} \tag{4}$$

where X_c, Y_c, Z_c - coordinates of a keypoint in the camera coordinate system; R - rotation matrix from equation 2; X_w, Y_w, Z_w - coordinates of the keypoint in the world coordinate system.

Calculation of coordinates of 3d human poses in camera space. For this we use the formula 5.

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = \begin{pmatrix} X_{c'} + X_p \\ Y_{c'} + Y_p \\ Z_{c'} - Z_{Lc'} \end{pmatrix} * \frac{X_{Lc}}{X_{Lc'} + X_p} + \begin{pmatrix} 0 \\ 0 \\ Z_{Lc} \end{pmatrix} \tag{5}$$

where X_c, Y_c, Z_c - coordinates of a person's keypoint in the camera coordinate system; $X_{c'}$, $Y_{c'}$, $Z_{c'}$ - estimated coordinates of the person's keypoint in pixels; X_p, Y_p - coordinates

of the upper left corner of the image patch containing the person in pixels; $X_{Lc'}$, $Z_{Lc'}$ - estimated coordinates of the person's keypoint lying on the surface; X_{Lc}, Z_{Lc} - coordinates of the person's keypoint lying on the surface in the camera coordinate system (formulas 3, 4).

The advantage of our proposed approach for estimating human poses in a common three-dimensional space with the aim of tracking human interaction is the generalization into unseen datasets, which is inherent in the neural network [17] we used for planes segmentation and normals estimation. In contrast to neural networks trained to directly estimate the depth.

IV. Human position estimation results

To test our proposed approach for estimating human position, we used a computer-graphic based dataset proposed in [11]. This set consists of 129,600 images of two interacting people on a city street, taken from different viewpoints. The dataset presents 4 different human models. A viewpoint is determined by three parameters: the distance from the camera to the interacting people (5, 10, 15, 20 meters), the tilt angle of the camera (10, 20, 50 degrees) and the angle of rotation of the interacting people in relation to the camera (30, 60, 90 degrees).

To estimate the error, for each image individual from the dataset absolute deviation (AD) and absolute deviation relative to the adjacent skeleton (ADAS) error metrics [11] were calculated. Since for the analysis of human interaction, only their relative position is important, before calculating the ADAS error the coordinates of human keypoints were multiplied by a coefficient calculated by the formula 6. This preprocessing is necessary to compare the ADAS error of the approach proposed in this work and the approach based on the assumption of the constant skeleton length [11].

$$r_{coef} = \frac{SL}{max(skel\_len(P_0), skel\_len(P_1))} \qquad (6)$$

where r_{coef} - resizing coefficient of a human pose; SL - constant human skeleton size, taken from [11]; P_0 and P_1 - estimated keypoints of interacting people in camera space; $skel\_len$ - function for calculating a human skeleton size by keypoints, taken from [11].

The tables I and II show the average error of all images and keypoints of people for each pair (camera distance, camera tilt angle) for two approaches. When calculating the average errors, 10% of the images with the largest ADAS error were excluded from consideration. For coordinates (x, y) and the depth coordinate z, the errors were calculated separately. Table I also shows the average error of the ground plane normal estimation, calculated as the angle between the estimated and the ground truth normals. Figures 5, 6, 7, 8 show graphs of average AD and ADAS errors versus camera distance and camera tilt angle for the normal estimation (NE) based approach and the skeleton length (SL) based approach [11]. Fig. 9 depicts a graph of the average error of the ground plane normal estimation, calculated as the angle between the estimated and the ground truth normals, versus camera tilt angle.

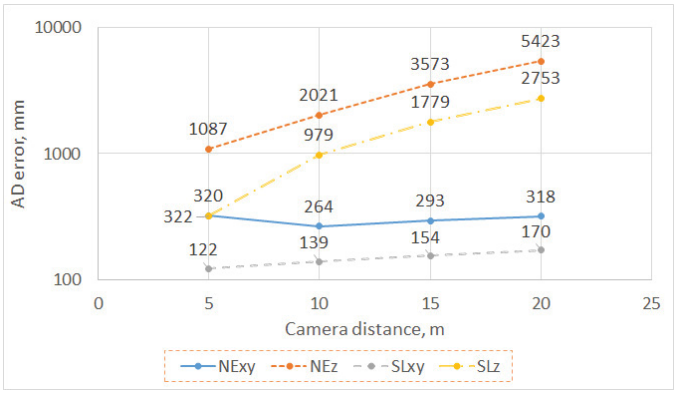

Fig. 5. AD error graphs versus the camera distance

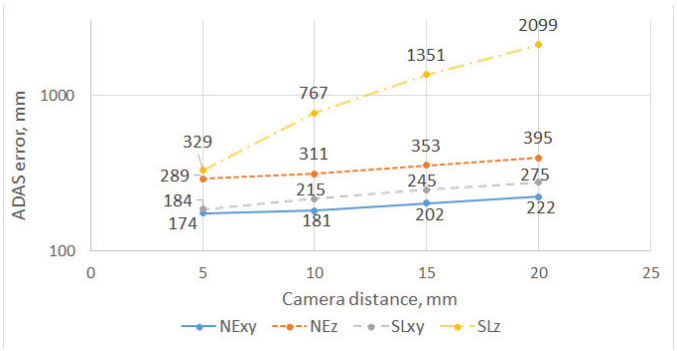

Fig. 6. ADAS error graphs versus the camera distance

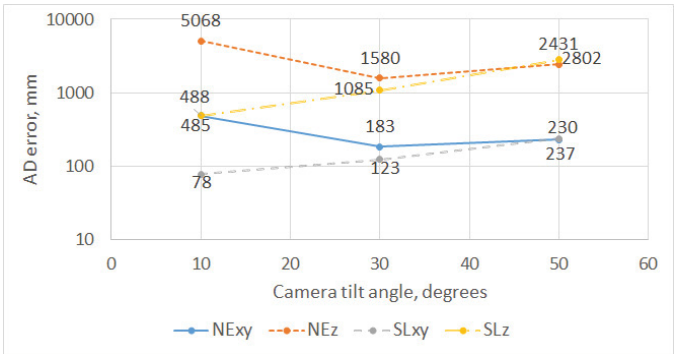

Fig. 7. AD error graphs versus the camera tilt angle

As a result of the analysis of the above data, the following observations were made:

- The error of person's location estimation in the camera coordinate system (AD) in both the normal estimation based approach and the skeleton length based approach mainly depends on the distance from the person to the camera. But in the former approach, the coefficient of proportionality is much higher, since it is caused by a high error the normal vector estimation, which can exceed 75% at low tilt angles.

TABLE I. AVERAGE AD AND ADAS ERRORS IN MILLIMETERS AND NORMAL ESTIMATION ERROR IN DEGREE FOR THE NORMAL ESTIMATION BASED APPROACH

View	(x,y)/z AD error	(x,y)/z ADAS error	Normal error
(5,10)	554/2424	139/359	10.13
(5,30)	213/471	142/239	1.17
(5,50)	193/368	241/270	1.09
(10,10)	451/3931	100/304	6.93
(10,30)	137/742	142/248	1.35
(10,50)	205/1390	301/380	5.83
(15,10)	478/6151	102/351	7.29
(15,30)	164/1690	161/279	2.72
(15,50)	238/2879	343/430	8.96
(20,10)	454/7768	108/337	6.6
(20,30)	216/3415	180/346	4.34
(20,50)	284/5086	379/502	11.71

TABLE II. AVERAGE AD AND ADAS ERRORS IN MILLIMETERS FOR THE CONSTANT SKELETON LENGTH BASED APPROACH

View	(x,y)/z AD error	(x,y)/z ADAS error
(5,10)	116/246	73/162
(5,30)	153/280	106/212
(5,50)	283/462	186/593
(10,10)	115/470	76/321
(10,30)	169/578	116/702
(10,50)	362/1254	226/1914
(15,10)	122/762	78/559
(15,30)	195/1046	130/1342
(15,50)	419/2244	254/3435
(20,10)	135/1140	87/911
(20,30)	220/1625	142/2083
(20,50)	472/3533	282/5265

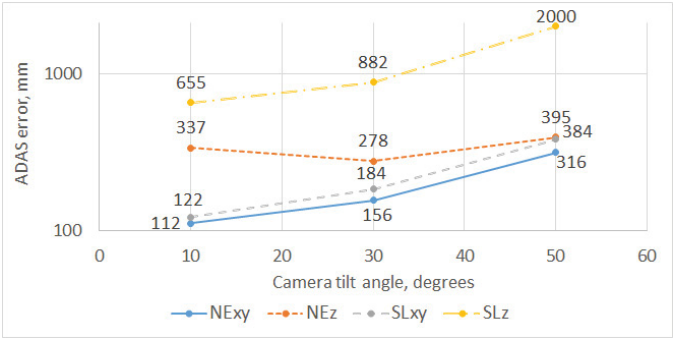

Fig. 8. ADAS error graphs versus the camera tilt angle

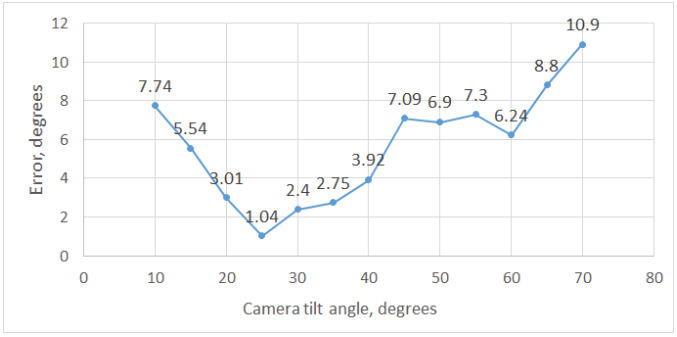

Fig. 9. A graph of the ground plane normal estimation error versus the camera tilt angle

- Despite the high AD error, the ADAS error of the normal estimation based approach is significantly lower than the skeleton length based approach. This is due to the fact that the ADAS error for the former approach depends majorly not on the distance to the camera, but on the distance between people. While for the latter approach, it depends on the distance to the camera with a coefficient of proportionality depending on the sum of errors in estimating the interacting people skeleton length. Nevertheless, the dependence of the former approach on the distance to the camera exists and is explained by the human pose estimation error, which increases with decreasing resolution of the human image [11].
- The error in estimating the ground plane normal vector is higher at low and high camera tilt angles. This may be due to the predominance of examples with a camera tilt angle close to 25 degrees in the training dataset. This is also reflected in the error in estimating the location of people of the normal estimation based approach.

Summing up, it can be concluded that the normal estimation based approach showed significantly better results in estimating the relative position of people than the skeleton length based approach. Nevertheless, the error remains high and can reach several hundred millimeters, which can affect the accuracy of the classification of human interactions. Additional

training of neural networks for estimating the pose of people and estimating the ground plane normal using datasets better reflecting the specificity of target scenes might reduce the error.

V. Conclusion

Automatic detection of deviant behavior of people via city's surveillance cameras is a promising direction in the field of ensuring the safety of residents. An important problem arising in this field is the estimation of poses of people in a common three-dimensional space. Where the poses are used to automatically detect a wide range of actions and interactions of people. Since the restoration of a 3d scene from a 2d image obtained from CCTV cameras is ambiguous, there is no a straightforward solution to this problem. In this work, we proposed an approach for estimating the distance from the camera to the 3d human pose based on the detection of the normal vector to the ground plane, which is used to determine the camera positioning. We tested this approach using a computer-generated dataset of interacting people and compared the results with the approach based on the assumption of a constant size of the human skeleton. The normal estimation based approach outperformed the constant skeleton length based approach in the problem of estimating relative position of people.

Our further work will be focused on improving the accuracy of the approach for estimating the relative position of people by creating additional training datasets that better reflect the specificity of target scenes viewed by city's CCTV cameras. As well as on the application of the obtained results to detect the deviant behavior of people.

Acknowledgment

This research has been supported by the grant of ITMO University (project № 620176) for related work analysis and results related to calculation of coordinates of 3d human poses in cameraspace, and calculation of the distance from the camera to the selected keypoint of each person. All other work has been funded by the RFBR, project 20-37-90118.

References

[1] FindFace Public Safety, Web: https://findface.pro/ru/face-recognition-public-safety.html.

[2] Smart Recognition System Reco3.26, Web: https://www.reco326.com/index.php/en/.

[3] ZeroEyes: Stop Threats at First Sight Not First Shot, Web: https://zeroeyes.com/.

[4] SmokeCatcher, Video Smoke Detection for critical environments, Web: http://www.araani.com/en/smokecatcher/.

[7] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, *Tracking Without Bells and Whistles,* in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 941–951.

[5] N. Teslya, I. Ryabchikov, and E. Lipkin, *The Concept of the Deviant Behavior Detection System via Surveillance Cameras,* in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2019, vol. 11624 LNCS, pp. 169–183.

[6] K. He, R. Girshick, and P. Dollar, *Rethinking ImageNet Pre-Training,* in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4917–4926.

[8] X. Sun, C. Li, and S. Lin, *An Integral Pose Regression System for the ECCV2018 PoseTrack Challenge*. Sep. 2018.

[9] L. Shi, Y. Zhang, J. Cheng and H. Lu, *Skeleton-Based Action Recognition With Directed Graph Neural Networks* in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7904-7913.

[10] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. -Y. Duan and A. C. Kot, *NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding* in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, vol. 42(10), pp. 2684-2701.

[11] I. Ryabchikov, N. Teslya and N. Druzhinin *Integrating Computer Vision Technologies for Smart Surveillance Purpose* in 2020 26th Conference of Open Innovations Association (FRUCT), Yaroslavl, Russia, 2020, pp. 392-401.

[12] L. Huang, T. Zhe, J. Wu, Q. Wu, C. Pei and D. Chen *Robust Inter-Vehicle Distance Estimation Method Based on Monocular Vision,* in IEEE Access, 2019, vol. 7, pp. 46059-46070.

[13] YC. Chen, TF. Su, SH. Lai *Integrated Vehicle and Lane Detection with Distance Estimation,* in Computer Vision - ACCV 2014 Workshops. ACCV 2014. Lecture Notes in Computer Science, 2014, vol 9010, pp. 473-485.

[14] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab *Deeper Depth Prediction with Fully Convolutional Residual Networks,* in 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 239-248.

[15] L. Ladický, J. Shi and M. Pollefeys *Pulling Things out of Perspective,* in 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 89-96.

[16] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa *Planenet: Piecewise planar reconstruction from a single rgb image,* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2579–2588.

[17] C. Liu, K. Kim, J. Gu, Y. Furukawa and J. Kautz *PlaneRCNN: 3D Plane Detection and Reconstruction from a Single Image,* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp.4445-4454.

[18] D. Fouhey, A. Gupta and M. Hebert *Data-Driven 3D Primitives for Single Image Understanding,* in Proceedings of the 2013 IEEE International Conference on Computer Vision, 2013, pp. 3392-3399.

[19] L. Ladický, B. Zeisl and M. Pollefeys *Discriminatively Trained Dense Surface Normal Estimation* in Proceedings of the 2014 Conference on Computer Vision, 2014, pp. 468-484.

[20] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price and A. Yuille *SURGE: Surface Regularized Geometry Estimation from a Single Image,* in Advances in Neural Information Processing Systems 29, 2016, pp. 172-180.

[21] F. Yang and Z. Zhou *Recovering 3D Planes from a Single Image via Convolutional Neural Networks,* in Proceedings of the 2018 Conference on Computer Vision, 2018, pp. 87-103.

[22] K. He, G. Gkioxari, P. Dollár and R. Girshick *Mask R-CNN,* in Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988.

[23] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2004.