Human Detection System by Fusing Depth Map-Based Method and Convolutional Neural Network-Based Method

Anh Vu LE[1], *Tran Tin PHU*[2,3], *Jong Suk CHOI*[4], *Jan SKAPA*[5], *Miroslav VOZNAK*[5]

[1]Optoelectronics Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, No. 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam
[2]Wireless Communications Research Group, Ton Duc Thang University, No. 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam
[3]Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, No. 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam
[4]Center for Robotics Research, Korea Institute of Science and Technology, Hawolgok-dong 39-1, Seongbuk-ku, Seoul 136-79, Republic of Korea
[5]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15, 708 00 Ostrava, Czech Republic

leanhvu@tdt.edu.vn, phutrantin@tdt.edu.vn, pristine70@gmail.com, jan.skapa@vsb.cz, miroslav.voznak@vsb.cz

DOI: 10.15598/aeee.v15i4.2377

Abstract. *In this paper, the depth images and the colour images provided by Kinect sensors are used to enhance the accuracy of human detection. The depth-based human detection method is fast but less accurate. On the other hand, the faster region convolutional neural network-based human detection method is accurate but requires a rather complex hardware configuration. To simultaneously leverage the advantages and relieve the drawbacks of each method, one master and one client system is proposed. The final goal is to make a novel Robot Operation System (ROS)-based Perception Sensor Network (PSN) system, which is more accurate and ready for the real time application. The experimental results demonstrate the outperforming of the proposed method compared with other conventional methods in the challenging scenarios.*

Keywords

Deep learning, fusion, human detection, ROS.

1. Introduction

Human detection is an important and interesting research area in the computer vision. The applications of human detection systems, in both academic and commercial environments, are implemented widely especially in the surveillance system, and for robot human interaction fields [1] and [2]. Recently, remarkable attempts such as [2] and [3] have been proposed to make the human detection tasks robust in the cluster environments. The challenges of these issues are that humans are normally occluded by others and the objects looking like humans are misclassified as humans. The widely-used methods are Histogram of Gradient (HOG)-based methods [4] and [5]. These methods try to build the histogram of the gradient maps from the training images. Subsequently, these features maps are used to train Support Vector Machine (SVM) features [6]. As SVM is calculated based on the gradients of images, the advantages of HOG-based methods include their simplicity, robustness to illumination changes and speed of detection of human beings. Accordingly, these methods are commonly used to detect pedestrians where the histogram of gradient maps of pedestrians indicates a difference with the objects in street [7] and [8].

The HOG-based methods, however, are not effective in detecting the humans who are not in the standard standing position, such as for instance a rear side appearing, or a sitting human being. In these methods, the humans are detected using only the colour information. To address the limitations of colour information such as light condition, similar colour of humans and other objects, other methods also consider the depth information to enhance the accuracy and efficiency of

human detection [9] and [10]. Note that recently, depth cameras such as Kinect [11] have become widespread. The depth blobs are classified according to which body part they belong and human joints are constructed from the training databases including a vast number of depth maps. The Openni method [12] leverages the advantage of depth information to enhance human detection in cluster environment. Because this method relies on depth information to detect any humans in the scenes, the three-dimensional (3D) information, and the region where the detected human occupied referred to as the Region of Interest (ROI), can be easy to be obtained. ROI is a two-dimensional (2D) rectangular with the locations at the top corner 2D coordinates, width and height. 3D information includes row, column index and depth value. This information is significantly useful in advance computer vision tasks such as human tracking, gesture, and action recognition. In addition, based on the depth information, containing different depth values for individual regions when it comes to humans in the foreground and other objects in the background in the cluster environment, these methods are obviously quick and require a simple hardware configuration. One of the most powerful features of Openni is that the human candidates identified by means of the depth images are available almost immediately after the human being has appeared in the scene thanks to the depth values discrimination in the depth maps. Although the human candidates are quickly available, in order to reduce the false detection rates, the Openni method takes a certain time to calibrate the human candidate skeleton joints against the standard joints distances already trained by the databases. The calibration time commonly takes up to two seconds from the moment the human being has appeared in the scene. Where the calibration is not successful, the detected human candidate is not assigned any ROI information. As the result of the available ROI, information such as information Identification (ID) and 3D location of this candidate is not available. The processes of Openni at clients are depicted in Fig. 1. To enhance the detection accuracy and speed, the burden of the calibration task of the Openni-based method must be removed.

The deep learning convolutional neural network has been developed over decades [13]. Human detection and human action recognition methods based on these networks have been researched intensively in theory. These methods have proved their powers in terms of accuracy [14]. The unique features of each object are established through a deep learning in large neural networks. The biggest difficulty, which makes the training processes of deep learning difficult to converge, is that there are a large number of filters coefficients, called free parameters, which need to be estimated from the convolutional neural network layers. Recently, the advanced technologies in computer graphic, such as Compute Unified Device Architecture CUDA [15], or NVIDIA CUDA Deep Neural Network CuDNN [16], enable parallel computing and multiple threads to be processed in separate graphic cards with their own Graphic Processing Unit (GPU). Consequently, the running time necessary to obtain the convergence of training these networks has dropped significantly, making deep learning-based methods' use in real applications viable. Since the features of the objects can be identified, the most significant achievement of these deep learning-based methods is their high accuracy rate in object classification. The advances in object detection, including object localizations and objects classification, are driven by the success of state-of-the-art-convolutional network (ConvNet)-based object detectors called the Region-based Convolutional Network method (RCNN) [17]. Comparing with the conventional methods such as colour HOG-based and depth-based Openni, these methods still have some drawbacks such as heavy computation time, or the need of expensive and powerful hardware configuration. These limitations pose a problem when implementing these methods at remote client sites where the requirement for high hardware configuration is hard to be satisfied. It is worth to know that in order to detect a human being at remote sites, the hardware configuration should have a compact size. The deployment of the advanced technologies of high configuration graphic hardware into a compact size requires a certain time. In addition, the cost of implementation of high configuration both computers and graphic cards have to be taken into account during the implementation. Furthermore, because these methods detect objects based on the colour information without any depth dimension, the 3D locations of detected humans are not available. In addition, the depth information is not used; these deep learning-based methods do not take into consideration the human joints which are obtained easily by the Openni method.

Recently, numerous efforts to reduce the complexity and the computational time of objection classification based on deep learning methods have been carried out. Simple features and effective region searching schemes are the basic ideas of the region proposal methods. The analysis of an image provided by the Selective Search method [18] to identify any high potential of objects locations usually returns only rough localization and must be refined to obtain precise localization. Solutions to these problems often compromise speed, accuracy, or simplicity of streamlining the training process. Spatial Pyramid Pooling networks (SPPnets) [19] were proposed to reduce the execution time of RCNN by sharing computation. SPPnets accelerate RCNN up to 100 times at the test time. The training time is also reduced three times because of the proposal feature extraction efficiency. Besides, there are still some notable issues of SPPnets. Similar to RCNN, training steps of

SPPnets include learning features, fine-tuning a network with log loss, training SVMs, and finally finding the best fitting bounding box. Note that the big amount of disk space is used to store the learned features. Simultaneous learning to classify object proposals and refinement of their spatial locations to reduce the execution time is achieved in a single stage training algorithm called Fast RCNN [20]. An entire image and a set of object proposals are fed to this network. Several convolutional (CONV) and max pooling layers process the input data to produce a CONV feature map. Then, for each Region of Interest (ROI) of the object proposal, pooling layer extracts a fixed-length feature vector from the feature map.

These features then are processed by fully connected (fc) layers that finally result into two sibling output layers. The first layer produces softmax probability estimates over L object classes plus a catch-all "background" class. The second layer returns four real-valued numbers of bounding-box positions (ROI) including 2D location row, column indexes in the input image, width, and height for each of the L object classes. The resulting method can train a very deep detection network VGG16 [21] nine times faster than RCNN and three times faster than SPPnets. The speed-up of runtime has been achieved. It takes 0.3 s to obtain detection network processes images (excluding the object proposal time). A simpler training scheme is presented in Faster RCNN [22] (FRCNN). Alternating between fine-tuning for the region proposal task and fine-tuning for object detection, while keeping the proposals fixed is the main idea of FRCNN to make its scheme converge more rapidly. As a result, these modifications produce a unified network with CONV features that are shared between both tasks. Running on GPU, the resulting detection with very deep network VGG16 has a frame rate of 5 fps, while achieving state-of-the-art object detection accuracy using 300 proposals per image. Thus, this method is a practical object detection system in terms of both speed and accuracy. The burden of detection time for the tested image with the big size is still present in the FRCNN. One of the solutions reduces the size of the tested image before conducting object detection but this solution also reduces the accuracy rate. The issues of Openni and FRCNN are summarized in Tab. 1. In this paper, the advantages of Openni and FRCNN are combined. A proposed Perception Sensor Network (PSN) system, including various Kinect cameras (PSN units) acting as system perception parts, is used to detected and track humans appearing in the Field of View (FOV) of Kinects in PSN. There are threefold aspects of the proposed method. Firstly, the client and master system with a simple hardware configuration at client sites is proposed to reduce the investment budget and network workload. Secondly, the detection time is reduced as the detection of a human candidate by Openni is implemented at the client with simple hardware configuration; and these candidates are verified at master with high hardware configuration by advanced faster FRCNN method. This system configuration reduces the complexity of the client, the burden of calibration steps of the Openni method and exploits the accuracy of FRCNN efficiently. Finally, the 3D locations are estimated from the final detected ROIs immediately after the successful human detection.

The structure of the paper is as follows: part 2 introduces the PSN system, and the next section describes the solution for detection of humans based on the proposed system. Then, the experimental results of the proposed method are provided. The last section contains the conclusion.

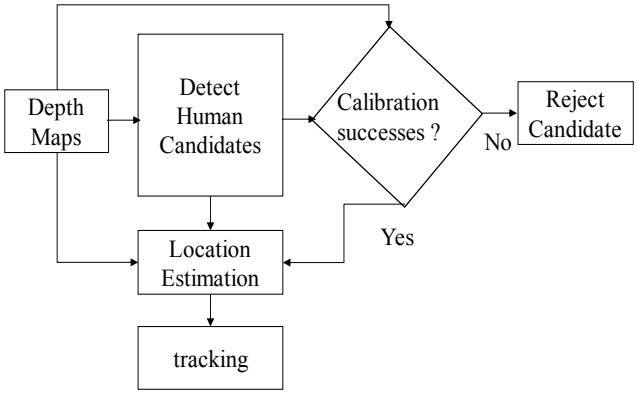


Fig. 1: Block diagram of Openni tracking method.

Tab. 1: The advantages and disadvantages of Openni and FRCNN.

Openni		FRCNN	
advantage	**disadvantage**	**advantage**	**disadvantage**
Need low hardware	High false detection rate	High accuracy rate	Need high hardware
Human mask and 3D available	Calibration slow	Multiclass detection	Human mask and 3D unavailable

2. Perception Sensor Network

To detect human, PSN system uses Openni, which is wrapped in a package of a Robot Operation System (ROS) [23]. ROS, an open source Linux and a master client communication platform, is widely implemented in robotics fields. Each PSN unit acts as the single node of ROS and communicates with other PSN units under the control of a server acting as the master of ROS. In PSN, the novel ROS-based package called Openni [12] is installed on the PSN units which use Kinects to detect the humans appearing in Kinect FOVs. Subsequently, 3D distance and 2D ROI of these detected

humans is provided. Specifically, the location of humans can be detected easily by finding the human body joints, which is supported very effectively by the available open source libraries in [12]. After presenting the detected humans, these humans are tracked by PSN and location and name is published as ROS topics in the system.

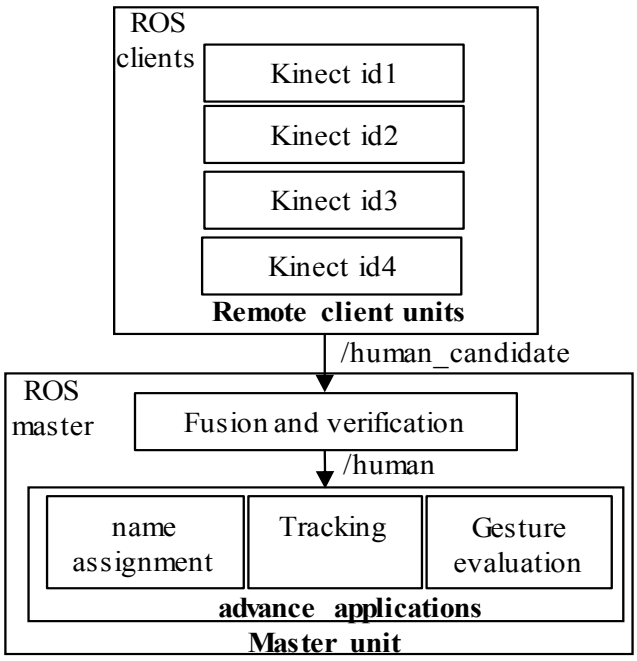


Fig. 2: Modules of PSN.

Fusion and verification modules are installed at the master side. By subscribing the /human_candidate topics which are advertised by all PSNs units, these modules accumulate all detected humans. After verifying, fusion rules will remove the redundant humans and keep the correct number of detected humans. The locations, ROIs of these humans are advertised in PSNs to use for the advanced application, such as name assignment, human tracking, and gesture evaluation. The module blocks, topics subscribing and publishing are depicted in Fig. 2. To detect and track humans, the PSN system uses four Kinects numbered as id1, id2, id3 and id4; two pairs (Kinects id1 and id2, and Kinects id3 and id4) are configured by the parallel setting; and two pairs (Kinects id1 and id3, and Kinects id2 and id4) are configured by the orthogonal setting.

3. Fusion Detected Human Method

Figure 3 contains the diagram of the proposed method. The figure shows that the system is divided into two main parts: the modules at the remote client site and the modules at the master server site. Note that the ROS-based method is the client and server platform where all the topics are controlled by the master node. The remote client side includes one Kinect connected with one mini pc such as Raspberry Pi [24]. The master node is installed on a workstation computer with a separate powerful graphic card. The simple processes, which require the low hardware configuration, are deployed at the clients. Complex and computational processes requiring a powerful hardware configuration are implemented at the master site.

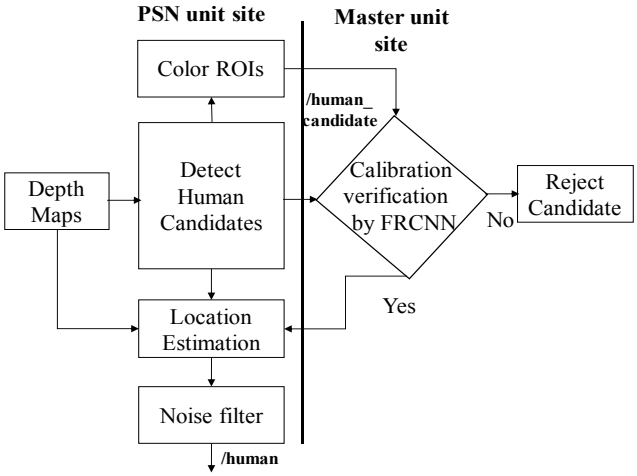


Fig. 3: Block diagram of proposed method.

At the client site, Openni-based human tracking is installed. Based on the depth images ROS topics provided by Kinect, human joint structures are obtained. Once the human candidates have been detected and human joints constructed, Openni requires certain amount of time to calibrate these joints structure with the standard joints structure in the trained database. Where the calibration has failed, the human location is not available as one can observe in Fig. 4. Otherwise, the location of detected humans is advertised as ROS topics call /detector. It is worth noting that the calibration is the main disadvantage of Openni. Besides, the human-look alike-objects can be falsely detected as humans, especially where there are objects such as chairs (see Fig. 5). To overcome the above-described limitations and to enhance the detection time as well as to reduce the detection errors for human-look alike-objects, the calibration step within Openni at the remote client side is ignored. Furthermore, to replace the verification roles of Openni calibration steps, the FRCNN installed at the server site will be used. Specifically, the flow chart of ROS topics running in this system can be observed in Fig. 2. The human candidates of Openni are packed and published as /human_candidate topics then subscribed by the master node. The detector candidate topics indicate the initial human ROI and location parameters, which possibly include the false errors. These parameters are refined and corrected at the master site. In order to do

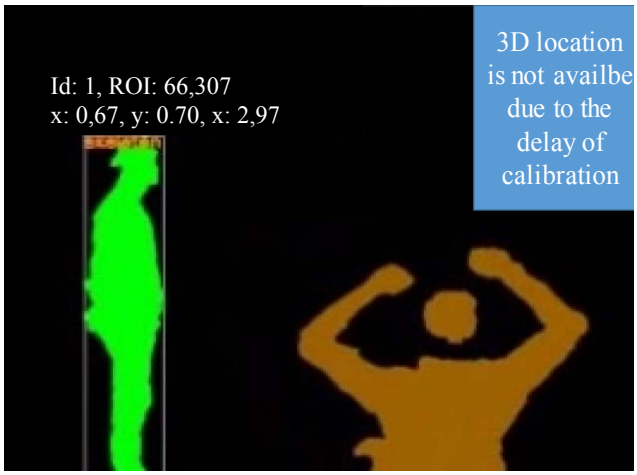
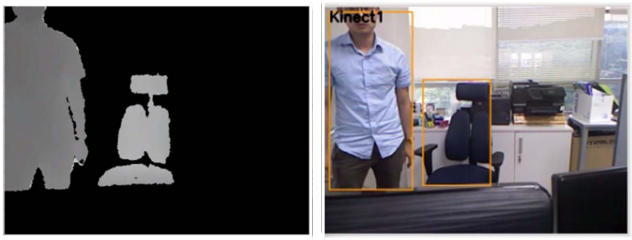


Fig. 4: 3D location of calibrating successfully and unsuccessful cases.



(a) (b)

Fig. 5: Object (chair) is considered as detected human. (a) human candidate masks in depth image, (b) human candidate ROIs are denoted as rectangles.

so at the master site, one module subscribes these topics as the inputs for faster FRCNN. Note that for object classification tasks of the original FRCNN, the first step is to perform object localization within the tested images. This step is the most time-consuming feature of the deep learning-based object classification methods. In the proposed method, object localization is replaced by the results of Openni. Specifically, from subscribing human candidate ROIs provided by Openni, the corresponding ROIs from a colour image of Kinect are published as ROS topics and are subscribed by the FRCNN module at the master.

The next step is that the FRCNN use these regions, considered as the human candidate in colour image, to make the classification decisions, i.e. whether this candidate is a real human being. Normally the FRCNN can classify up to 300 proposal classes with VGG16 networks [21]. The assignment of the tested objects to individual classes is performed based on the percentage scores given by FRCNN networks. The sum of all classes' scores is 100 %. The class with the highest score is the class of the tested object. The proposed system uses FRCNN to ascertain whether the two human candidate locations are human or not. The human location is refined by only taking the ROI of the human candidate. This will relieve the region proposal step of FRCNN in finding the object candidate. As a result, these proposed modifications of the original FRCNN reduce the execution time significantly and make the system suitable for a simpler convolutional neural network, allowing addressing the human detection issues. In brief, our system is a detection system where Openni is in charge of the human localization tasks and FRCNN is in charge of human verification tasks. The candidates of Openni are accepted as humans if FRCNN confirms these candidates are humans. After refining ROI, the final /human topics including 3D location obtained from the correspondence ROI in depth image and refined ROI parameters are published in the ROS system.

To further reduce the complexity, the human candidate identified by Openni and confirmed by FRCNN will be recognized by the system as the human for several consecutive frames. If the 3D Euclidian distance calculated using Eq. (2) between the human location verified by FRCNN and location of human candidate provided by Openni is q^t, the verification processes of this candidate q^t by FRCNN at master, as described above, will be triggered. This process reduces the complexity and execution time because the ROS nodes, which instated FRCNN at the master, have to subscribe the human candidate topic after a certain period of time. The human location can be tracked at the next frame by the conventional tracking method such as particle filter or Kalman filter.

$$\text{dis}\left(p^t, q^t\right) = \sqrt{\left(x_p^t - x_q^t\right)^2 + \left(y_p^t - y_q^t\right)^2 + \left(z_p^t - z_q^t\right)^2}. \tag{1}$$

The ROI sizes provided by FRCNN are quite different between two consecutive frames of video sequences. In the real situation, the change of the ROI size between these frames is very little. To make the ROI of the final detected human more stable, it is recalculated by averaging the size of ROIs in the previous frame p^{t-1} and the current frame. The size of ROI of the previous frame does not exceed threshold T_r in comparison with the current frame ROI of person p^t as shown in Eq. (2). In this paper, we set $T_r = 400$. Where this threshold is exceeded, the current size returned by Openni or FRCNN is ignored.

$$\text{size}\left(\text{ROI}\left(p^t\right)\right) - \text{size}\left(\text{ROI}\left(p^{t-1}\right)\right) < T_r. \tag{2}$$

In addition, to differentiate between human and chair object during the human candidate verification processes of FRCNN, the score for the chair object is considered. If this score exceeds the predefined threshold, this human candidate will be discarded by the system.

4. Experimental Results

The PSN system uses four Kinects whose Fields of View (FOV) partially overlapped. Note that the parallel configurations are applied to the pair of Kinects 1 and 2, and the pair of Kinects 3 and 4. The orthogonal configurations are applied to the pair of Kinects 1 and 3, and the pair of Kinects 2 and 4. This configuration increases the FOV of the system to be robust in detecting any partially overlapping humans. The detected human from Kinects is fused to remove the redundant and keep the human with the bigger ROI. The results of the proposed method are compared with the conventional Openni method and FRCNN method in terms of subjective evaluations and numerical evaluations. We set up the test scenarios, which include the group of people appearing in the FOV of Kinects. The setting can be observed in Fig. 6(b), with the two human staying close and touching together creating one group. In addition, in the testing environment, there are several human-alike-objects, such as chair, door, and table. In terms of subjective comparisons, the results returned by the Openni method are displayed in Fig. 6. Note that the calibration step of Openni is the step to remove the false detections but this burden takes considerable time to finish. We modified the original Openni by removing the calibration step of Openni. As the calibration steps in Openni have been disregarded to speed up detection processes, look-like-humans, such as chair, table, and door, can easily be falsely detected as humans. Furthermore, detecting a group of humans touching each other as a single human being is another problem and can be observed for the group of two humans in FOV of Kinect 1. These errors can be clearly seen in Fig. 7(a). The results of the proposed method, which combines Openni and FRCNN, fix the false detection and touching objects issues of the Openni method (see Fig. 7(b)). The human candidates of the Openni method (the rectangular ROI) in Fig. 7(a) are sent to the PSN units to be verified by FRCNN at the remote master unit. Note that at the master unit, the powerful hardware configuration that supports the advanced deep learning computations, is deployed. Due to the accuracy of FRCNN, the falsely detected ROI regions of the door, chair, and table are eliminated, allowing identification of the human being. Furthermore, the ROIs of the touching human group are classified as two humans with separated ROIs by FRCNN. These ROIs detected by FRCNN are sent back to the PSN units to calculate 3D locations. Figure 8 displays other results, showing the benefits of the proposed method. The human and the chair in the first frame are detected by Openni as humans. After the verification by FRCNN, only blue ROIs are classified as belonging to a real human being.

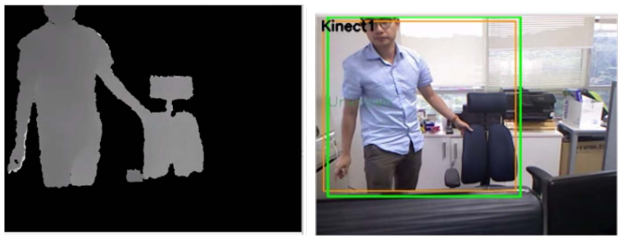


Fig. 6: Touched objects are considered to belong to human. (a) human candidate mask in depth image, (b) human candidate ROI is denoted as rectangle.

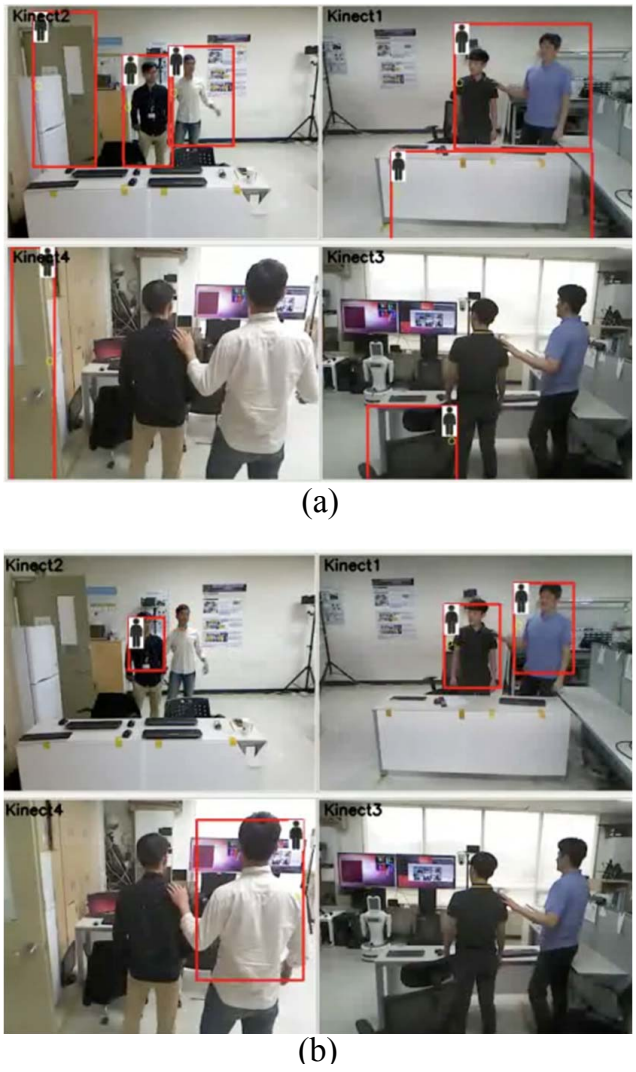


Fig. 7: Touched objects are considered to belong to human. (a) human candidate mask in depth image, (b) human candidate ROI is denoted as rectangle.

Table 2 displays the numerical comparisons between the proposed method and the Openni method. The criterions Miss Rate (MR), False Positive Rate (FPR), are calculated for each tested tracking method. Specifically, the miss rate indicates the ratio of the number of frames where the human is mistracked to the num-

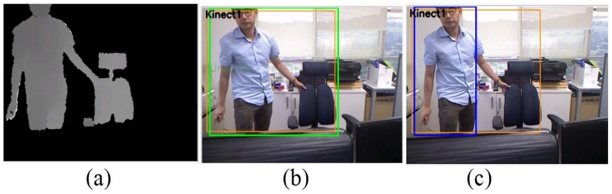


Fig. 8: ROI separation by FRCNN for touching objects (a) human candidate mask in depth image, (b) human candidate ROI is denoted as rectangle, (c) the real human ROI is denoted as blue rectangle.

Tab. 2: Accuracy Comparisons.

Method	Miss rate	False positive rate
Openni with calibration	9.86 %	8.19 %
FRCNN	4.46 %	4.69 %
Proposed method	5.53 %	5.22 %

ber of the frames where a human is tracked. The false positive rate indicates the ratio of the false positives to the sum of false positives and true negatives. One can observe that the results of the proposed method are significantly higher than those obtained by Openni and slightly lower than those obtained by FRCNN.

As regards the computational complexity (Tab. 3), the detection time of the proposed method is compared with FRCNN and Openni with calibration. The detection time of the proposed method includes the human candidate detection time at the PSN units by Openni without any calibration and the verification time by FRCNN at the master units. As regards FRCNN method, the execution time includes the time of human localization and classification within the colour images given by the Kinect cameras. Note that the colour topic from the Kinect camera is the colour video sequence with the resolution 640×480 and the ROI size provided by Openni is much smaller than the original size of the Kinects colour image. Since the detection step of FRCNN in the proposed method is reduced to a smaller size of human candidate ROIs provided by Openni, (see Tab. 3) the running time of proposed method is about two times lower than that of the original FRCNN. One of the advantages of the proposed method is that the full resolution colour images topics of Kinects do not need to be published by the PSN units in the PSN system. If these topics have to be transmitted in the system to be used for human detection by the original FRCNN method, the network load will increase significantly. The network load table for the cases of the proposed method and original method is provided in Tab. 4 for cases the PSN system includes from one to four Kinects. For the proposed method, the network load is reduced sharply.

Tab. 3: Execution time comparisons.

Method	Processing time
Openni with calibration	0.80 s
FRCNN	0.32 s
Proposed method	0.18 s

Tab. 4: Network Load Comparisons.

Method	FRCNN	Proposed method
1 Kinect	2 Mb·s^{-1}	1.5 Mb·s^{-1}
2 Kinect	2.8 Mb·s^{-1}	1.9 Mb·s^{-1}
3 Kinect	3.5 Mb·s^{-1}	2.5 Mb·s^{-1}
4 Kinect	4.1 Mb·s^{-1}	2.9 Mb·s^{-1}

5. Conclusion

This paper addresses the issue of the detection of multiple humans by means of multiple Kinect cameras using the fusion approach between depth and colour information. The respective advantages of Openni and FRNN have been selectively incorporated. This approach can yield accurate detection results even in challenging environment such as the noise of 3D human positions when multiple humans stay close together. In addition, the proposed method integrated as a package within the Robot Operation System will help robots to deal with the human beings tracking efficiently.

Acknowledgment

The research was partly supported by the R&D programs of MOTIE (10041629 [SimonPiC] and 10077468 [DeepTasK]) and by ICT R&D programs of IITP (2015-0-00197 [LISTEN] and 2017-0-00432 [BCI]) and also received a support from the SGS grant No. SP2017/174, VSB–Technical University of Ostrava.

References

[1] NGUYEN, D. T., W. LI and P. O. OGUNBONA. Human detection from images and videos: a survey. *Pattern Recognition*. 2016, vol. 51, iss. 1, pp. 148–175. ISSN 0031-3203. DOI: 10.1016/j.patcog.2015.08.027.

[2] NASEER, T., J. STURM and D. CREMER. FollowMe: Person following and gesture recognition with a quadrocopter. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Tokyo: IEEE, 2013, pp. 624–630. ISBN 978-1-4673-6358-7. DOI: 10.1109/IROS.2013.6696416.

[3] ZHAO, T. and R. NEVATIA. Tracking multiple humans in crowded environment. In:

IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2004, pp. 406–413. ISBN 0-7695-2158-4. DOI: 10.1109/CVPR.2004.1315192.

[4] DALAL, N. and B. TRIGGS. Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington: IEEE, 2005, pp. 886–893. ISBN 0-7695-2372-2. DOI: 10.1109/CVPR.2005.177.

[5] ZHU, Q., M.-C. YEH, K.-T. CHENG and S. AVIDAN. Fast human detection using a cascade of histograms of oriented gradients. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2006, pp. 1491–1498. ISBN 0-7695-2597-0. DOI: 10.1109/CVPR.2006.119.

[6] SUYKENS, J. A. K. and J. VANDEWALLE. Least squares support vector machine classifiers. *Neural processing letters*. 1999, vol. 9, iss. 3, pp. 293–300. ISSN 1370-4621. DOI: 10.1023/A:1018628609742.

[7] DOLLAR, P., C. WOJEK, B. SCHIELE and P. PERONA. Pedestrian detection: A benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Miami: IEEE, 2009, pp. 304–311. ISBN 978-1-4244-3992-8. DOI: 10.1109/CVPR.2009.5206631.

[8] DOLLAR, P., C. WOJEK, B. SCHIELE and P. PERONA. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2012, vol. 34, iss. 4, pp. 743–761. ISSN 0162-8828. DOI: 10.1109/TPAMI.2011.155.

[9] XIA, L., C.-C. CHEN and J. K. AGGARWA. Human detection using depth information by kinect. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Colorado Springs: IEEE, 2011, pp. 15–22. ISBN 978-1-4577-0530-4. DOI: 10.1109/CVPRW.2011.5981811.

[10] ZHANG, H. and L. E. PARKER. CoDe4D: Color-Depth Local Spatio-Temporal Features for Human Activity Recognition From RGB-D Videos. *IEEE Transactions on Circuits and Systems for Video Technology*. 2016, vol. 26, iss. 3, pp. 541–555. ISSN 1051-8215. DOI: 10.1109/TCSVT.2014.2376139.

[11] ZHANG, Z. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia*. 2012, vol. 19, iss. 2, pp. 4–10. ISSN 1070-986X. DOI: 10.1109/MMUL.2012.24.

[12] Package Summary. In: *ROS.org* [online]. 2016. Available at: `http://wiki.ros.org/openni_kinect`.

[13] SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural networks*. 2015, vol. 61, iss. 1 pp. 85–117. ISSN 0893-6080. DOI: 10.1016/j.neunet.2014.09.003.

[14] JI, S., W. XU, M. YANG and K. YU. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013, vol. 35, iss. 1, pp. 221–231. ISSN 0162-8828. DOI: 10.1109/TPAMI.2012.59.

[15] WENINGER, F., J. BERGMANN and B. SCHULLER. Introducing CURRENNT: The Munich open-source CUDA RecurREnt neural network toolkit. *Journal of Machine Learning Research*. 2015, vol. 16, iss. 3, pp. 547–551. ISSN 1532-4435.

[16] CHETLUR, S., C. WOOLLEY, P. VANDERMERSCH, J. COHEN, J. TRAN, B. C. CATANZARO and E. SHELHAMER. CuDNN: Efficient Primitives for Deep Learning. In: *CoRR* [online]. 2014. Available at: http://arxiv.org/abs/1410.0759.

[17] GIRSHICK, R., J. DONAHUE, T. DARRELL and J. MALIK. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2016, vol. 38, iss. 1, pp. 142–158. ISSN 0162-8828. DOI: 10.1109/TPAMI.2015.2437384.

[18] UIJLINGS, J. R. R., K. E. A. VAN DE SANDE, T. GEVERS and A. W. M. SMEULDERS. Selective Search for Object Recognition. *International Journal of Computer Vision*. 2013, vol. 104, iss. 2, pp. 154–171. ISSN 0920-5691. DOI: 10.1007/s11263-013-0620-5.

[19] HE, K., X. ZHANG, S. REN and J. SUN. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2015, vol. 37, iss. 9, pp. 1904–1916. ISSN 0162-8828. DOI: 10.1109/TPAMI.2015.2389824.

[20] GIRSHICK, R. Fast R-CNN. In: *IEEE International Conference on Computer Vision*. Santiago: IEEE, 2015, pp. 1440–1448. ISBN 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.169.

[21] SIMONYAN, K. and A. ZISSERMAN. Very deep convolutional networks for large-scale image recognition. In: *CoRR* [online]. 2014. Available at: `http://arxiv.org/abs/1409.1556`.

[22] REN, S., K. HE, R. GIRSHICK and J. SUN. Faster R-CNN: Towards real-time object detection with region proposal networks. In: *CoRR* [online]. 2015. Available at: http://arxiv.org/abs/1506.01497.

[23] QUIGLEY, M., K. CONLEY, B. GERKEY, J. FAUST, T. FOOTE, J. LEIBS, R. WHEELER and A. Y. NG. ROS: an open-source Robot Operating System. In: *Willowgarage* [online]. 2009. Available at: http://www.willowgarage.com/sites/default/files/icraoss09-ROS.pdf.

[24] HALFACREE, G. and E. UPTON. *Raspberry Pi User Guide*. Chichester: John Wiley & Sons Ltd, 2014. ISBN 978-1-118-46446-5.

About Authors

Anh Vu LE is a member of the Optoelectronics Research Group, at the Faculty of Electrical and Electronics Engineering of the Ton Duc Thang University in Ho Chi Minh City, Vietnam. He obtained his M.Sc. and Ph.D. degrees in Electronics and Electric from the Dongguk University in 2012 and 2015, respectively. He has been a visiting scientist at center for robotic research, Korea Institute of Science and Technology. His current research interests include robotics vision, human detection feature matching, 3D video and optoelectronics.

Tran Tin PHU (Corresponding author: phutrantin@tdt.edu.vn) was born in Khanh Hoa, Vietnam, in 1979. He received the B.Sc. degree (2002) and M.Sc. degree (2008) from Ho Chi Minh City University of Science. Currently, he is a lecturer at the Faculty of Electronics Technology (FET), Industrial University of Ho Chi Minh City. Since 2015, he has been participating in Ph.D program that had been collaborated between VSB–Technical University of Ostrava, Czech Republic and Ton Duc Thang University, Ho Chi Minh City. His major research interests are computer vision, wireless communication in 5G, energy harvesting, and performance of cognitive radio, physical layer security, and optoelectronics.

Jong Suk CHOI received Ph.D. degree from the Korea Advanced Institute of Science and Technology (KAIST) in 2001. He is currently a Principal Research Scientist at the Korea Institute of Science and Technology (KIST), and Professor of the Korea University of Science and Technology (UST). His research interests include robot audition, sound classification, and sound source localization, audio-visual perception for human, human-robot interaction, and mobile robot navigation.

Jan SKAPA obtained his Ph.D. degree in telecommunication engineering from VSB–Technical University of Ostrava, Czech Republic in 2009. Since then, he has worked as an assistant professor in Department of Telecommunications. His research interests include digital signal processing, speech and image processing and the use of the wavelet transform in engineering applications.

Miroslav VOZNAK obtained his Ph.D. in 2002 in the telecommunication engineering and was appointed an associate professor in 2009 at the Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava (VSB–TUO). Since 2013, he has led the Department of Telecommunications at the VSB–TUO as the department chair. He is an IEEE Senior member actively engaged in numerous IEEE conference committees. His research interests focus generally on information and communication technology, particularly on Voice over IP, quality of experience, network security, wireless networks and in the last couple years also on Big Data analytics in mobile cellular networks.