Development of Field View Monitor 2

-An assisting function for safety check around a hydraulic excavator using real-time image recognition with monocular cameras

Yoshihisa Kiyota[a] , Shunsuke Otsuki[a], Susumu Aizawa[a] , Danting Li[a]

[a]Information & Communication Technology Dept, Technology Research Center, Sumitomo Heavy Industries, Ltd.
E-mail: yoshihisa.kiyota@shi-g.com, shunsuke.otsuki@shi-g.com,
susumu.aizawa@shi-g.com, danting.li@shi-g.com

Abstract –

We have developed a function to process the images of three monocular cameras mounted on a hydraulic excavator in real time, and to notify the user of any human-like images on the monitor and by sound. This function is composed of the following four functions. (1) Detection of the human head by random forest, (2) Normalization of the region of interest, (3)Evaluation of human presence based on luminance gradient features, and (4) Tracking based on the human and the excavator motion model. The algorithm is implemented in a FPGA unit, and the 3-camera images are input into the FPGA unit and processed by time-division manner to achieve both high-speed processing and low-cost.

Keywords –

Excavator; Image recognition; Human detection; Electronic Control Unit

1 Introduction

Hydraulic excavators are large, heavy, and fast-moving, and they do not only run but also turn. Operators are required to check the surrounding area reliably and over a wide area in order to ensure safety at sites where workers and obstacles are mixed in. However, there are many blind spots in the rear and right side of the vehicle that are not directly visible to the operator, because the engine hood and counterweight block the view from the operator's cabin at the front left of the vehicle. Therefore, it is necessary to make it possible to check the surroundings with mirrors and cameras. We developed a function to process the images of three monocular cameras mounted on the hydraulic excavator in real time. After processing the images, the function notifies the operator of any human-like images with a monitor display and sound. An image of the function is shown in Figure 1.

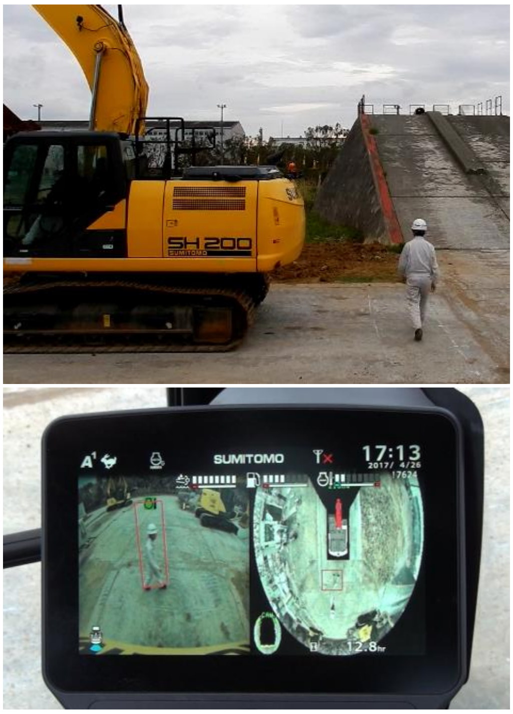

Figure 1. An example of the image recognition function.

The conventional function developed by us [1] [2] consists of three monocular cameras with wide angle, high sensitivity and wide dynamic range at the rear, right and left sides of the hydraulic excavator, which are combined in real time and displayed on a monitor in the operator's cabin. This function allows the operator to see 270 degrees view behind the excavator at a glance while sitting on the operator's seat. An example of the camera arrangement and a composite image is shown in Figure 2. The pipeline processing using FPGA allows for the composite display of images without frame rate degradation. In 2011, this function was first sold as an option for hydraulic excavators and later it becomes standard equipment.

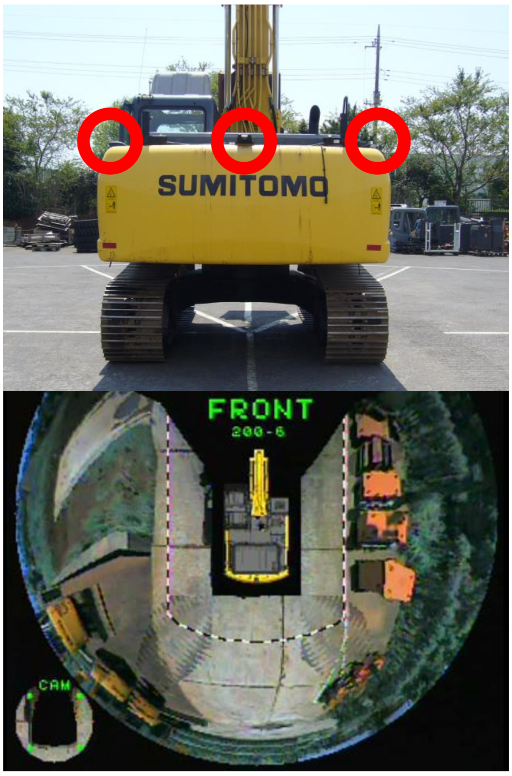

Figure 2. An example of a three-camera arrangement and image composition

However, the timing of when operators check the monitor is different depending on the operators. For example, it is assumed that workers in the vicinity may approach the hydraulic excavator while you are concentrating on your work and taking your eyes off the monitor. We developed a function to recognize camera images and notify the operator of the presence of human-like images with a monitor display and sound so that, the operator checks the area. The development's goal was to achieve a low-cost system configuration with only a monocular camera and to process the composite image that covered the entire 270-degree rearward area. In this paper, we introduce the technical issues and solutions to realize these problems, as well as examples of actual operations.

2 Targets and issues

The development goals were set as follows. This chapter describes the technical challenges for achieving the goals.

Target 1: Evaluation of the human presence using only monocular image recognition

Target 2: Processing of camera images with three different directions at least 10 times per second.

Target 3: Development compact and inexpensive in-vehicle equipment

2.1 Apparent change of a person in an image

The apparent appearance of the person in the image changes in a variety of ways, such as the color contrast of the background and the clothing, clothing with different colors for the upper and lower body, personal belongings, position and orientation to the camera, and pose. In addition, since the camera is mounted on the top of the hydraulic excavator body, which is taller than the person's height, the ratio of the image of the head to the body changes when the person approaches the camera and when he or she moves away from the camera. An example of this image is shown in Fig. 3. Although we assumed that only pedestrians appear in the images in this development, it is still difficult to manually design an algorithm to quantify human-like appearance by assuming all kinds of apparent changes of pedestrians.

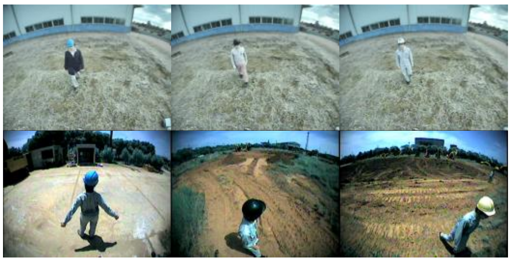

Figure 3. Examples of changes in the appearance of a person on an image

2.2 Distortion of the wide-angle camera image

A wide-angle camera is used to capture the boundary between the rear camera and the side camera without any gaps, and it is installed at an angle downward to cover the entire distance from the very near to the far side of the excavator. As a result, the image of an object may be tilted at the left and right ends of the field of view, while the image of the excavator may be reflected at the lower end of the field of view. An example of the image is shown in Figure 4. When we set the region of interest to evaluate the human presence if there is a human in the area, a part of the attention region might be outside the image or covered by the excavator. This means that image information is missing.

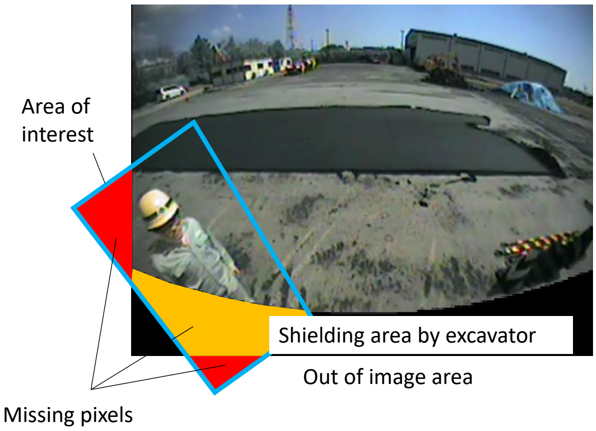

Figure 4. The missing pixels at the edge of the field of view

2.3 Faster and more stable image recognition processing cycle

It is necessary to keep executing image recognition processing, which requires a large number of computations, at a stable time cycle; FPGA can be used for stable processing time, but low-cost FPGA is limited by resources such as multipliers and internal memory.

2.4 Dense placement of the attention area

Even if the area of image for the human detection is limited to reduce the computations, thousands of regions of attention are still needed for each camera to evaluate human presence. But if we evaluate the human presence for all of regions, it is too computationally intensive to achieve the target processing cycle.

2.5 Reduction of substrate area

If the FPGA or CPU that implements image synthesis and image recognition processing is mounted separately, the board area becomes larger and more expensive, and the external appearance of the device becomes larger and more difficult to install in the excavators.

3 Evaluation of the human presence by Monocular Image Recognition

We adopts an evaluation model based on machine learning technique for the technical issues from 2.1 to 2.2 in this development. However, publicly available databases and evaluation image cropping methods [3] are not configured to address the issues described in 2.1 and 2.2, especially when a person comes directly under the camera.

In order to obtain the evaluation model for human presence in machine learning technique, a large number of supervised images are required. To collect the supervised images, we assumed a simulated environment where there is a hydraulic excavator equipping a laser sensor and a camera. The camera images were collected while the laser sensor measured the position of a person. Thousands of images were collected in the simulated environment, varying the background, weather, people's clothing, time of day, season. Then, based on the measurement results of the position of the person, a region of interest was set on the camera image, and the image was cropped and stored in the database. The position of the person in relation to the camera, the date and time of shooting, the background images and the weather, the region of interest outside the image range or the image of the car, and the original camera image were also stored in the database.

We divided the database into two categories. Category one is for the data taken very close distance from the camera, and category two is for the data taken away from the camera. We randomly selected samples from the data of the two categories. The supervised images for human detection were obtained by excluding pixels of outside the image range from the sample images or removing sample images with a large proportion of pixels of the excavator. This is to prevent the learning of pixels outside of the image range or the image of an excavator body as a characteristic of human. It is also possible to generate and use separate training models for the vicinity of the camera and for other areas to accommodate differences in the way people are reflected. The non-personal images are randomly selected by setting an arbitrary region of interest in the image without a person.

We tried to increase the number of combinations of clothing and backgrounds for supervised images, but there was a limitation. Therefore, as a preprocessing method for machine learning, we performed a feature vector transformation based on the distribution of the luminance gradient direction and normalized it for each 84-segmented partial feature vector to make it less susceptible to effects such as the contrast between background and clothing and clothing with different colors in the upper and lower body. Machine learning is ensemble learning in which the normalized sub-feature vectors are selected for a maximum of 99 logistic regressions.

4 Image recognition processing and overall structure

For the technical issues 1.3 to 1.4, we have adopted the overall structure shown in Figure 5.

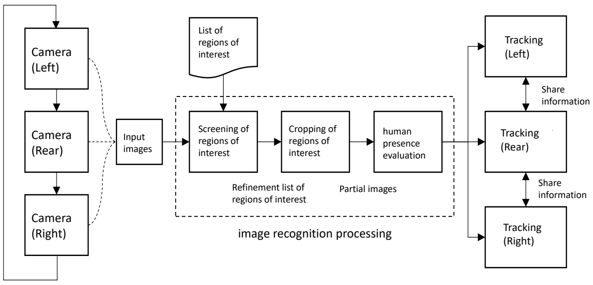
Camera (Left)
Camera (Rear)
Camera (Right)
Input images
List of regions of interest
Screening of regions of interest
Cropping of regions of interest
human presence evaluation
Refinement list of regions of interest
Partial images
image recognition processing
Tracking (Left)
Tracking (Rear)
Tracking (Right)
Share information
Share information

Figure 5. Overall structure of the image recognition process

The region of interest is the rectangle that surrounds the person in the image. The position of rectangles are calculated in advance for each position of the person in relation to the camera.

The image recognition process is the FPGA module that inputs grayscale images extracted from the camera images and outputs information on the area of interest that is judged to be human-like. It includes the processing of area of interest screening, the area of interest cropping, and the human presence evaluation.

In the screening of regions of interest process, all the region of interest in the field of view are input and the candidates are narrowed down by simple image recognition without trimming the area. In this case, we assume that there is a human-like image in the region of interest and transform the feature vectors of the region corresponding to the human head based on the luminance distribution. After that, we evaluate the image recognition with a machine learning model. The machine learning model is a lightweight random forest.

The region of interest extraction process extracts the pixels in the region of interest narrowed down by the region of interest screening process and normalizes them to a partial image with a uniform size.

The human presence evaluation process applies the human presence evaluation model described in Chapter 3 to the partial image generated by the region of interest cropping process, and judges whether the image is humanness or not by comparing it with the threshold value.

The tracking process obtains the location, time and camera ID of the region of interest that is determined to be humanness. The position of the region of interest is determined to be within the range of a human movement by comparing it with the results of the judgement made up to the previous time. This movable range is calculated from the relative speeds of the man and the excavator. If the judgment is positive, the positional information of the region of interest is output. This positional information is a rectangle that surrounds the object in the image and is fine-tuned to take into account the deformation of the edge of the field of view and the speed of movement. It can be done. In addition, as the person may move between cameras, a function to share location information between adjacent cameras is provided. The system is equipped with a function to select a person in the same field of view in order of near distance from the camera if more than one person is detected in the same field of view. When two or more persons are detected in the same field of view, a function to select them in the order of their distance from the camera is provided. This set of processes is a lightweight algorithm that applies a histogram filter [4] and does not affect the speed of overall image recognition cycle.

5 System Configuration and Hardware

In this development, the structure shown in Figure 6 was adopted to solve the technical problem of 2.5. The outline specifications of the configuration are shown in Table 1.

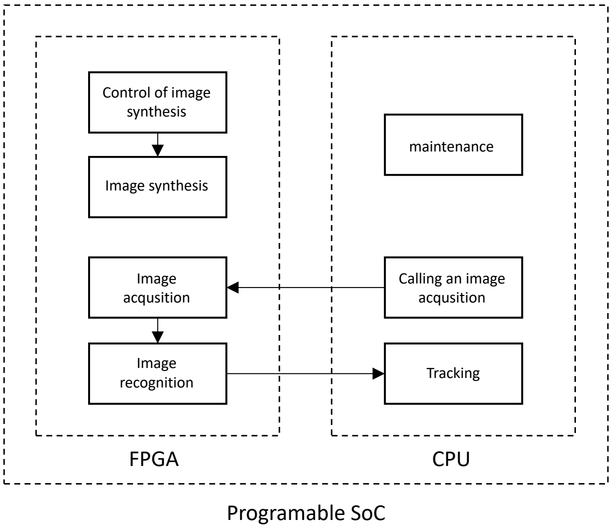

Figure 6. Overall configuration using a programmable SoC

Table 1. System specifications

In-vehicle unit	Specification	
Image Signal form	NTSC	
CPU	800	MHz
RAM	DDR3L SDRAM 2048	MB
Image synthesis memory	SDR-SDRAM 64	Mbit
Dimensions	Width 215	mm
	Length 128	mm
	Height 36	mm

The programmable SoC consists of the FPGA part, a CPU and a microcomputer part with various peripheral interfaces in the same chip.

In the FPGA part, an image synthesis module, a control module, an image recognition module, and an image acquisition module that supplies grayscale images to the module are integrated. In the microcomputer, a software that calls the image acquisition module, tracking software, and maintenance software for setting and adjusting each FPGA module and software were installed.

The evaluation of various scenes was based on the part of the database described in Chapter 3. Scenes in which people were present in the vicinity of 2 m from the camera and should be correctly detected were selected as the dataset for human detection evaluation. In addition, scenes that did not contain people were selected as a dataset for false alarm evaluation. While analyzing the processing results of these data sets, we adjusted the balance between the sensitivity to detect people and the false alarm rate. As shown in Figure 10, we were able to detect people with simple backgrounds with almost no misses. On the other hand, as shown in Figure 11, when the background contains high contrast and complex edges, the number of missed images and false alarms increases. Specifically, this includes the shadows of the excavated ground, plants and trees, other work equipment, cars, and man-made objects such as drums and benches.

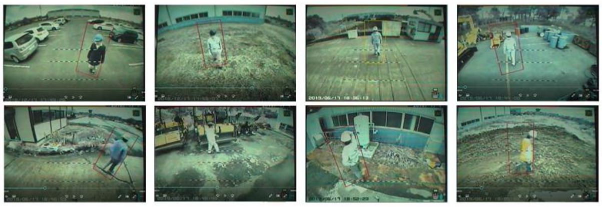

Figure 7. Dataset for human detection and evaluation

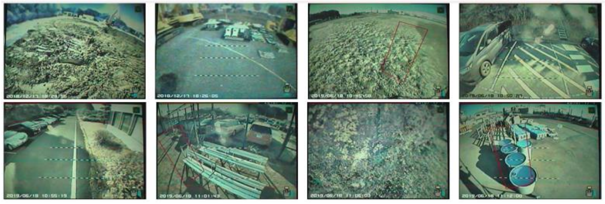

Figure 8. Dataset for false alarm evaluation

6 Examples of system operation

Figure 9 is an example of a monitor display of the recognition results for various orientations and distances of people from the camera, and for various types of clothing and backgrounds. If there is a human-like image, we can see that the image can be displayed on the monitor regardless of the change in appearance on the image.

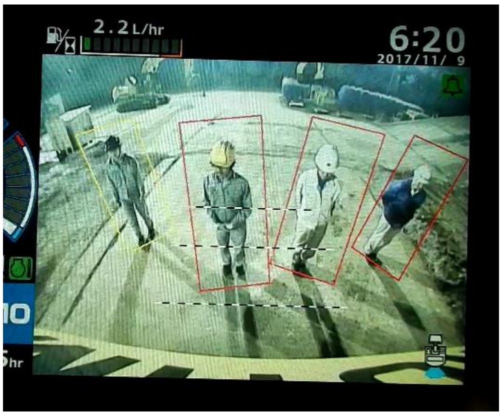

Figure 9. Perceptional results of people with diverse appearances

Figure 10 is an example of a monitor display of the recognition result when a person is approached by walking along the boundary between adjacent cameras. Even if the image appears to be tilted near the edge of the field of view, if there is a human-like image, it can be displayed on the monitor because of the image distortion of the wide-angle camera.

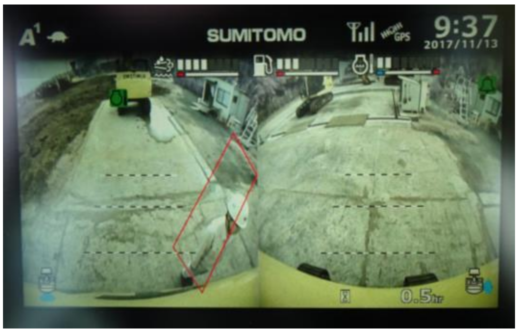

Figure 10. Recognition results at the edge of the field of view

Figure 11 is an example of the conventional function of displaying recognition results on a composite image. The system not only allows the operator to check the rear 270 degrees of the excavator at a glance, but also notifies the operator of any human-like images on the monitor display, and at the same time the system notifies the operator by sound, etc., which can be a trigger for the operator to check the area around the excavator.

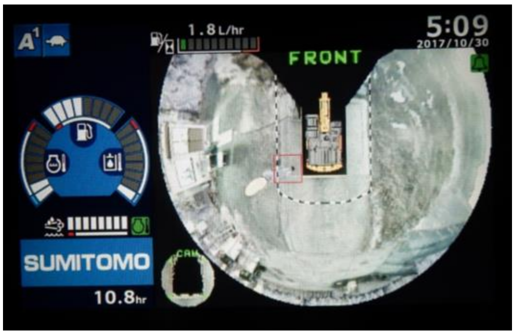

Figure 11. Display of recognition results on the composite image

Figure 12 is an example of the monitor display of the recognition result when a non-human object is judged as a human.

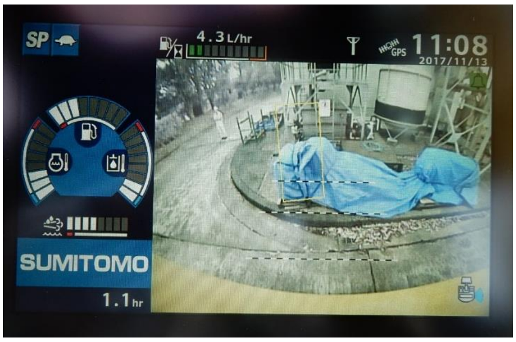

Figure 12. Examples of not being a person, but recognizing a person as a person

7 Conclusion

We have developed and put into practical use a function that recognizes images from a monocular camera mounted on a hydraulic excavator and notifies the operator of any human-like images with a monitor display and sound, giving the operator the opportunity to check the surroundings. The system was also approved for registration in the New Technology Information System (NETIS) [5].

References

[1] Yoshihisa Kiyota. and Masato Indo. and Hidehiko Kato. Development of a Field View Monitor System - A Support System for Confirming the Surroundings of Hydraulic Excavators by Video Synthesis. *SUMITOMO HEAVY INDUSTRIES TECHNICAL REVIEW*, No.179 Aug.2012, pages 5-8.

[2] New Technology Information System：NETIS，https://www.netis.mlit.go.jp/netis/pubsearch/details?regNo=KT-110057%20, 2018

[3] Yuan Li, Masaya Itoh, Masanori Miyoshi, Hironobu Fujiyoshi. Human Detection using Smart Window Transform and Edge-based Classifier. *The Japan Society for Precision Engineering Fall Conference*, Sep.2011, pages920-921.

[4] Sebastian Thrun, Wolfram Burgard, Dieter Fox, Probabilistic Robotics, The MIT Press. simulated environment

[5] New Technology Information System：NETIS，https://www.netis.mlit.go.jp/netis/pubsearch/details?regNo=KT-190106%20, 2020