Fusing Detected Humans in Multiple Perception Sensors Network

Hoang Quang Minh TRAN, Anh Vu LE*

Optoelectronics Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

*leanhvu@tdt.edu.vn

Abstract. *A fusion method is proposed to keep a correct number of humans from all humans detected by the robot operating system based Perception Sensor Network (PSN) which includes multiple partially overlapped Field of View (FOV) Kinects. To this end, the fusion rules are based on the parallel and orthogonal configurations of Kinects in PSN system. For the parallel configuration, the system will decide whether the detected humans staying in FOV of single Kinect or in overlapped FOV of multiple Kinects by evaluating the angles formed between their locations and Kinect original point on top view (x, z plane) of 3D coordination. Then, basing on the angles, the PSN system will keep the person stay in only one FOV or keep the one with biggest ROI if they stay in overlapped FOV of Kinects. In the case of Kinects with orthogonal configuration, 3D Euclidian distances between detected humans are used to determine the group of humans supported to be same human but detected by different Kinects. Then the system, keep the human with a bigger Region of Interest (ROI) among this group. The experimental results demonstrate the outperforming of the proposed method in various scenarios.*

Keywords

Human detection, robot operation system, sensor fusion.

1. Introduction

Detecting and tracking individual human in the cluster environments such as in the group of humans is an interesting area in recent years. In these scenarios, detecting the appearances of humans is challenging problem since the humans are normally partially or fully overlapped by other humans in the scene and humans appearing out of sensor FOV [1]. Multiple sensors fusion is one of efficient approaches to solve the these problems [2]. Sensor fusion in computer vision is classified into several categories: scene segmentation, representation, 3-D shape estimation, sensor modeling, autonomous robots, and object detection and tracking. In [3], a survey of the state of the art in multi-sensor fusion is presented. The research of [4] proposed a method in which 3D positions of detected humans by multiple sensors are used to track the detected one. Using multiple sensors in [5] enhances the robot abilities to track human in rescues situations. The works of [6] have proved that information from multiple sensors increases the accuracy of the human actions recognition. In [7] and [8], pedestrian tracking using the multiple sensors and optical follow are researched and these proved that cluster environments are the challenge problem of human detection and tracking. The issues of multimodal fusion are revised in the works of [9]. In the works of [10], a number of fusion strategies have been employed to combine sensor outputs from RGB-depth camera and depth sensor to enhance the integrity of

depth data in detection object. To improve the navigation of robotic cars in cluster environment with obstacles, [11] propose a vision multisensor fusing framework, which fuses color, depth, and laser information consistently via both geometrical and semantic constraints. Sensor fusion is used in [12] to address the problem of object detection and tracking in occlusion scenario. The redundant data is determined if the Euclidian distance between objects is lower than threshold. This fusion rule easily lead to miss detection after fusing in cases different objects staying closely each other. A proposed PSN system in [13] and [14] including various Kinect cameras (PSN units) [15] playing a role as system perception parts is used to detected and tracking humans appearing in the FOV of Kinects in PSN. Since the PSN system consists of multiple Kinects and the FOV of two Kinects is configured to overlap partially each other, it is possible that someone staying in the overlapped region is detected by both Kinects. The fusion rules of [13] and [14] is similar with [12], they use the distances between detected objects to find the same object but are detected by different Kinent cameras. This fusion rule easily lead to miss detection after fusing in cases different objects staying near each other. The requirement of keeping a correct number of humans from all detected humans including redundant detected humans is the crucial requirement to enhance the detecting and tracking tasks. To this end, in this paper a novel fusing methods will accumulate the location information of all detected humans from PSN units and then remove redundant information if any. Moreover, these fusion methods are proposed to tolerate the calibration errors of Kinects setting Specifically, Kinect configurations are classified into parallel and orthogonal. The fusion rules of parallel configuration are based on the decisions whether detected human belong to FOV of one Kinect or belong to both FOV of two Kinect. On another hand, 3D locations of detected humans are used in the orthogonal configuration to determine the group of humans supporting the same human but detected by different Kinects.

The following is the structure of the paper. Section 2. will introduce the PSN system. Section 3. presents the fusion method. The experimental results of proposed method are provided in Sec. 4. .

2. PSN System Overview

The Robot Operation System (ROS)-based system [16] are used to detect humans. Note that ROS basing on master and clients platform provides the fundamental communication environments for nodes. Each PSN unit acts as the single node of ROS and communicates with other PSN units under the controls of server playing a role as master of ROS. In PSN, the novel ROS based package called Openni [17] installed in PSN units which use Kinects to detect the humans appearing in Kinect FOVs then to provide the 3D distance and 2D ROI of these detected humans. Specifically, the locations of humans can be detected easily by finding the human body joints which are supported very effectively by the available open source libraries [18] and [19]. After the presenting of one detected human, this human is tracked by PSN and location and name is published as ROS topics in the system.

Each software module acting as a node of ROS publishes as well as subscribers the ROS topic messages. The PSN system consists of WHERE module installed at PSN units using Kinect Openni to detect humans. The human locations are advertised as ROS topics on the system and ROS nodes can subscribe to utilize this information conveniently. The module blocks, topics subscribing and publishing are depicted in Fig. 1. WHERE fusion module installed at master will accurate all detected humans by subscribing the /detector topics advertised by PSNs units and uses fusion rules to keep the correct number of humans. The final human locations ROIs after fusing (/where topics) are advertised in PSNs for advanced application such as tracking and gesture evaluating.

Specifically, in PSN, k_n ($n \in \{1,2\}$) is the 3D origin position of Kinect n and and the 3D position of a person i at frame t detected by Kinect k^n defined as $p_i^t(k^n)$ and ROI of this person at frame t is defined as ROI$\{p_i^t(k^n)\} = (x_i^t, y_i^t, w_i^t, h_i^t)$ where w_i^t and h_i^t mean width and

height respectively as in Fig. 2. To detect and track human, PSN system uses four Kinects numbered as $id1$, $id2$, $id3$ and $id4$ in which the pair Kinects $id1$, $id2$ and the pair Kinects $id3$ and $id4$ are configured by parallel setting and the pair Kinects $id1$, $id3$ and the pair Kinects $id2$, $id4$ are configured by orthogonal setting. Note that in parallel setting, the two local origin of Kinects e.g. k^n and global origin O of PSN system are on the same line and in orthogonal setting, the local origins of 2 Kinects are configured so that they face oppositely each other. The parallel setting of two Kinects of PSN is configured as shown in Fig. 3 which depicts a top view (xz-plane) and the orthogonal setting of two Kinects of PSN is configured as shown in Fig. 4 which also depicts a top view (xz-plane) of global 3D coordination.
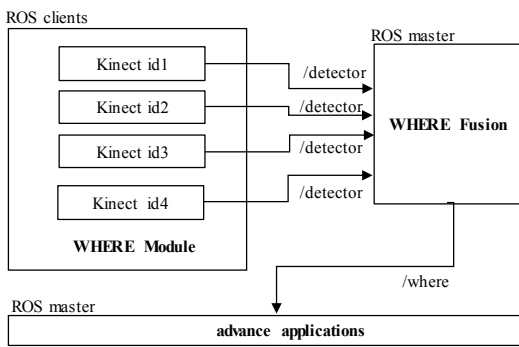

Fig. 1: Modules and message transmissions of PSN.

$$p_i^t(k^n) = (x_i^t, y_i^t, z_i^t) \qquad ROI\{p_i^t(k^n)\} = (x_i^t, y_i^t, h_i^t w_i^t)$$
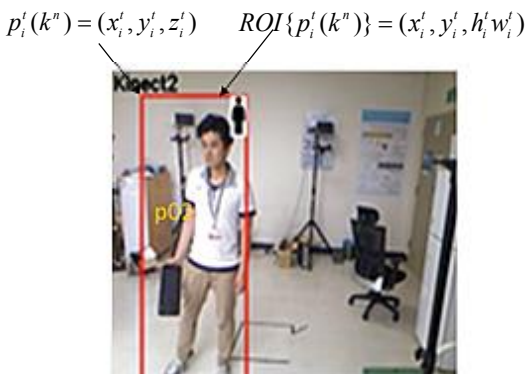

Fig. 2: Features of detected human in FOV of Kinect n: 3D location at top left corner and ROI.

3. Proposed Fusion Method

The Kinects in PSN system are roughly calibrated manually in advance. Since the effective FOV of Kinect is about three meter, the offset value after calibration in horizontal axis of Kinects is set to 3 meters and in vertical axis is set to 5 meters. The origin of global coordinate is set at the middle point on the line connected local origins of Kinects $id1$ and $id2$. Then the calibration parameters are stored in the database of PSN system and published as ROS topics. The PSN units subscribe these calibrations parameters topics to convert the local coordinate of each Kinect to global coordinate of PSN system. As the results, the 3D locations of detected human of Kinects are converted to global 3D locations. If the calibration steps are implemented perfectly and the accurate calibrating parameters can be obtained, the detected people supporting to be the same person but detecting by different Kinects will have the same 3D locations. In this situation, removing redundant humans is straightforward by keeping the detected humans with the biggest ROI. In reality, doing the calibration for a system with multiple sensors is the challenging problem, calibration errors normally occur in this system. In our PSN system, the Kinects are calibrated roughly basing on orthogonal configuration and parallel configuration which denoted in Fig. 3 and Fig. 4 respectively. Specifically, the offset in horizontal and vertical axis (xz-plane) of Kinects with origin is measured and stored as calibration parameters of each Kinects. Fusion rules are proposed to tolerate the calibration errors which makes the offsets of 3D distances of humans supporting the same person detected by difference Kinects.

The parallel setting of a Kinect pair in PSN system is depicted in Fig. 3. As one can observe, Kinect pair forms a panorama view. To overcome the problem of calibration errors, the fusion rules of the parallel configuration of Kinects pair are based on the angles, which are formed between the positions of detected humans and the Kinect global origin positions. Specifically, the angle between a person p_i^t and k^n is θ_i^n. The angle between one detected human and

Kinect can be calculated from the position values, for instance $\tan(\theta_i^n) = (Ok^n - |x_i^t|) / |z_i^t|$. The FOV of Kinect id n is shown in terms of angle θ^n Here, default FOV parameters of Kinect are used so that each θ^n is $43°$. In case of human i belongs to the FOV of only Kinect $id1$ denoted as $p_i^t(k^1)$ (region 1 in Fig. 3), or belongs FOV of only Kinect $id2$ denoted as $p_i^t(k^2)$ (region 2 in Fig. 3), the fusion rules can be expressed as Eq. (1) and Eq. (2):
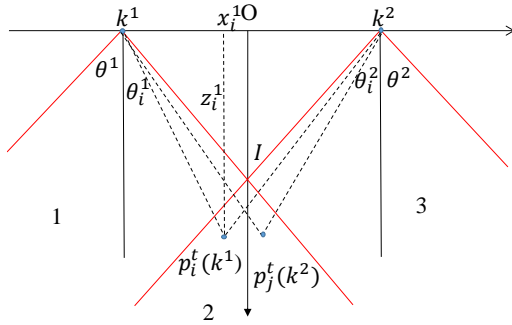

Fig. 3: Parallel cofiguration.

$$\text{if } \theta_i^1 \leq \theta^1 \wedge \theta_i^2 \geq \theta^2$$
$$\text{then } p_i^t \in \text{FOV}\left(k^1\right) \wedge p_i^t \notin \text{FOV}\left(k^2\right) \quad (1)$$
$$\wedge \text{ fusionrulekeeps } p_i^t\left(k^1\right),$$

$$\text{if } \theta_i^1 \geq \theta^1 \wedge \theta_i^2 \leq \theta^2$$
$$\text{then } p_i^t \notin \text{FOV}\left(k^1\right) \wedge p_i^t \in \text{FOV}\left(k^2\right) \quad (2)$$
$$\wedge \text{ fusionrulekeeps } p_i^t\left(k^2\right).$$

In another case, one human belongs to both FOV of Kinect $id1$ and Kinect $id2$ (region 3 in Fig. 3), these humans are detected as $p_i^t\left(k^1\right)$ and $p_j^t\left(k^2\right)$, the fusion rules will select the human from either Kinect $id1$ or Kinect $id2$. The human with the bigger ROI will be kept after fusing as depicted in Eq. (3). Note the size of ROI of people are calculated by multiplying the width and the high of their ROI:

$$\text{if } \theta_i^1 \leq \theta^1 \wedge \theta_i^2 \leq \theta^2$$
$$\text{then } p_i^t \in \text{FOV}\left(k^1\right) \wedge p_j^t \in \text{FOV}\left(k^2\right)$$
$$\wedge \text{ fusionrulekeeps } p_i^t\left(k^1\right)$$
$$\text{if size}\left(\text{ROI}\left\{p_i^t\left(k^1\right)\right\}\right) > \text{size}\left(\text{ROI}\left\{p_j^t\left(k^2\right)\right\}\right)$$
$$\text{then fusionrulekeeps } p_i^t\left(k^1\right)$$
$$\text{else fusionrulekeeps } p_j^t\left(k^2\right). \quad (3)$$

In the case of orthogonal configuration, the Kinects pairs are set to face oppositely each other and they do not form a panorama view as can be observed in Fig. 4. As the results, it is difficult to determine that the person appears in the FOV of one Kinect or in FOVs of both Kinect as described in the parallel configuration. The fusion rules of the orthogonal setting of two Kinects in PSN system as Fig. 3 to remove redundant detected people are based on the relative distances between 3D global locations of all detected peoples. To remove redundant humans after humans are detected by all Kinects of PSN system (/detector topics), firstly the 3D Euclidian distance between two humans i and j at frame t is calculated as in Eq. (4):

$$\text{dist}\left(p_i^t, p_j^t\right) = \sqrt{\left(x_i^t - x_j^t\right)^2 + \left(y_i^t - y_j^t\right)^2 + \left(z_i^t - z_j^t\right)^2}. \quad (4)$$

Secondly, detected person j is considered as the same person with person i if the relative distance of 3D positions of these people are lower than threshold Td and Kinect ID of person i which indicates the name of Kinect detects this person is different from Kinect ID of person j. Using only the 3D distance condition can lead to unwanted removing detected human in cases of there are many detected humans staying close together. To overcome this limitation, considering each Kinect ID which is different from Kinect ID of to be checked person i, the following conditions are used to further check whether the other detected people are the same. Note that for the person appears in FOV of one Kinect, this Kinect can detect this person as one object. That is, if there are more than one people having same this Kinect ID and 3D distances of these people and person i are lower than distance Td, only one person who yields the smallest distance with person i will be chosen as person i detected by this Kinect ID. The same method is used for all Kinect in PSN system. The Eq. (5)

describes the above conditions. In our system, Kinects are calibrated to convert each local 3D position detected by Kinect coordinate to global PSN coordinate. If Kinects are calibrated perfectly, the 3D Euclidian distances calculated by Eq. (4) approach to zero. In our PSN system, we set $Td = 0.5$ to tolerate the calibration errors between Kinects. Finally, supporting that for one detected person there are numbers of people detected by PSN Kinects who are considered as the same person after above considerations, only one person with biggest ROI will be kept and others will be removed after fusing. Eq. (5):

$$\text{if dist}\left(p_i^t, p_{j*}^t\right) < Td$$
$$\wedge \text{ kinect id}\left(p_i^t\right) \neq \text{kinect id}\left(p_{j*}^t\right)$$
$$\wedge j = \underset{j* \in D}{\arg\min}\left(\text{dist}\left(p_i^t, p_{j*}^t\right)\right)$$
$$\text{then } p_j^t \text{same as } p_i^t \qquad (5)$$
$$\text{if size}\left(\text{ROI}\left\{p_i^t\right\}\right) > \text{size}\left(\text{ROI}\left\{p_j^t\right\}\right)$$
$$\text{then fusionrulekeeps } p_i^t$$
$$\text{else fusionrulekeeps } p_j^t,$$

where D is the set of detected humans whose the distance calculated as Eq. (4) between each of their 3D locations and 3D location of to be checked person i is lower than threshold Td.
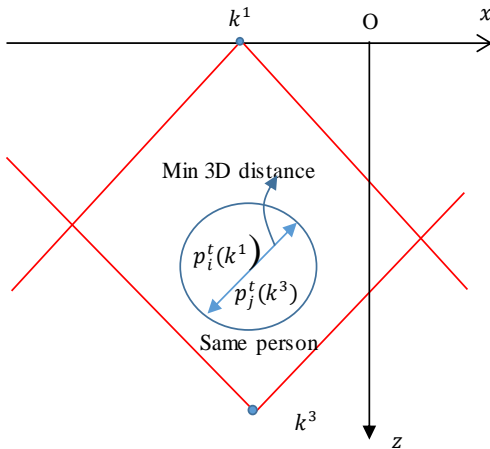

Fig. 4: Orthogonal configuration.

Obviously, the distant threshold Td affects the fusion results. If fusion rules applied for orthogonal configuration based on relative distances between 3D positions are used for parallel configuration, the detected humans can be removed incorrectly after fusing. The errors happen easily if the two different humans yield the distance that lower than the distant threshold. If this situation occurs, then only one person with the biggest ROI is kept and the other person is removed incorrectly after fusing. Therefore, instead of using 3D distance between humans, the fusion rules of parallel configuration of two Kinects in PSN system as Fig. 4 are based on the angles, which are formed between the human positions and the Kinect global origin positions. To overcome the fusion errors when applying distance-based rules to the orthogonal configuration, we should implement the calibration steps precisely to reduce the calibrations errors.

4.   Experimental Results

The sensitivity of Openni method, which is used to detect human, is about 0.5 seconds. This means that when human appear in the Kinect FOV, Openni takes 0.5 seconds to detect this human. In our experiments scenarios, the humans enter the Kinect FOV and stay static about one second at the defined positions before moving. Note that staying in these positions, which in the FOV of multiple Kinect guarantees that these humans are detected by at least one Kinect. The results of WHERE fusion for the two Kinects with parallel configurations are shown in Fig. 5. The detected humans are denoted as rectangular boxes. Note that the detected humans after fusing are denoted as the green boxes with a number and the redundant detected humans are denoted as yellow rectangle without numbering, respectively. In Fig. 5 the $p09$ belongs to both FOV of Kinect1 and FOV of Kinect2 after fusing the ROI of the human in Kinect1 is kept since the size of the ROI of this person is bigger as described in Eq. (3). On the other hand, $p01$, $p02$, $p03$ belong only to one Kinect FOV so it is kept by fusion rules.

The results of WHERE fusion for orthogonal configuration of two Kinect with $id1$ and $id3$ are shown in Fig. 6. There are six detected humans and after doing fusion the system removes three redundant detected ones and keep the corrects the number of humans appearing in the FOV of

Fig. 5: WHERE fusion of the parallel configuration (keeps correctly 4 detected humans (red ROI rectangular) from total 5 detected humans).

orthogonal configuration Kinect $id1$ and Kinect $id3$ as three people.

Fig. 6: WHERE fusion of orthogonal configuration (keeps correctly 3 detected humans (green ROI rectangular) form total 6 detected humans).

In Fig. 7, we can observe the problem when we use distance based rules to remove the redundant detected humans in the parallel configuration. Selecting the distance threshold Td affects the results of fusion. If the large threshold Td is selected there are more detected people will

be considered as the same human. The people who stay near the border of Kinects will have the small 3D distances. These distances can be lower than distance threshold Td (are set as 0.5) which use to determine the same person from detected people. The people are denoted as $p01$, $p02$, $p03$ are considered as the same person and after using fusing rules based on 3D location, the human $p03$ are removed incorrectly and is not shown as the green boxes in final fusion results.
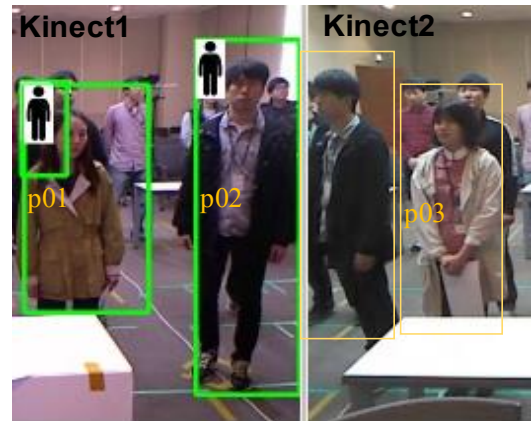

Fig. 7: WHERE fusion keeps only 1 detected human instead of two human staying near the border of two Kinects.

The results with four Kinects configurations are showed in Fig. 8 and Fig. 9. Note that the parallel configurations are applied for the Kinect $id1$ and Kinect $id2$ and the Kinect $id3$ and Kinect $id4$ pairs. The orthogonal configurations are applied for the Kinect $id1$ and Kinect $id3$ and the Kinect $id2$ and Kinect $id4$ pairs. The detected humans are denoted as rectangular boxes together with their correspondence names. As the results of where fusion method, the redundant humans that detected by Kinects are removed and the number of detected humans and their ROIs as well as names are displayed correctly by the PSN system.

The proposed method is compared with the method used in reference [14]. Note that the fusion rules of [14] are the same as [12] and [13]. The results are provide in Fig. 9. As one can see in Fig. 9(b), total numbers of humans in FOV of four Kinects are 16 and after using proposed fusion rules, the number of humans are kept correctly from 20 detected humans. On the other

Fig. 8: WHERE fusion keeps correctly 4 detected humans (green ROI rectangular) from total 9 detected humans in four Kinects with parallel and orthogonal configurations

hand, using methods in [14], two human at the border of Kinects are not kept after fusion. Since the reference [14] use only one distance criterion of detected humans to determine the same human and the redundant data is determined if the 3D Euclidian distance between objects is lower than threshold, this fusion rule remove humans wrongly if different humans stay closely each other. One other hand, the fusion rules of proposed method are based on the Kinects configuration. For orthogonal configuration, beside the distance criterion, the proposed method uses other criterions such as Kinect id of detected human and smallest distant to determine the same human detected by different Kinects. For parallel configuration, the fusion rules are based on the angles formed between detected human and Kinect origin. The combination of these fusion rules eliminates unwanted cases of the removing detected humans.

5. Conclusions and Future Works

Using the distances between humans in orthogonal configuration and angles between detected humans and Kinects origins the PSN system in parallel configuration, the proposed fusion methods can remove the redundant detected humans and keep correct number of detected humans ef-

ficiently. The proposed fusion method enhances the human detection and tracking. Testing the calibration errors is the potential future research topics.

Fig. 9: Comparison between proposed method with method [14]. (a) Result of [14], (b) result of proposed method.

Acknowledgment

References

[1] 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.

[2] ONG, K. S., Y. H. HSU and L. C. FU. Sensor fusion based human detection and tracking system for human-robot interaction. In: *International Conference on Intelligent Robots and Systems.* Vilamoura: IEEE, 2012, pp. 4835–4840.

[3] AGGARWAL, J. K. *Multisensor fusion for computer vision.* New York: Springer-Verlag, 1993.

[4] KNOOP, S., S. VACEK and R. DILLMANN. Sensor fusion for 3D human body tracking with an articulated 3D body model. In: *International Conference on Robotics and Automation.* Orlando: IEEE, 2006.

[5] NOURBAKHSH, I. R., K. SYCARA, M. KOES, M. YONG, M. LEWIS and S. BURION. Human-robot teaming for search and rescue. *IEEE Pervasive Computing.* 2005, vol. 4, iss. 1, pp. 72–79.

[6] TANG, C., C. ZHOU, W. PAN, L. XIE and H. HU. Fusing mixed visual features for human action recognition. *International Journal of Modelling, Identification and Control.* 2013, vol. 19, iss. 1, pp. 13–22.

[7] NOAMAN, R. A. K., M. A. M. ALI and N. ZAINAL. Enhancing pedestrian detection using optical flow for surveillance. *International Journal of Computational Vision and Robotics.* 2017, vol. 7, iss. 1–2, pp. 35–48.

[8] JIANG, Y., H.-G. WANG and N. XI. Target Object Identification and Location Based on Multi-sensor Fusion. *International Journal of Automation and Smart Technology.* 2013, vol. 3, no. 1, pp. 57–65.

[9] ISMAIL, A. W. and M. S. SUNAR. Multimodal fusion: progresses and issues for augmented reality environment. *International Journal of Computational Vision and Robotics.* 2017, vol. 7, no. 3, pp. 240–254.

[10] SONG, H., W. CHOI and H. KIM. Robust Vision-Based Relative-Localization Approach Using an RGB-Depth Camera and LiDAR Sensor Fusion. *IEEE Transactions on Industrial Electronics.* 2016, vol. 63, iss. 6, pp. 3725–3736.

[11] XUE, J., D. WANG, S. DU, D. CUI, Y. HUANG and N. ZHENG. A vision-centered multi-sensor fusing approach to self-localization and obstacle perception for robotic cars. *Frontiers of Information Technology & Electronic Engineering.* 2017, vol. 18, iss. 1, pp. 122–138.

[12] CHAVEZ-GARCIA, R. O. and O. AYCARD. Multiple Sensor Fusion and Classification for Moving Object Detection and Tracking. *IEEE Transactions on Intelligent Transportation Systems.* 2016, vol. 17, iss. 2, pp. 525–534.

[13] AN, K., J. PARK, M. D. HOANG and J. CHOI. Dispensing materials of mobile robot cooperating with perception sensor network. In: *11th International Conference on Ubiquitous Robots and Ambient Intelligence.* Kuala Lumpur: IEEE, 2014, pp. 496–499.

[14] PARK, J., K. AN and J. CHOI. Realistic 3D simulation of multiple human recognition over Perception Sensor Network. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication.* Edinburgh: IEEE, 2014, pp. 507–512.

[15] ZHANG, Z. Microsoft Kinect Sensor and Its Effect. *IEEE MultiMedia.* 2012, vol. 19, iss. 2, pp. 4–10.

[16] 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.

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

[18] A Software Architecture for RGB-D People Tracking Based on ROS Framework for a Mobile Robot. LEE, S., K.-J. YOON and J. LEE. *Frontiers of intelligent autonomous systems.* Studies in computational intelligence, v. 466. New York: Springer, 2013, pp. 53–68.

[19] XIA, L., C.-C. CHEN and J. K. AGGARWA. Human detection using depth information by Kinect. In: *Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* Colorado Springs: IEEE, 2011, pp. 15–22.

About Authors

Hoang Quang Minh TRAN is lecturer of Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. He is currently a member of Optoelectronics Research Group, Ton Duc Thang University. His Areas of expertise include Power system, Relay Protection, and Optoelectronics.

Anh Vu LE is at Faculty of Electrical and Elec-tronic Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. He received his B.Sc. in Electronics and Telecommunications from Ha Noi University of Technology, M.Sc. and Ph.D. degrees in Electronics and Electrics from the Dongguk University in 2007 and 2012, respectively. He is currently a member of Optoelectronics Research Group, Ton Duc Thang University. His current research interests include Robotics vision, human detection feature matching, 3D video processing, Optoelectronics.