Integrating Computer Vision Technologies for Smart Surveillance Purpose

Igor Ryabchikov, Nikolay Teslya
SPIIRAS
St.Petersburg, Russia
i.a.ryabchikov@gmail.com, teslya@iias.spb.su

Nikita Druzhinin
ITMO University
St.Petersburg, Russia
kleverteo@gmail.com

Abstract—**Automatic detection of dangerous situations in order to ensure the safety of residents is a new step in the development of video surveillance systems in cities. And dangerous situations are often caused by deviant behavior of people: robbery, brawl, vandalism and etc. But due to the strong variability of such scenes, their detection is a challenging problem, which still remains unresolved. The key to solving this problem is the recognition of fine-grained features and events of scenes and the application of knowledge management technologies. In this paper, three computer vision technologies for detecting people, tracking people and estimating three-dimensional human poses were integrated with the aim of recognizing the actions and interactions of people in three-dimensional space. For all technologies an open source implementations were used that showed high results in popular computer vision challenges. A dataset was also created using computer graphics to test the developed system, containing scenes of the interaction of people in the city, shot under different point of views. This dataset showed that additional teaching of the human pose estimation component to handle challenging poses of people and camera viewpoints is required.**

I. Introduction

CCTV systems are widespread in modern cities. For instance, in Chongqing (China) there are more than 2.5 million surveillance cameras, in London (UK) - more than 500 thousand, and in Moscow (Russia) - more than 170 thousand. They have had a huge impact on ensuring the safety of residents, helping authorities in solving crimes and collecting evidence. But their potential is not limited to this. Thanks to modern intelligent technologies, the video surveillance system can be used to automatically detect dangerous situations in real time. Examples of successful use of intelligent technology for this purpose at the moment are the detection of wanted criminals [1], detection of abandoned objects in public places [2] and smoke detection [3].

But the range of dangerous situations is much wider, and often a dangerous situation is caused by deviant behavior of people - actions that violate social norms or people's rights, and which can lead to the endangerment of people and property damage. Examples of a deviant behaviour are robbery, brawl, vandalism and kidnapping. The rapid detection of such situations will make it possible to prevent further escalation, provide timely assistance to victims and detain suspects. The complexity of the deviant behavior detection task is that such scenes can have strong variability and for their recognition it is necessary to take into account individual signs for a relatively long period of time (from tens of seconds to tens of minutes). For example, in a robbery, offenders can approach a pedestrian, prevent him from leaving, search, pick up valuables and escape by vehicle.

Existing approaches to detecting deviant behavior of people are mainly focused on detecting punches, falls and other short-term actions of people with distinctive patterns of velocity and acceleration, for example, [4], [5]. But to detect complex long scenes, these approaches cannot be used. In recent years, technologies for the three-dimensional human pose estimation have been actively developed [6], which potentially allow recognition of a huge range of human actions and interactions [7], for example, talking on the phone, transferring an item, searching, or punching.

In accordance with the concept proposed in [8], the detection of deviant behavior of people should be carried out by logical or probabilistic inference in accordance with the descriptions of such scenes written in a formal language. Descriptions should be obtained as a result of the process of extracting knowledge from the thematic literature, police reports and videos of offenses. The basis of descriptions will be facts of the occurrence of actions and knowledge about the joint positioning of people and objects obtained through the use of various computer vision technologies (Fig. 1). At the same time, the basic knowledge that is sufficient to detect a wide range of actions and determine joint positioning of people with certain accuracy can be obtained using technologies for detecting people [9], tracking people and objects [10] and estimating three-dimensional human poses [11].

This work continues the previous [8] and presents the next stage of the proposed concept implementation for detecting deviant behavior of people via video surveillance cameras of a city. In this paper, we have integrated open source technologies for detecting people, tracking people and estimating three-dimensional human poses, which showed high results in modern computer vision challenges. The main goal of a such integration is to track the interaction of people in three-dimensional space. A dataset have also been developed using computer graphics for testing the estimation of three-dimensional poses of people based on materials from city's surveillance cameras. This dataset shows the interaction of people filmed from different viewpoints inherent in city's surveillance cameras. Camera parameters (focal length, matrix size and resolution) correspond to the popular surveillance camera ISON AHD20F-CD.

The rest of the paper is structured as follows. Section II

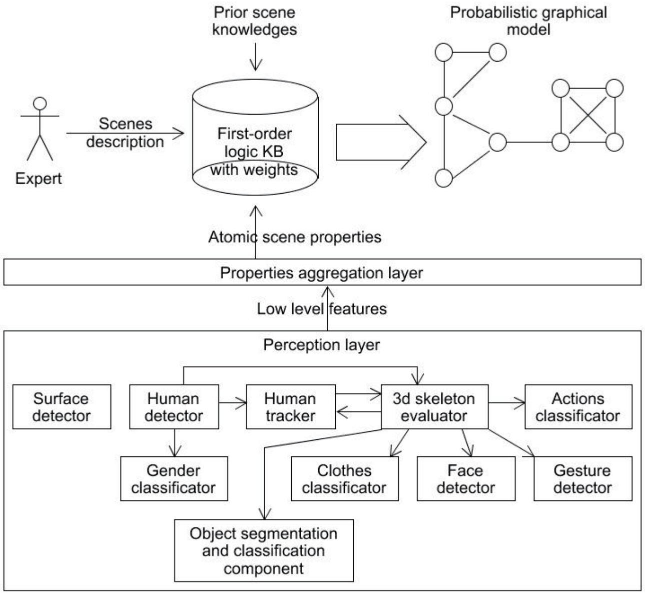

Fig. 1. Diagram of the deviant behavior detection system concept

provides a description of the used implementations of computer vision technologies. Section III presents the principles of integrating these technologies in order to track the interaction of people in three-dimensional space. Section IV describes the developed dataset for testing the three-dimensional human pose estimation. The test results are presented in section V.

II. Review of computer vision technologies

One the most well-known challenges dedicated to the tasks of detecting people, tracking people and estimating the three-dimensional human poses are COCO [12], [13], MOT [14], [15], and Human3.6M [16], [17].

One of the best results on the COCO Detection 2017 (47.7 mAP) and COCO Segmentation 2017 (41.7 mAP) datasets were achieved by the implementation of a convolutional neural network [9] based on the Mask-RCNN [18] (Fig. 2).

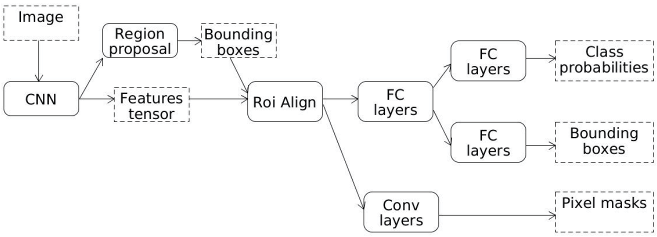

Fig. 2. Mask R-CNN architecture

Within the architecture, the initial image is fed to the backbone convolutional neural network (ResNet-101 [19]) to obtain a set of spatial features of size $W \times H \times C$, where W is the width, H is the height, C is the number of filters in the last convolutional layer. Based on the obtained features, the initial set of regions (bounding boxes) in which objects of interest (including people) can be present is determined (Region proposal stage). To do this, two convolutional layers with filter sizes 3×3 and 1×1 are applied to the feature set. The outputs of the last convolutional layer are the confidence coefficients. These coefficients shows the confidence of objects presence in regions of predefined sizes and shapes positioned in image according to the position of the extracted spatial features. The resulting regions undergo the Non-maximum suppression procedure. After the procedure the regions with the highest confidence coefficients are selected for further consideration. The following is the Roi align stage. In this stage each region is considered separately. From the set of spatial features, features that fall into the region under consideration are extracted. Then the region is divided into a fixed predetermined number of subregions of the same size, and for each subregion one single feature is calculated by performing the average pooling operation on features that fall into the subregion. Thus, for each region, a fixed number of characteristics is extracted. Further, those characteristics are fed to a neural network with fully-connected layers to obtain probabilities of an object presence in the considered region and the final coordinates of refining bounding boxes for each object class. Thus, $C * 4 + C$ features in total are presented as output, where C is the number of detected classes of objects. In parallel, the region characteristics are fed to another convolutional neural network to obtain segmentation masks (bitmaps) for each object class. At the last stage, these masks undergo a resizing procedure to match the input image.

The work [10] presents an implementation of a people tracker that showed one of the best results among open source implementations (MOTA 56.3) on the MOT 2017 dataset with a public testing protocol (bounding boxes of objects for each video frame were provided in advance). The main difference of this approach is the use of Faster R-CNN [20] (with ResNet 101 as a backbone) to move the bounding boxes of the previous frame based on the current one and the use of Siamese CNN [21] (with ResNet-50 as a backbone) to calculate the appearance characteristics of objects for their re-identification.

The correspondence of objects detected in previous frames with objects detected in the current frame is determined in several stages. First, bounding boxes detected in the previous frame are shifted based on the current frame to account for movement. For this, Faster R-CNN, trained to detect objects, is used. The current frame is fed to a convolutional neural network to calculate spatial features, than the Region proposal stage (Fig. 2) is ignored and bounding boxes detected in the previous frame are considered (Fig. 3). The following stages are performed according to the original Faster-RCNN to calculate refined bounding boxes and probabilities. If the resulting probability of an object given the current frame is too small, the object is marked as “inactive”. Inactive objects are stored in memory for some time for re-identification purpose. Then, between the refined bounding boxes and those detected in the current frame the intersection areas are calculated, based on which a decision on their correspondence is made. In case of new objects were discovered whose intersection with existing ones is too small, re-identification takes place - the appearance characteristics of inactive objects are compared with new ones. If characteristics of two objects are similar enough (in accordance with a given threshold), inactive object goes into the “active” state. If there is no inactive objects with similar characteristics, it is believed that a new object has been detected. To calculate the appearance characteristics, the Siamese CNN input is supplied with an image patch and optical flow calculated based on an adjacent frame. The network has been trained to produce appearance characteristics agnostic to viewpoints.

Thus, for differentiating people, the proposed approach uses positions of the bounding boxes on adjacent frames and appearance features when a person was lost from view due to occlusions or visual limits, and re-identification is required.

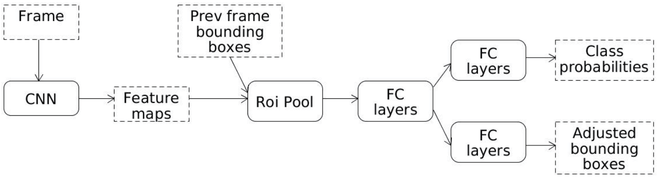

Fig. 3. Faster R-CNN architecture for refining bounding boxes of objects from the previous frame

The work [11] presents an open source implementation of a three-dimensional human pose estimator which took second place in the Human3.6M challenge with an error of 47 MPJPE. The network architecture is shown in Figure 4. The network takes as input a resized patch containing a single person. The network is based on the ResNet-152, after which several convolution and deconvolution layers are added, the output of which are volumetric heat maps for each key point. Each two-dimensional heat map is a slice of space at a certain depth. To regress the coordinates of key points the soft-argmax function (equation 1) is applied to the heatmaps.

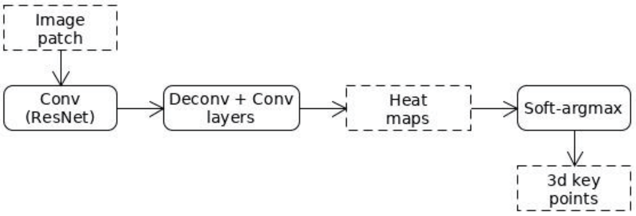

Fig. 4. Neural network architecture for estimating the coordinates of three-dimensional human body key points

$$J_k = \sum_{j \in \Omega} \frac{j * e^{H_k(j)}}{\sum_{j \in \Omega} e^{H_k(i)}} \tag{1}$$

where J_k is a three-dimensional coordinate of the k-th key point, $H_k(j)$ is a value of the volumetric heat map of the k-th key point at the position j, Ω is the set of all discrete points of the volumetric heat map.

III. Integration of computer vision technologies

A diagram of the system for tracking three-dimensional people interaction in videos developed by integrating various implementations of computer vision technologies is presented in Fig. 5.

Each frame goes to the human detection component developed on the basis of the Mask-RCNN neural network from [9]. The output of the component is a set of bounding boxes of people and a tensor of spartial features extracted with underlying ResNet-101. Since the human tracking component is based on Faster-RCNN, instead of the implementation proposed in [10], the Mask-RCNN neural network was reused as well as output features from ResNet-101. In addition to bounding boxes and features, the people tracking component uses information about tracks obtained in previous iterations: track identifier; status (active or inactive); last track frame; track deactivation time (frame number) for inactive tracks and appearance characteristics used to re-identify people. Since an optical flow is used to calculate the appearance characteristics, the component also receives the previous frame. The output of the component is the updated tracks information.

Then, based on the resulting bounding boxes, patches with detected people are cut out of the frame. Patches are fed to the component of the three-dimensional poses evaluation, the output of which is the coordinates of 18 key points of a human body. The x and y coordinates correspond to the shift from the upper left corner of the patch in pixels, and the z coordinate corresponds to the depth relative to the center of the human pelvis in pixels.

To obtain coordinates of human body key points in the camera space, coordinates are shifed and rescaled. For that purpose, coordinates of the patch are added to the x and y coordinates of each key point and half of the image width / height is subtracted. Then all coordinates are multiplied by the pixel size of the video camera matrix, and the camera's focal length is added to the z coordinate. Then each coordinate is multiplied by a scaling factor to obtain coordinates in real scale (Equation 3). To calculate the scaling factor, we assume that the length of each person's skeleton is the same and equal to a certain value calculated experimentally. Thus, the scaling factor is calculated by the Equation 2. An example of detecting three-dimensional skeletons of people using developed system is presented in Fig. 6.

$$r_{coef} = 1 + \frac{B_{len}}{bones_{len}(J) * S_{pix}} \tag{2}$$

where r_{coef} is the scaling factor, B_{len} is the experimentally calculated length of the skeleton in millimeters, J is the set of coordinates of the human body key points in pixels, $bones_{len}$ is the function for calculating the skeleton length by key points, S_{pix} is the pixel size in millimeters.

$$J_a = r_{coef} * ((J_k + (x_{patch} - \frac{w}{2}, y_{patch} - \frac{h}{2}, 0)) * S_{pix} + (0, 0, f)) \tag{3}$$

where J_a are the absolute coordinates of the human body key points in millimeters, J_k are the coordinates of the human body key points in pixels, x_{patch} and y_{patch} are the coordinates of the upper left corner of the cropped image in pixels, w and h are the width and height of the original image in pixels, S_{pix} is the pixel size in millimeters, f is the focal length of the camera lens in millimeters, r_{coef} is the scaling factor calculated by the equation 2.

To calculate the length of the human skeleton, the distances between the following key points were summed: pelvis center, thorax; right shoulder, thorax; left shoulder, thorax; right elbow, right shoulder; left elbow, left shoulder; right wrist, right elbow; left wrist, left elbow; right thigh, pelvic center;

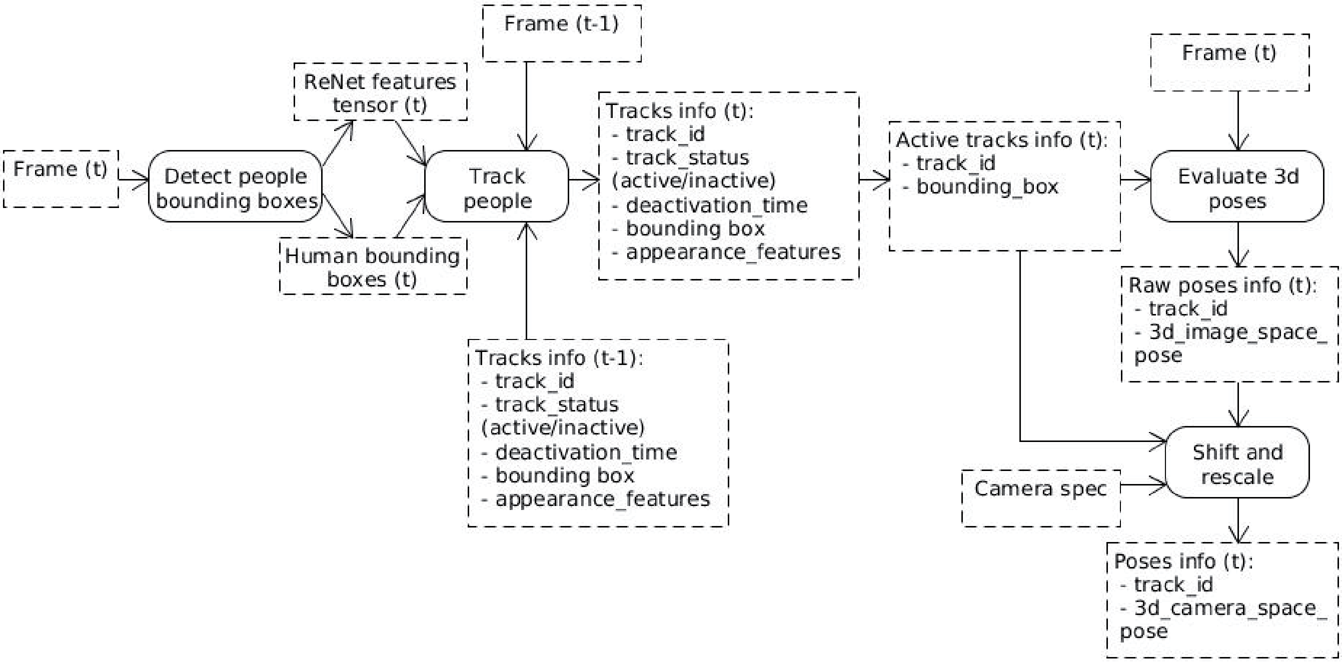

Fig. 5. A diagram of the system for tracking three-dimensional people interaction in videos

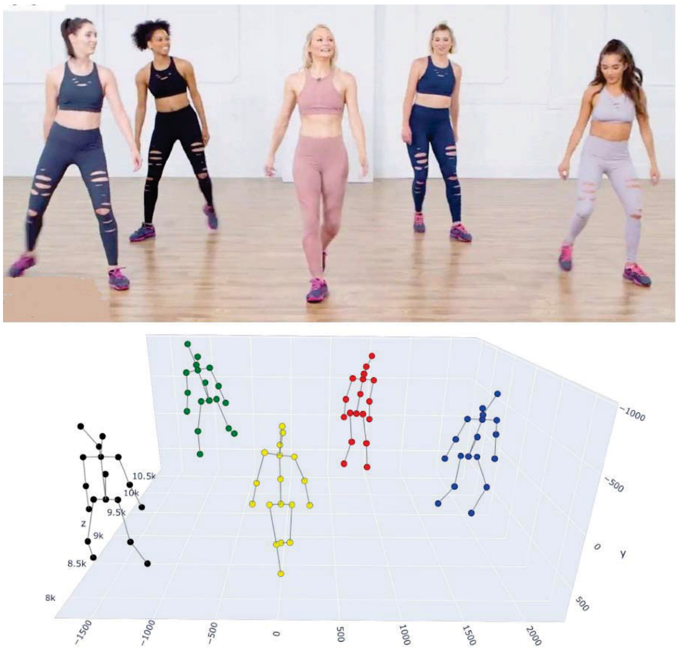

Fig. 6. The result of detecting a three-dimensional skeleton of people in the image

left thigh, center of the pelvis; right knee, right thigh; left knee, left thigh; right ankle, right knee; left ankle, left knee. These key points describe the human skeleton with the exception of the head.

The reference value of the skeleton length was obtained experimentally. For this, a model was chosen - a man with a height of 179 cm (Fig. 7). Several pictures of the man were taken at a distance of 3, 5 and 7 meters from the camera. Images were fed to the neural network to estimate the body key points coordinates. For each image the skeleton length was calculated by equation 4. As a reference length of the skeleton the median value of all images equal to 3846 mm was taken.

$$B_{len} = (d - f) * bones_{len}(J) * S_{pix} * \frac{1}{f} \quad (4)$$

where B_{len} is the size of the human skeleton in millimeters, f is the focal length of the camera lens in millimeters, d is the distance from the camera to the center point of the human pelvis in millimeters, J is the coordinates of the detected human body key points in pixels, $bones_{len}$ is the function for calculating the length of the skeleton, S_{pix} is the pixel size in millimeters.

IV. Dataset generation

A computer-graphic based dataset was developed based on materials from the city's security cameras to evaluate estimation of the three-dimensional human poses. A street model was created on which two people were placed to perform punches, kicks and other actions towards each other. Four different human models were used (Fig. 8), a total of 12 pairs of models were considered. The interaction was captured from various viewpoints given by three parameters: the angle of rotation of people relative to the point of the center of interaction (30, 60 and 90 degrees) (Fig. 9); the distance from the camera to people (point of the center of interaction) (5, 10, 15 and 20 meters) (Fig. 10); camera lifting angle (10, 20, 50 degrees) (Fig. 11). At any viewpoint, the camera is always directed to the center point of interaction. The maximum distance between people during interaction is 1500 millimeters, the distance decreases when performing action's

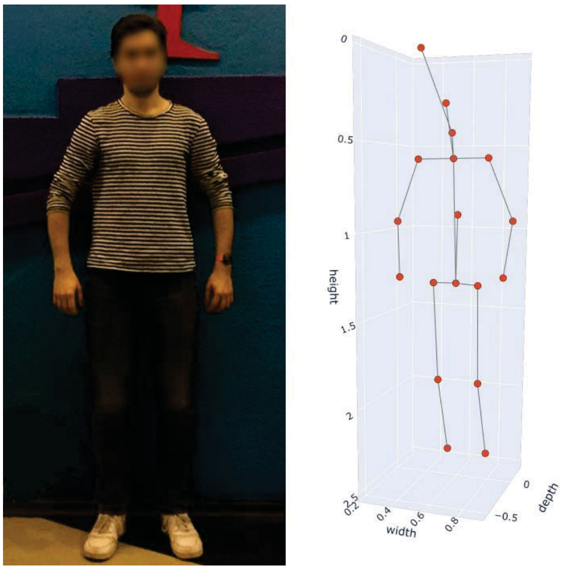

Fig. 7. Image of a model for evaluating the reference length of a human skeleton

animation. For each pair of models and each viewpoint, 300 frames were obtained. Each frame was captured by a camera, the characteristics of which correspond to the popular ISON AHD20F-CD surveillance camera: focal length - 2.8mm, pixel size - 2.8mkm, resolution - 1920x1080. The sizes of the models were selected so that the length of the skeleton was equal to the reference value - 3846 mm. As a result, the dataset of 129600 images have been obtained. Each image is accompanied by an annotation: the identifier of the person's model, the progress time of the animation, and the coordinates of the body key points are specified for each person, as well as the angle of people rotation, camera distance and camera lifting angle are specified for the entire frame. The center of the coordinate system is the point of the people interaction center.

V. 3D HUMAN POSE ESTIMATION RESULTS

When testing the estimation of three-dimensional poses of people, the error was calculated separately for each camera viewpoint (distance camera, camera lifting angle, angle of people rotation). The error was averaged over other parameters (various models and moments in the animation). When cutting out image patches with people the ground truth bounding boxes of people were used. For each human body key point three error metrics were calculated: absolute deviation (AD), average deviation relative to one's own skeleton (ADOS), average deviation relative to the adjacent skeleton (ADAS).

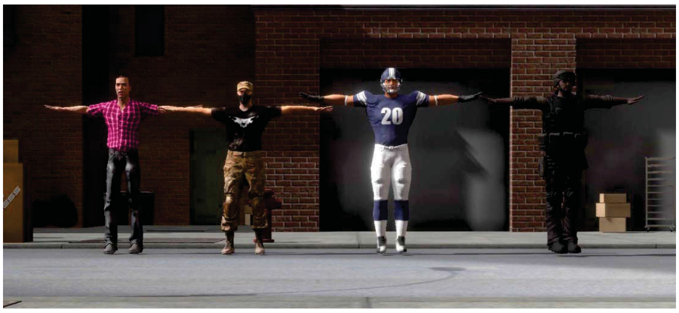

Fig. 8. Human models for dataset generation

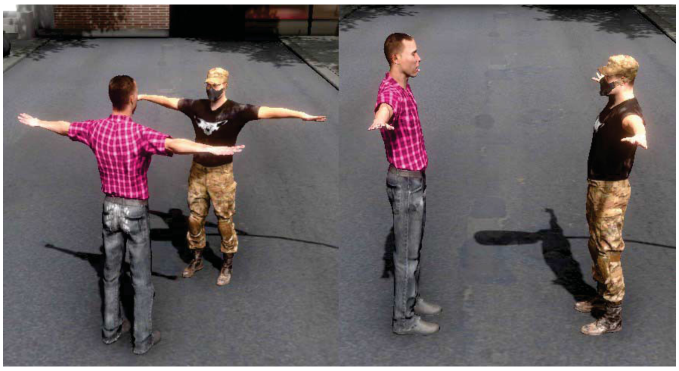

Fig. 9. Images of interacting people with a people rotation angle of 30 and 90 degrees relative to the point of the interaction center

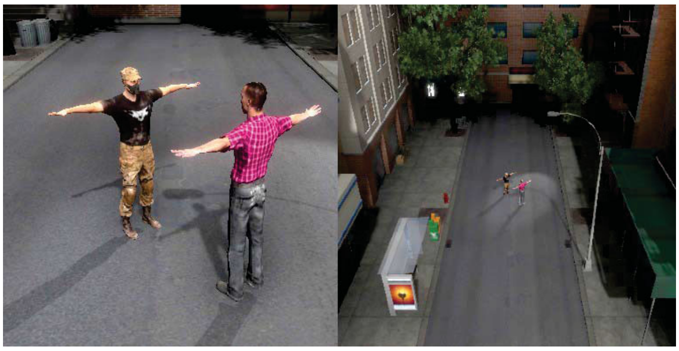

Fig. 10. Images of interacting people with a distance to the camera of 5 and 20 meters

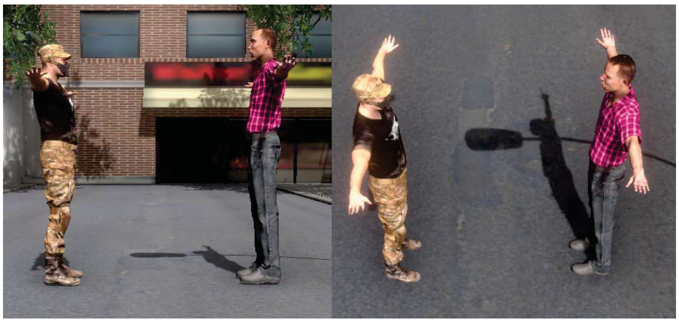

Fig. 11. Images of interacting people with a camera lifting angle of 10 and 50 degrees

The absolute deviation is calculated by the equation 5 and is used to calculate the mean per joint position error (MPJPE) metric. MPJPE was used to evaluate applicants in the Human3.6M challenge and shows the absolute deviation of the detected body key points in space. Since one of the main purposes of detecting body key points in this work is to classify the actions of individuals and the interactions of closely located people, in addition to the absolute deviation, two metrics were introduced that reflect the error of the relative location between key points within one's own skeleton (equation 6) and between key points of adjacent skeletons (equation 7). For each key point the total weighted length of deviation vectors is calculated. Each deviation vector is calculated as a difference between the vector of two ground truth key points and the vector of the two corresponding estimated key points. These errors are not sensitive to the similar displacement of all key points by one vector. Fig. 12 shows the vectors between the key points of the one's own skeleton, considered to calculate an estimation error of the right wrist's key point.

$$E(k)_{AD} = ||j(k) - j'(k)||_2 \tag{5}$$

where $E(k)_{AD}$ is the absolute deviation of the k-th human body key point in millimeters, $j(k)$ are the ground truth coordinates of the k-th human body key point in millimeters, $j'(k)$ are the estimated coordinates of the k-th human body key point in millimeters.

$$E(k)_{ADOS} = \sum_{i \in KP(k)} \frac{||v_{ki} - v'_{ki}||_2}{||v_{ki}||_2 * \sum_{p \in KP(k)} \frac{1}{||v_{kp}||_2}} \tag{6}$$

where $E(k)_{ADOS}$ is the average deviation of the k-th key point relative to one's own skeleton in millimeters, v_{ki} is the vector between the k-th and i-th key point of the ground truth skeleton, v'_{ki} is the vector between the k-th and i-th key point of the estimated skeleton, KP(k) is the set of all skeleton key points without k-th point.

$$E(k)_{ADAS} = \sum_{i \in KP} \frac{||v_{ki} - v'_{ki}||_2}{||v_{ki}||_2 * \sum_{p \in KP} \frac{1}{||v_{kp}||_2}} \tag{7}$$

where $E(k)_{ADAS}$ is the average deviation of the k-th key point relative to the adjacent skeleton in millimeters, v_{ki} is the vector between the k-th ground truth key point of the one person and i-th ground truth key point of the adjacent person, v'_{ki} is the vector between the k-th estimated key point of the one person and i-th estimated key point of the adjacent person, KP is the set of all skeleton key points.

The test results are presented in Tables I, II and III. When calculating average errors, 10 percent of human pose estimations with the highest ADOS error, averaged over all key points of the human body, were excluded from consideration. For the depth coordinate (z), errors were calculated separately from the (x, y) coordinates, since the estimation of (z) is a more difficult task, and it is more error prone taking into account the method used to estimate the absolute coordinates (equation 3) and the location of interacting people in the center of the view. As a result of statistical processing of the obtained data the graphs of average error for camera distance, camera lifting angle and characters rotation angle were obtained (Fig. 13, 14 and 15).

During the analysis of the test results, the following observations were made:

- Increasing the camera distance reduces the accuracy of the pose estimation (average AD 122-170 mm for x, y and 322-2753 mm for z; average ADOS 120-137 mm for x, y and 136-144 mm for z; average ADAS 182-272 mm for x, y and 329-2100 mm for z). Since the absolute values of the coordinates are calculated in accordance with equation 3, the absolute error is directly proportional to the camera distance and the skeleton length calculation error. At the same time, a decrease in image quality could affect the accuracy of key points estimation and, thus, the skeleton length estimation. A significant increase in ADAS may indicate the inefficiency of using absolute values of key points coordinates in the analysis of people interaction.
- Changing the camera lifting angle had a significant impact on the error (average AD 79-237 mm for x, y and 488-2802 mm for z; average ADOS 73-195 mm for x, y and 106-177 mm for z; average ADAS 120-380 mm for x, y and 655-1874 mm for z). The possible reason is that in the training dataset of the neural network used for poses evaluation, viewpoints were similar to the viewpoint of the testing dataset with 10-degrees camera lifting angle. And the network is trying to choose a pose corresponding to the familiar viewpoint (Fig. 16) which leads to the increase of the

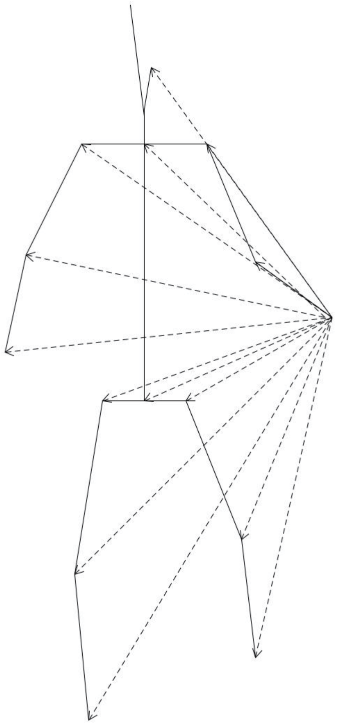

Fig. 12. Vectors considered when calculating the ADOS error of the right wrist's key point

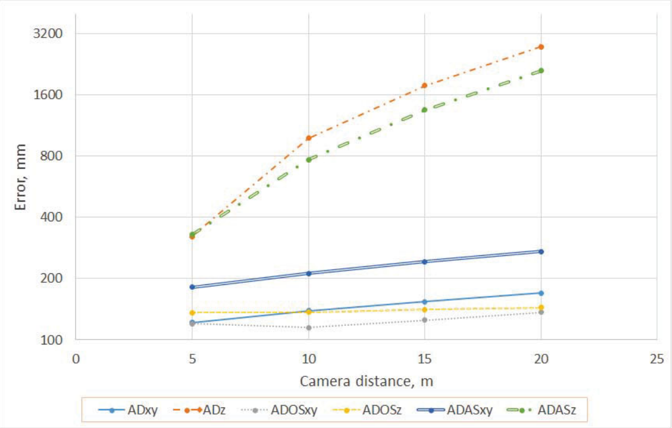

Fig. 13. Average three-dimensional pose estimation error depending on the camera distance

TABLE I. AVERAGE AD ERROR FOR EACH CAMERA VIEWPOINT (CAMERA DISTANCE IN METERS, CAMERA LIFTING ANGLE, PEOPLE ROTATION ANGLE) AND EACH BODY KEY POINT. FOR x, y AND z COORDINATES ERRORS ARE CALCULATED SEPARATELY

view	pelvis	rHip	rKnee	rAnkle	lHip	lKnee	lAnkle	nose	lShould	lElb	lWrist	rShould	rElb	rWrist	thorax
(5,10,30)	86/163	90/173	51/183	46/198	88/158	57/208	47/208	86/175	78/157	114/175	130/210	79/158	70/171	72/192	73/148
(5,10,60)	77/133	83/139	59/159	61/175	76/133	52/186	53/189	64/144	64/114	82/128	86/140	84/133	90/148	85/176	64/120
(5,10,90)	67/150	67/152	66/175	68/164	72/157	46/194	53/186	69/173	61/143	66/152	71/162	79/145	101/149	126/172	61/140
(5,30,30)	85/174	88/172	69/171	90/202	95/175	83/187	79/212	158/271	148/278	138/275	148/275	134/256	85/233	93/231	127/270
(5,30,60)	79/150	87/154	68/135	75/163	83/148	54/148	59/185	164/249	152/264	124/257	137/255	149/254	117/246	110/251	135/264
(5,30,90)	73/154	80/152	80/160	87/188	80/159	58/160	67/203	162/237	130/237	105/223	102/222	152/239	132/241	141/249	128/243
(5,50,30)	114/511	112/502	143/397	205/339	137/519	140/407	161/365	251/739	245/719	206/714	212/698	215/688	136/646	135/617	214/717
(5,50,60)	143/574	148/580	134/423	142/324	150/569	98/392	83/302	303/821	293/795	223/782	224/748	283/776	192/747	178/709	271/799
(5,50,90)	145/536	157/544	123/397	128/277	142/525	97/343	78/256	317/795	272/748	211/714	203/677	300/754	222/735	223/716	274/765
(10,10,30)	94/382	93/384	54/373	38/383	97/380	60/382	44/392	92/380	88/393	100/404	112/427	88/394	78/394	78/398	83/388
(10,10,60)	85/278	91/277	54/289	48/306	82/280	49/265	44/274	78/276	73/286	72/281	84/286	90/310	93/322	88/329	74/294
(10,10,90)	83/277	85/274	59/283	53/287	84/278	57/275	47/282	81/277	69/279	65/286	70/284	87/292	98/302	109/310	71/282
(10,30,30)	108/717	104/714	71/612	78/548	123/714	110/632	97/587	156/832	159/836	148/845	158/840	144/814	110/789	112/771	137/830
(10,30,60)	106/697	111/701	71/597	63/503	109/690	72/567	54/507	160/814	159/827	136/835	138/825	159/813	126/808	115/785	144/826
(10,30,90)	103/625	110/626	83/531	67/464	105/620	74/501	61/452	159/739	134/742	123/730	113/711	159/739	134/751	134/740	133/747
(10,50,30)	160/1798	153/1793	154/1584	241/1445	186/1804	220/1609	228/1488	260/2055	257/2035	254/2013	251/2004	245/2005	191/1963	184/1939	226/2032
(10,50,60)	191/1951	196/1958	185/1790	233/1639	200/1946	182/1744	184/1613	295/2205	289/2177	251/2157	254/2142	291/2161	244/2126	225/2100	271/2182
(10,50,90)	204/1881	220/1889	200/1691	193/1541	198/1872	168/1661	149/1521	311/2137	277/2097	240/2061	233/2034	309/2096	257/2075	254/2057	280/2112
(15,10,30)	102/678	102/678	60/665	43/677	106/677	67/656	50/671	94/657	93/685	101/691	110/711	91/686	83/682	85/680	86/682
(15,10,60)	87/492	93/489	54/494	50/510	85/493	54/461	45/470	83/478	76/497	73/500	80/501	91/513	92/522	85/521	75/502
(15,10,90)	83/499	87/493	63/501	56/503	85/503	63/488	55/492	87/496	69/500	70/505	71/503	88/515	95/521	114/515	72/507
(15,30,30)	120/1369	115/1369	84/1254	93/1183	135/1365	125/1262	109/1212	158/1483	166/1491	159/1497	166/1493	155/1469	131/1440	129/1411	143/1486
(15,30,60)	120/1379	124/1382	89/1274	80/1170	125/1371	92/1240	73/1174	166/1494	167/1509	153/1512	155/1505	174/1492	147/1485	132/1463	153/1507
(15,30,90)	119/1230	127/1231	108/1115	89/1035	120/1223	95/1095	82/1027	168/1345	143/1352	139/1335	131/1316	170/1344	147/1350	152/1338	141/1355
(15,50,30)	174/3225	171/3221	168/3004	262/2860	200/3230	244/3018	250/2884	274/3491	268/3474	281/3429	282/3423	259/3444	220/3382	215/3355	233/3471
(15,50,60)	211/3435	213/3441	221/3261	278/3107	223/3432	231/3224	244/3077	304/3691	292/3667	274/3636	283/3625	301/3648	271/3602	260/3576	276/3671
(15,50,90)	235/3544	249/3551	247/3346	246/3194	232/3539	227/3330	231/3181	333/3801	296/3769	274/3715	278/3693	329/3765	294/3722	293/3694	298/3781
(20,10,30)	109/1044	110/1042	68/1032	46/1047	111/1041	75/1018	59/1034	103/1008	99/1045	105/1049	116/1061	99/1049	92/1046	95/1037	91/1043
(20,10,60)	93/862	101/858	62/857	58/863	90/862	63/828	59/841	95/837	81/865	85/867	93/872	101/877	108/883	95/876	81/869
(20,10,90)	90/834	94/826	72/829	63/837	91/837	71/812	61/821	105/816	78/832	77/835	79/830	99/845	104/856	118/845	80/837
(20,30,30)	122/2019	119/2018	92/1907	108/1834	136/2015	135/1910	127/1862	158/2133	166/2141	166/2139	172/2131	161/2124	145/2090	138/2059	143/2137
(20,30,60)	133/2188	139/2192	110/2086	98/1988	135/2181	111/2044	96/1977	176/2304	173/2322	165/2318	171/2310	184/2305	167/2288	150/2266	160/2319
(20,30,90)	135/1999	143/2000	124/1883	106/1801	135/1993	115/1863	103/1796	179/2114	157/2124	152/2102	145/2084	180/2115	166/2108	169/2094	152/2126
(20,50,30)	185/4815	185/4811	185/4590	276/4442	215/4823	271/4602	277/4469	287/5084	277/5077	308/5023	307/5007	272/5036	243/4962	254/4932	239/5068
(20,50,60)	231/5163	232/5170	254/4973	314/4828	247/5161	277/4945	287/4805	320/5420	300/5399	300/5348	312/5338	316/5385	301/5329	293/5294	284/5405
(20,50,90)	276/5720	286/5728	292/5513	291/5367	278/5714	284/5498	281/5347	365/5983	323/5955	313/5891	315/5869	354/5948	336/5897	339/5866	321/5966

TABLE II. AVERAGE ADOS ERROR FOR EACH CAMERA VIEWPOINT (CAMERA DISTANCE IN METERS, CAMERA LIFTING ANGLE, PEOPLE ROTATION ANGLE) AND EACH BODY KEY POINT. FOR x, y AND z COORDINATES ERRORS ARE CALCULATED SEPARATELY

view	pelvis	rHip	rKnee	rAnkle	lHip	lKnee	lAnkle	nose	lShould	lElb	lWrist	rShould	rElb	rWrist	thorax
(5,10,30)	52/91	62/101	75/110	83/163	64/92	77/106	79/118	83/135	78/105	107/109	127/121	89/96	85/98	93/115	77/118
(5,10,60)	51/82	59/106	79/102	85/144	63/84	74/114	80/122	74/141	77/96	99/100	102/104	82/92	91/102	94/122	72/113
(5,10,90)	52/73	58/87	79/88	85/106	65/85	71/98	78/107	77/139	85/79	92/89	92/104	79/102	103/97	124/124	74/91
(5,30,30)	74/86	88/99	129/151	153/209	88/99	124/145	139/191	109/137	107/122	125/128	137/152	110/135	101/102	101/119	96/110
(5,30,60)	72/93	82/108	118/149	134/204	87/112	105/174	123/221	103/112	106/133	117/125	118/126	110/128	112/129	105/125	92/111
(5,30,90)	72/97	82/108	122/153	142/185	87/111	109/163	128/208	106/99	106/126	108/125	97/130	112/122	126/121	128/141	95/115
(5,50,30)	136/116	155/141	236/241	295/332	157/133	212/238	253/321	166/125	168/111	182/152	193/165	163/109	153/133	149/130	144/100
(5,50,60)	125/124	141/147	216/246	248/306	142/150	179/249	215/351	162/125	165/121	163/155	164/150	168/116	159/137	150/143	143/103
(5,50,90)	119/115	134/131	203/239	245/294	136/145	176/251	218/320	163/115	152/118	144/134	143/130	169/111	166/146	183/164	145/94
(10,10,30)	47/87	55/96	67/112	73/170	58/93	70/100	71/115	74/131	69/102	83/110	99/126	73/92	70/96	80/111	68/110
(10,10,60)	45/79	52/102	67/101	73/141	55/82	64/109	74/124	66/132	66/93	77/100	88/109	68/93	77/100	87/125	63/109
(10,10,90)	44/70	50/82	67/89	72/106	56/76	65/90	76/106	67/130	66/76	70/83	76/103	66/99	83/99	102/126	64/81
(10,30,30)	72/87	85/97	127/154	150/216	86/99	124/148	148/200	96/136	97/116	108/131	121/161	96/130	91/101	95/123	85/107
(10,30,60)	64/91	74/106	105/148	119/211	79/108	95/171	119/223	87/114	93/130	94/132	97/135	93/125	93/129	89/128	80/110
(10,30,90)	63/93	73/105	106/155	118/183	77/105	101/165	127/207	91/97	89/123	90/125	87/132	97/116	104/121	107/145	82/109
(10,50,30)	145/117	169/140	247/246	330/337	165/134	240/244	301/326	165/128	171/115	195/147	197/168	176/115	175/140	173/148	150/105
(10,50,60)	128/124	144/147	219/240	281/311	148/147	200/244	254/353	163/130	162/126	166/154	176/163	166/122	172/143	166/164	140/108
(10,50,90)	113/115	129/131	201/257	248/309	133/143	181/252	237/319	154/114	144/119	146/134	154/144	158/116	155/148	178/173	134/99
(15,10,30)	48/92	55/99	68/115	75/173	59/96	72/102	76/118	74/138	70/105	83/113	98/133	71/96	68/101	81/118	68/110
(15,10,60)	46/83	53/107	66/106	73/147	56/84	65/110	77/127	66/140	64/95	74/100	81/111	67/95	75/99	84/125	63/113
(15,10,90)	46/73	53/83	69/95	75/108	57/79	66/93	82/113	70/135	67/78	75/86	77/112	68/103	85/103	110/130	65/83
(15,30,30)	75/91	90/102	138/160	166/217	91/102	135/159	164/212	101/138	102/119	116/135	131/169	101/135	100/107	104/131	88/113
(15,30,60)	67/94	77/108	112/153	127/219	82/108	102/176	133/228	90/117	96/131	100/134	106/139	99/128	100/132	97/135	83/113
(15,30,90)	65/94	76/104	118/164	129/194	80/108	108/166	143/211	95/99	91/125	96/125	95/136	101/113	106/121	115/148	85/112
(15,50,30)	156/118	183/139	271/247	360/336	176/136	256/246	320/336	182/129	187/122	218/146	223/168	193/122	198/145	195/155	162/112
(15,50,60)	140/126	159/150	248/244	318/320	163/149	229/249	293/355	177/130	176/135	194/157	207/173	183/128	194/150	196/178	153/116
(15,50,90)	126/115	144/133	230/259	288/318	150/140	213/252	284/321	172/116	162/127	170/139	182/148	173/121	177/148	204/181	148/107
(20,10,30)	49/93	58/100	71/116	82/173	60/98	77/105	87/121	79/142	71/107	87/116	103/139	74/97	72/101	88/120	70/107
(20,10,60)	48/87	56/105	72/106	84/146	58/87	72/114	90/131	72/147	68/101	79/107	87/121	72/97	84/101	88/126	66/111
(20,10,90)	47/76	54/85	73/94	81/109	58/85	70/96	90/118	78/135	71/80	75/86	81/117	72/106	88/106	112/130	68/84
(20,30,30)	79/93	94/103	147/166	181/223	95/105	142/167	177/219	110/136	108/119	128/133	142/165	110/137	114/113	119/138	94/114
(20,30,60)	71/95	83/109	124/157	140/215	87/109	116/181	153/233	101/118	103/131	112/132	122/141	107/128	113/130	109/142	89/116
(20,30,90)	68/96	80/107	124/165	136/197	84/109	119/166	161/208	104/101	99/127	104/125	104/145	108/115	119/121	129/151	90/117
(20,50,30)	169/122	197/143	286/252	381/345	192/139	275/253	343/342	200/132	203/129	246/148	252/175	214/125	224/151	223/161	177/116
(20,50,60)	153/128	172/152	271/254	356/329	180/152	258/256	326/360	192/130	190/137	214/156	232/177	201/133	216/156	220/180	166/121
(20,50,90)	140/120	161/139	263/263	339/323	166/146	240/263	313/325	190/120	180/136	194/149	210/161	188/129	200/157	234/193	163/117

TABLE III. AVERAGE ADAS ERROR FOR EACH CAMERA VIEWPOINT (CAMERA DISTANCE IN METERS, CAMERA LIFTING ANGLE, PEOPLE ROTATION ANGLE) AND EACH BODY KEY POINT. FOR x, y AND z COORDINATES ERRORS ARE CALCULATED SEPARATELY

view	pelvis	rHip	rKnee	rAnkle	lHip	lKnee	lAnkle	nose	lShould	lElb	lWrist	rShould	rElb	rWrist	thorax
(5,10,30)	114/279	116/285	95/272	98/282	119/274	105/317	100/326	125/291	122/283	158/304	171/330	114/270	114/272	122/290	116/271
(5,10,60)	107/214	111/218	102/220	106/239	109/210	98/242	104/254	106/212	105/217	123/225	126/232	118/221	127/231	128/244	105/215
(5,10,90)	108/211	109/211	113/220	118/226	112/215	100/220	108/222	115/228	108/220	117/223	122/224	118/223	139/224	163/238	108/223
(5,30,30)	128/282	128/282	140/307	166/348	141/281	147/337	157/368	166/330	168/321	183/330	194/353	145/300	133/293	142/317	146/309
(5,30,60)	127/245	134/251	135/264	146/303	133/235	124/280	139/303	168/260	169/258	161/257	168/272	167/275	159/272	156/288	153/271
(5,30,90)	127/235	136/233	143/254	152/255	130/248	129/254	141/280	181/256	155/257	148/256	148/261	180/256	180/262	188/271	154/252
(5,50,30)	221/517	214/514	258/588	317/650	245/517	263/569	295/625	293/536	294/536	281/521	288/501	256/527	219/517	219/499	260/537
(5,50,60)	245/391	251/391	250/441	270/508	254/398	227/429	239/486	340/456	342/449	296/429	294/419	330/443	278/432	265/422	318/448
(5,50,90)	264/395	281/379	262/389	264/408	259/419	233/414	242/445	369/439	330/423	294/426	286/413	370/410	326/386	326/395	337/417
(10,10,30)	115/563	114/567	93/552	94/551	123/560	105/588	98/596	125/574	126/568	143/584	155/609	118/558	118/563	121/577	119/557
(10,10,60)	109/407	114/406	96/428	97/441	109/409	92/404	97/406	111/413	107/407	113/400	125/403	119/435	130/441	135/445	109/418
(10,10,90)	110/420	113/416	100/423	100/425	111/423	98/421	102/425	117/426	107/421	112/422	119/422	119/426	134/425	152/432	109/425
(10,30,30)	151/639	144/637	145/636	166/653	171/641	175/673	184/692	179/668	187/657	197/665	208/679	158/648	149/646	156/664	163/651
(10,30,60)	148/509	154/513	145/520	150/529	155/506	141/519	151/533	178/531	185/527	173/525	175/536	181/532	170/532	164/539	168/531
(10,30,90)	151/542	162/539	149/541	144/539	151/551	141/547	152/554	193/566	169/562	170/558	167/556	191/563	186/561	189/560	167/563
(10,50,30)	298/1276	281/1278	302/1308	377/1337	327/1272	373/1304	404/1324	352/1280	354/1275	367/1263	365/1242	316/1270	299/1254	298/1239	312/1278
(10,50,60)	339/1301	340/1299	332/1335	381/1385	350/1310	354/1319	383/1347	399/1359	398/1358	383/1336	385/1322	388/1345	369/1314	357/1296	374/1352
(10,50,90)	351/1143	370/1132	358/1105	365/1130	344/1158	335/1132	353/1137	418/1201	382/1185	366/1178	360/1169	414/1171	390/1135	390/1131	384/1180
(15,10,30)	124/897	123/899	101/885	103/883	132/894	114/918	108/923	130/904	133/899	147/909	158/937	126/891	124/895	129/913	126/889
(15,10,60)	110/646	115/644	98/664	100/676	111/647	95/636	100/637	119/649	112/644	114/635	124/637	123/667	130/671	135/671	112/653
(15,10,90)	117/735	121/730	110/737	108/739	118/737	108/731	114/729	132/738	117/733	123/732	126/730	129/741	143/733	166/728	119/738
(15,30,30)	169/1091	161/1088	165/1095	188/1112	189/1093	198/1108	208/1118	189/1092	200/1092	215/1089	226/1097	176/1099	172/1102	178/1110	177/1093
(15,30,60)	176/1004	181/1007	174/1016	176/1022	182/1008	171/996	183/1009	203/1030	209/1029	207/1020	213/1028	210/1032	203/1027	194/1033	193/1028
(15,30,90)	182/1016	193/1011	186/998	177/1000	182/1025	176/1012	187/1006	221/1045	195/1038	200/1030	199/1026	217/1039	213/1027	221/1028	193/1039
(15,50,30)	328/2053	314/2057	332/2076	407/2088	358/2050	411/2067	438/2078	389/2065	387/2056	415/2038	418/2024	347/2055	341/2037	348/2022	342/2060
(15,50,60)	395/2343	392/2336	384/2369	438/2413	411/2356	434/2358	466/2378	453/2391	443/2396	450/2371	457/2355	432/2381	426/2341	424/2320	419/2389
(15,50,90)	422/2309	438/2298	426/2271	436/2295	418/2323	427/2300	462/2291	486/2360	444/2346	441/2338	446/2326	473/2337	459/2295	462/2285	444/2343
(20,10,30)	130/1259	131/1260	110/1252	112/1249	138/1258	122/1275	122/1277	139/1257	141/1258	155/1263	167/1284	135/1256	135/1261	139/1275	133/1251
(20,10,60)	124/1053	131/1051	113/1072	118/1079	124/1054	111/1044	121/1041	139/1055	126/1050	131/1039	144/1042	141/1071	155/1071	153/1069	126/1058
(20,10,90)	126/1107	131/1102	120/1109	117/1110	127/1108	119/1103	123/1101	152/1104	128/1103	133/1101	138/1099	141/1107	154/1101	175/1095	130/1105
(20,30,30)	183/1636	176/1633	180/1637	208/1640	202/1638	216/1649	230/1652	207/1637	215/1636	236/1633	246/1637	196/1646	197/1647	201/1656	191/1640
(20,30,60)	202/1598	209/1599	203/1607	205/1613	207/1602	202/1593	217/1598	230/1625	229/1623	235/1611	246/1619	233/1626	233/1616	224/1619	214/1622
(20,30,90)	211/1622	222/1615	216/1600	204/1609	210/1633	209/1614	221/1599	249/1652	223/1647	227/1635	227/1627	242/1647	244/1629	252/1627	218/1646
(20,50,30)	359/3045	347/3048	362/3063	431/3068	391/3044	451/3055	476/3056	423/3061	417/3056	460/3048	463/3041	376/3049	378/3032	403/3026	368/3054
(20,50,60)	439/3646	434/3641	429/3674	483/3709	459/3656	497/3657	526/3672	498/3697	481/3700	499/3671	511/3666	472/3687	477/3652	481/3635	455/3693
(20,50,90)	500/3875	511/3865	496/3850	501/3860	501/3889	521/3873	549/3861	559/3922	511/3909	523/3896	530/3881	535/3903	538/3862	545/3848	507/3907

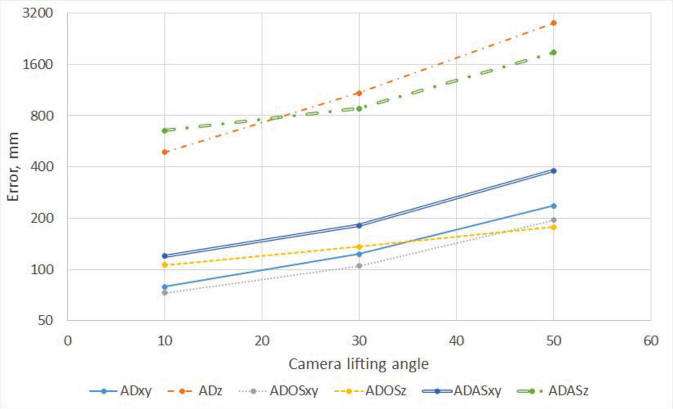

Fig. 14. Average three-dimensional pose estimation error depending on the camera lifting angle

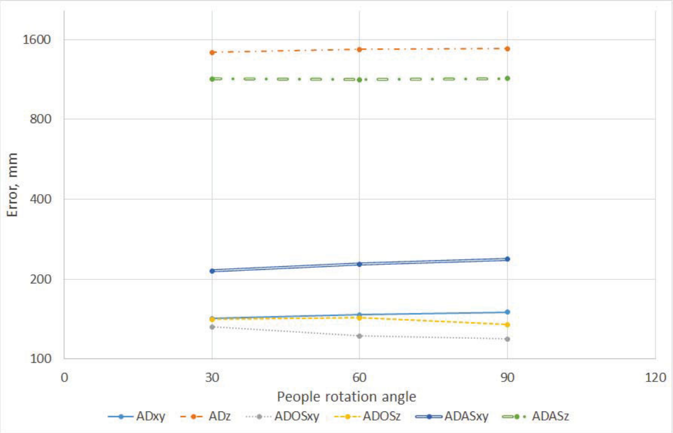

Fig. 15. Average three-dimensional pose estimation error depending on the people rotation angle

skeleton length calculation error.

- Changing people rotation angle did not have a significant effect on the error.
- The absolute error for the z coordinate turned out to be substantially larger than the error for the x, y coordinates. The reason for this is the method used to estimate the absolute coordinates (equation 3) and the fact that interacting people were placed in the center of the camera view. Moving people to the edge of the camera view should transfer part of an error to x, y coordinates.
- An analysis of the 10 percent of estimations with the highest ADOS error revealed that the highest error with the same camera view is achieved when there are occlusions of people (Fig. 17) and when there are poses with lifted legs (Fig. 18), which were absent in the training dataset of the neural network. Another common error is where the key points of the left and right legs are swapped (Fig. 19).

Summing up, we can conclude that it is necessary to retrain the neural network to estimate three-dimensional human poses using a dataset containing camera viewpoints inherent in the city's video surveillance cameras and poses inherent in the deviant behavior of people. Using the absolute coordinates of key points in the analysis of people interactions can be ineffective due to a high error. A possible solution is to use the coordinates of key points, relative to each person, with additional information - the approximate proximity of people and the rotation angle of a body. The utilized neural network is not able to handle strongly occluded people images, which can lead to the omission of important information. The solution can be the use of alternative approaches to the three-dimensional human pose estimation more robust to occlusions, for example, bottom-up approaches [22], [23], in which all individual body parts are detected in the original image and associated with a specific person.

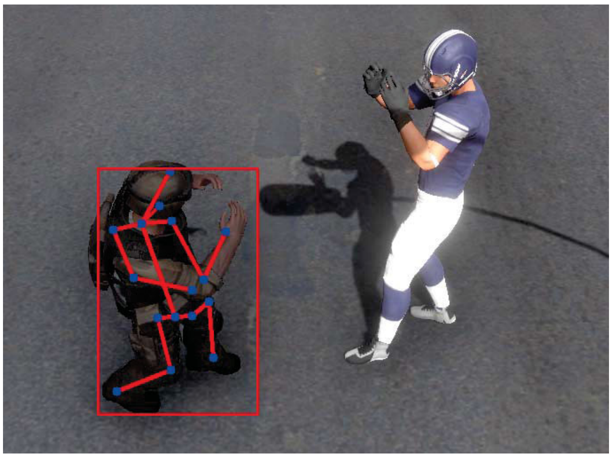

Fig. 16. An example of incorrect estimation of a pose in image with a 50-degree camera lifting angle

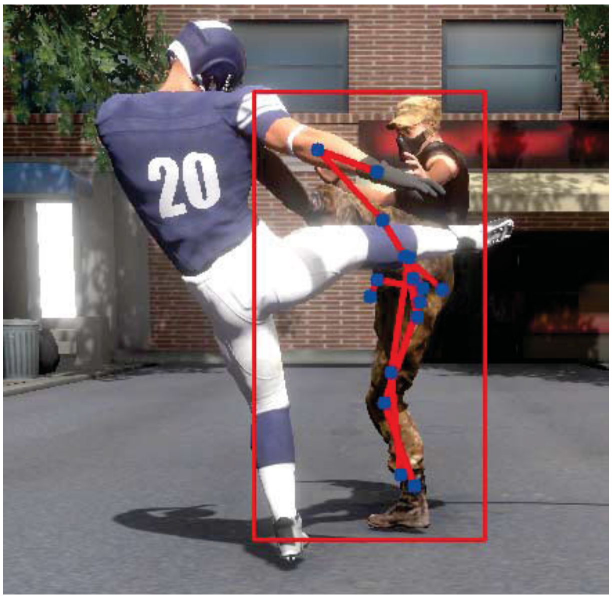

Fig. 17. An example of incorrect estimation of a pose in presence of occlusions

VI. Conclusion

Detecting the deviant behavior of people is an urgent task nowadays. The solution of that problem will improve the safety of residents. And to solve it, it is necessary to improve video surveillance systems of cities by the integration with modern intelligent technologies, such as object recognition, three-dimensional human poses estimation and knowledge management. In this work, a system was developed for tracking interactions of people in three-dimensional space. The system is based on the integration of open source implementations of technologies for people detection, object tracking, and three-dimensional human poses estimation that have showed high results in modern computer vision challenges. This system will be used as a basis for the recognition of the actions and interactions of people via city's surveillance cameras. Testing of the three-dimensional poses estimation was done based on artificial images from outdoor surveillance cameras. Weaknesses were identified that should be taken into account when implementing recognition of the actions and interactions of people. Some weaknesses can be eliminated by additional training of the utilised neural network using a dataset that contains problematic poses of people and camera viewpoints.

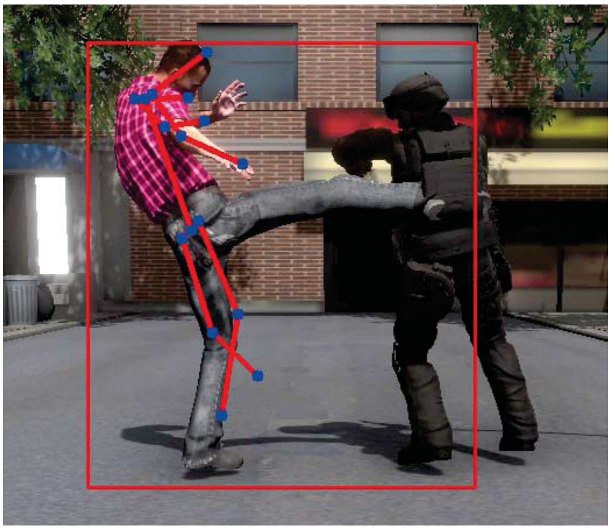

Fig. 18. An example of incorrect estimation of a kick pose

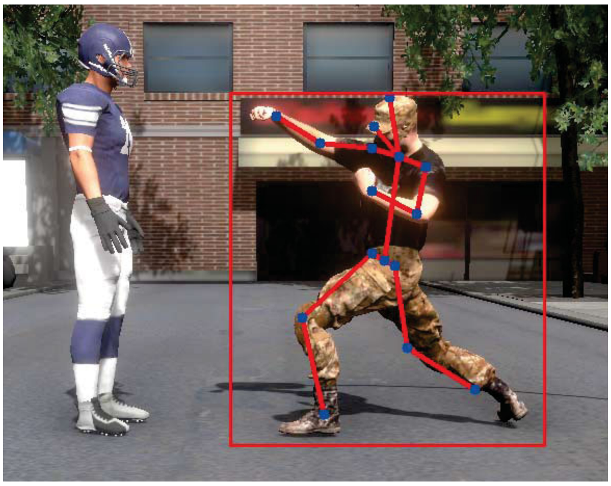

Fig. 19. An example of incorrect estimation of legs key points

The aim of further work is to eliminate the shortcomings identified during testing of the obtained solution, and to continue the implementation of the concept of deviant behavior detection via city's surveillance cameras.

Acknowledgment

The reported study was partly funded by RFBR, project number 20-07-00904.

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

[4] P. Zhou, Q. Ding, H. Luo, and X. Hou, *Violence detection in surveillance video using low-level features* PLoS One, vol. 13, no. 10, p. e0203668, Oct. 2018.

[5] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, *Violence detection in video using computer vision techniques,* in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6855 LNCS, no. PART 2, pp. 332–339.

[6] A. Zanfir, E. Marinoiu, and C. Sminchisescu, *Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes: The Importance of Multiple Scene Constraints,* in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 2148–2157.

[7] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, *RGB-D-based human motion recognition with deep learning: A survey,* Comput. Vis. Image Underst., vol. 171, pp. 118–139, Jun. 2018.

[8] 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.

[9] 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.

[10] 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.

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

[12] T. Y. Lin et al., *Microsoft COCO: Common objects in context,* in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014, vol. 8693 LNCS, no. PART 5, pp. 740–755.

[13] *COCO - Common Objects in Context.* Web: http://cocodataset.org/#home. [Accessed: 01-Mar-2020].

[14] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, *MOT16: A Benchmark for Multi-Object Tracking,* Mar. 2016.

[15] *MOT Challenge.* Web: https://motchallenge.net/. [Accessed: 01-Mar-2020].

[16] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, *Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,* IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, 2014.

[17] *Human3.6M Dataset.* Web: http://vision.imar.ro/human3.6m/
description.php. [Accessed: 01-Mar-2020].

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask R-CNN,* IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 386–397, Feb. 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition,* in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-December, pp. 770–778.

[20] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,* IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[21] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, *Learning by Tracking: Siamese CNN for Robust Target Association,* in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 418–425.

[22] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, *Realtime multi-person 2D pose estimation using part affinity fields,* in Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, vol. 2017-January, pp. 1302–1310.

[23] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,* IEEE Trans. Pattern Anal. Mach. Intell., pp. 1–1, Jul. 2019.