シミュレータで作成された訓練データのデータ拡張による油圧ショベルの動作認識 *

ルイ笠原純ユネス ** 沈 鎭赫 † 小松 廉 † 筑紫彰太 ** 永谷圭司 †
千葉拓史 †† 山本新吾 †† 茶山和博 †† 山下 淳 ** 淺間 一 **

Action Recognition of Excavator with Data Augmentation of Simulator-Generated Training Data

Jun Younes LOUHI KASAHARA, Jinhyeok SIM, Ren KOMATSU, Shota CHIKUSHI, Keiji NAGATANI,
Takumi CHIBA, Shingo YAMAMOTO, Kazuhiro CHAYAMA, Atsushi YAMASHITA, and Hajime ASAMA

In construction sites, construction machinery such as excavators plays a critical role. The management of such equipment, notably the monitoring of actions conducted by each construction machinery, is, therefore, key to high productivity and efficiency. This time-consuming and laborious task is currently conducted manually by humans and thus, its automation is highly sought after. Previous works on this issue have achieved high performance using deep learning-based approaches and cameras. However, the investments needed to obtain the training data critical to such approaches are often prohibitive. Using a simulator to generate the training data appears therefore as an alternative to allow fast and easy gathering of training data. However, models trained using such training data perform poorly on real data. The purpose of this study is therefore to increase the performance of action recognition of construction machinery such as excavators using simulator-generated training data. A data augmentation process using background images gathered from actual construction sites is used to reduce the gap between simulator-generated data and real-world data. Experiments with data collected in an actual construction site showed the effectiveness of the proposed method.

Key words: computer vision, automation in construction, action recognition, data augmentation, sim2real

1. Introduction

In the construction industry, heavy machinery represents a large portion of a project's budget. Those are usually machines such as excavators, dump trucks, bulldozers, ... Despite the recent advances that pushed forward the automation of many processes, the construction industry is known to still have lower efficiency compared to other fields such as the manufacturing industry [1]. One key aspect to improve efficiency is the monitoring of the aforementioned construction machinery. However, this is still a task conducted by human workers, either on-site during the actual project or afterward using video recordings. Therefore, the automation of construction machinery monitoring, i.e., action recognition of construction machinery, is highly desirable.

Previous works that focused on the action recognition of construction machinery can be distinguished between those using onboard sensors and those using outboard sensors. Commonly used onboard sensors are encoders and GPS [2] and have the merit of providing reliable and low dimensional data for action recognition. However, the use of such sensors involves modifications on each of the monitored construction machinery. This can be prohibitive for scenarios involving a large fleet of machines or rented machines. Commonly used outboard sensors are cameras and microphones [3)1)]. Those have the merit of being usually already installed on construction sites for other purposes such as surveillance and archiving. Furthermore, a single sensor can potentially allow the monitoring of several machines simultaneously. However, the data dimensional is higher and those sensors are less reliable: cameras are affected by lighting conditions and microphones are affected by wind noise.

Researches on action recognition using cameras have seen a great surge in recent years, due to two factors. First is the boom of cameras, resulting in higher availability at lower costs and higher resolutions. Second is the progress in machine learning, namely the advent of deep learning, which allowed the processing of camera information at higher levels. While human action recognition is predominant [4)5)], some deep learning-based approaches for action recognition of construction machinery using cameras were also proposed [6)-8)]. One unsolved issue for such approaches is the gathering of training data. Indeed, while deep learning approaches can boast high performance, they are critically dependent on the available training data, i.e., labeled data. Compared to human action recognition, for which more labeled data is made available on almost a daily basis, labeled data on specific targets such as construction machinery hardly exists. Additionally, since construction machinery usually operates in construction sites where access is restricted, the gathering of large training data required for deep learning approaches is also very difficult from a logistical perspective.

In our previous works [9)10)], we have attempted to tackle this is-

* 原稿受付 令和 3 年 5 月 28 日
掲載決定 令和 3 年 10 月 18 日
** 正会員 東京大学 (〒113-8656 東京都文京区本郷 7-3-1)
† 東京大学
†† 株式会社フジタ (〒151-8570 東京都渋谷区千駄ヶ谷 4-25-2)

sue by a sim2real approach consisting of using training data generated by a simulator instead of training data gathered at an actual construction site. Promising results were obtained. However, the achieved performance was lacking, mainly due to the domain shift between simulator-generated data and actual data, i.e., differences in image features between the simulator and the real world.

Therefore, the objective of this paper is to achieve higher performance in action recognition of construction machinery without using real-world training data. This is achieved by data augmentation using background images obtained in actual construction sites of simulator-generated training data.

2. Concept and Overview

Video data generated using a simulator and data gathered in the real world differ in many aspects. Therefore, a learning model matching class labels with input data would learn different relationships and ultimately perform poorly if trained exclusively using a simulator and tested on real-world data. This is an issue globally known as *domain shift*, which refers to the mismatch between the training and testing data. This is caused by several inconsistencies between the simulator environment and the real world: the appearances being affected by varying lighting, background clutter, etc.

Video samples can be roughly divided into foreground objects and background. Among all the causes for the domain shift, in this study, we focused on the background portion of the video samples. In the proposed method, we argue that the domain shift can be filled in large part by using actual background images of the construction site. While obtaining actual training data from the construction site where deployment is expected, i.e., labeled footage of the excavator operating in the real world, has high costs due to the need to label the accumulated data, obtaining background images is easy since there is no need for tedious labeling. Therefore, in this study images obtained in actual construction sites where the system is expected to be deployed are used to bridge the differences between simulator-generated training data and real-world data. Concretely, data augmentation is conducted by substituting the background in simulator-generated training data with images of actual construction sites.

3. Method

3.1 Action Recognition of Excavator

Excavators are one of the most common construction machinery. They are also critical in earth moving work, consisting of displacing soil, which is fundamental in construction [11]. In this process, as illustrated in **Fig. 1**, the excavator digs the ground and loads the extracted soil onto a dump truck [12]. The workflow can be distinguished into the following steps:

1. Dig soil
2. Rotate to face the dump truck
3. Load soil onto the dump truck
4. Rotate back to face the digging location

Steps 2 and 4 essentially consist of the same excavator action, they only differ by whether the bucket is full or empty. The earth moving work can therefore be decomposed into three excavator action classes: ”digging”, ”turning” and ”loading”. Those are illustrated in **Fig. 2**.

Fig.1 An excavator conducting earth moving work

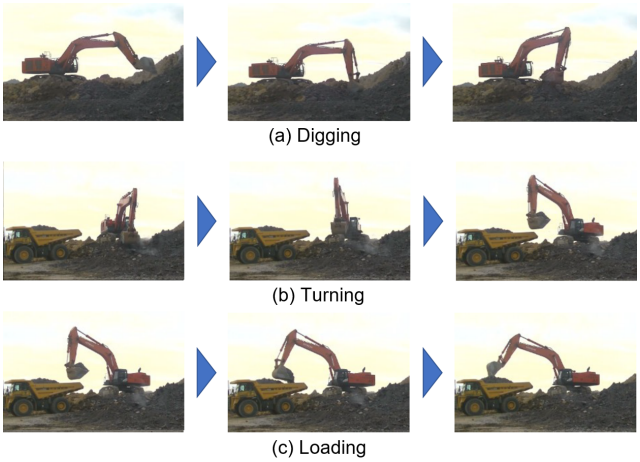


Fig.2 The three action classes considered in this study

3.2 Training Data Generation Using a Simulator

While gathering training data in actual construction sites requires heavy investments in logistics and manpower, generating training data for action recognition of excavator in a simulator is comparatively easy. In this study, Vortex Studio was used as a simulator to generate the needed training data. Vortex Studio is a real-time simulator for operating mechanical systems and allows the operation of construction machinery such as excavators among others [13].

Since a simulator is used here, there is virtually no limit in the number and diversity of camera viewpoints that can be realized. Most scenarios for automated action recognition of construction machinery involve the use of fixed cameras in the construction site. The locations where those cameras can be placed are limited in practice and can be reasonably expected to be decided in advance, i.e., in the planning phase of the construction project. Therefore, in our experiments, we generated training data with viewpoints very roughly matching the viewpoints of the data collected in the actual construction site.

In this study, a model of an excavator was controlled in Vortex Studio manually to continuously perform excavation, turning, and loading, as illustrated in **Fig. 3**. The simulator was set up in what could be described as the most neutral conditions with a decent light source illuminating the excavator model from the top. For the action class ”turning”, video segments with the bucket full

Fig.3 Excavator in the simulator environment of Vortex Studio

and empty were generated. Recordings were also taken from several viewpoints. Those recordings were then segmented to separate each action and labeled accordingly.

3.3 Data Augmentation

In our previous work [9], a video filter was used to attempt to bridge the simulated-to-real gap between the training data generated using a simulator and the test data gathered in actual construction sites. However, the accuracy of action recognition was low due to the influence of the background present in real construction sites. This issue could not be solved using our previous approach. Therefore, it is necessary to add various backgrounds to the training data generated in the simulator [14]. However, the addition of backgrounds that are not related to the test data will have the opposite effect of further lowering the action recognition performance [15]. Therefore, in this study, data augmentation of the simulator-generated training data is conducted using background images of the actual construction site.

Several images were taken from a construction site in Motomiya City, Fukushima Prefecture, Japan, for data augmentation. Those are shown in **Fig. 4**.

Those images correspond to the backgrounds where excavators of the test dataset operate in. In the scenario considered in the present study, the monitoring of construction equipment is conducted using cameras. The locations where cameras can be positioned in a construction site are limited so as not to impede construction activities and safety. Those can therefore be known in advance: for example, the location and orientation of fixed cameras for surveillance purposes are usually decided at the planning stage of a construction project. Those images were selected in order to reduce the domain shift between the simulator-generated training data and the test data obtained in a real construction site.

The simulator offers the advantage of allowing rapid and easy generation of excavator motion along with the corresponding activity labels. However, the simulator differs from the real world in many aspects. This causes the domain shift and impedes the effectiveness of the trained model. Those images were selected in order to reduce the domain shift by matching the background in both data.

It is worth noting that, unlike gathering labeled video segments, gathering those background images is relatively easy. Using those background images and a video editing software, the background in the training data generated in the simulator was replaced with background images from the construction site, as shown in **Fig. 5**. By including those background images in order to match the background images between the training and test data, the domain shift can be largely reduced.

3.4 Learning Model

In this study was used a combination of Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) [16], which showed initially high recognition accuracy on human action recognition but also on excavator action recognition [10]. CNN is a network in which the intermediate layer is composed of a convolutional layer and a pooling layer, and a feature map containing spatial information can be extracted. LSTM is a network able to learn long-term dependencies and is therefore suitable for time-series data. The use of both CNN and LSTM allows recognition of the excavator's actions considering both spatial and temporal information.

This architecture can be thought of potentially suffering from the gap in camera perspective on the excavator itself between the training and testing data. However, as mentioned earlier, it can be reasonably be expected that the camera perspective of the test data would be at least roughly known in advance. In the cases where this assumption does not hold, since our approach has the advantage of using a simulator to generate the training data, i.e., can generate training data of virtually any camera perspective, the number of camera perspectives in the training data could be expanded at very little cost to bridge the gap.

In **Fig. 6** is shown an outline of the learning model. The training data generated using a simulator is used as input data. Training data consists exclusively of video data. First, RGB image data is gathered by extraction from each video frame contained in the training data sample. Then, the extracted RGB data is inputted to the CNN, and image features are extracted. Those are inputted into the LSTM. The LSTM consisted of three layers and the last softmax layer classified each sample between each excavator action class.

4. Experiments

In our experiments, all video segments were taken at a resolution of 1920*1080 and at a framerate of 30 fps. The duration of each video segment was not unified: the average duration was 7s, with the shortest being 4s and the longest being 13s.

During learning, RGB data is first extracted from each frame of a video segment of the training data. The RGB data is then converted to a size 298*298*3. Following that, it is inputted into a CNN for feature extraction. The used CNN was Inception V3 [17], a network pre-trained on more than 1 million images. The extracted features are then finally passed to the LSTM. Training of both the CNN and the LSTM were done using the Adam optimizer with a batch size setting of 32 and for 150 epochs.

Three methods were considered in our experiments:

1. CNN+LSTM, serving here as a baseline, differing with the proposed by only the absence of the data augmentation step.
2. The method of [10].
3. The proposed method.

Fig.4 Background images collected in actual construction sites

Fig.5 Data augmentation using background images from actual construction sites

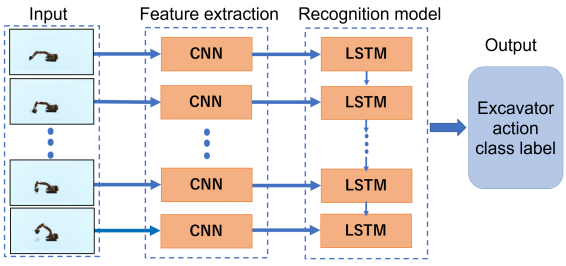


Fig.6 Outline of the learning model

(a) Test data collected in an actual construction site

(b) Data collection process

Fig.7 Test data gathered in actual construction site

4.1 Data Gathering in Actual Construction Site

To measure the performance of our system, test data was collected in an actual construction site located in Motomiya City, Fukushima Prefecture, Japan, where a standard excavator was conducting earth moving work, as shown in **Fig. 7**. Due to the activity of the construction site, resulting in various heavy machinery operating in the close vicinity, the number of available recording locations was limited. Nonetheless, recording of the excavator conducting earth moving work was successfully conducted from three camera viewpoints. Segmentation and labeling of the obtained data into the three action classes presented earlier were conducted manually. The resulting data contained 90 video segments, comprised of 30 video segments per action class.

4.2 Data Generation in Simulator

As mentioned previously, training data was generated using Vortex Studio, a real-time simulator. The camera viewpoints in the simulator were set up to roughly match the ones that were available during the test data gathering in the actual construction site, as illustrated in **Fig. 8**. 180 labeled video segments were generated, consisting of 60 video segments per action class. Additionally, a smaller dataset of 60 video segments, comprised of 20 video segments per action class, was separately generated for testing purposes.

For each training video segment obtained using a simulator, data augmentation was conducted with 3 randomly selected background images from an actual construction site. Therefore, following data augmentation, the training data generated using a simulator contained 540 video segments, comprised of 180 video segments per action class.

5. Results and Discussions

To measure the performance of the action recognition system, the classification accuracy was computed as in (1).

$$\text{accuracy} = \frac{\text{Number of correctly classified video segments}}{\text{Total number of video segments}} * 100 \quad (1)$$

In **Fig. 9** are reported the accuracy values for the considered methods for both the test data generated in the simulator and the test data gathered in the actual construction site.

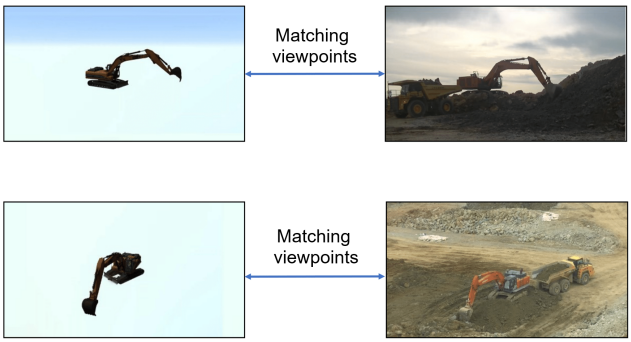


Fig.8 Viewpoints in simulator and in actual construction site

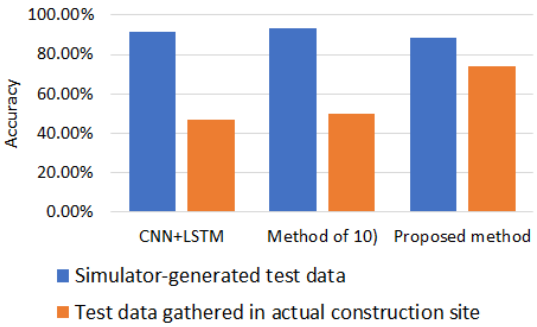
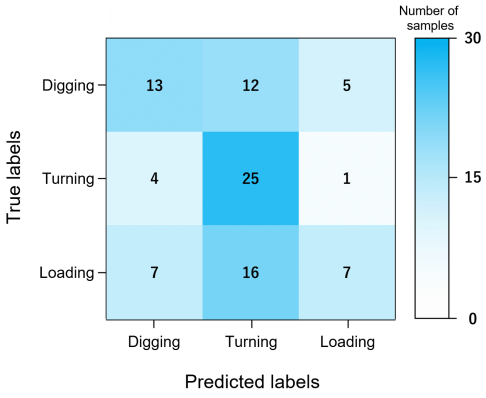
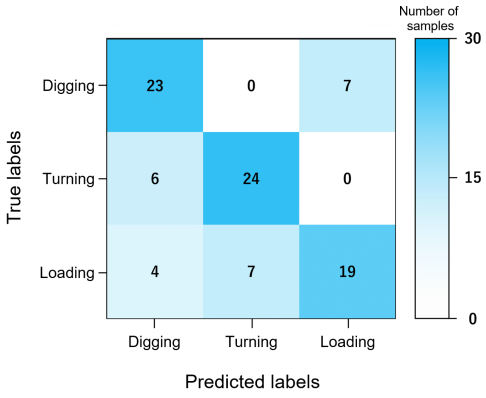


Fig.9 Accuracy comparison



(a) Confusion matrix of the method of [10)]



(b) Confusion matrix of the proposed method

Fig.10 Confusion matrices of the considered methods in experiments with test data gathered in actual construction site

CNN+LSTM as well as the method of [10)] achieved high accuracy of over 90% on the test data generated in the simulator. This shows that the learning process itself was successful: when training and testing are conducted with simulator-generated data, although there is no overlap between them since they were both generated in the same simulator environment, both methods showed a solid performance. However, when tested on the data collected in the actual construction site, both models see their performance drop considerably, down to about 50%. This is due to the domain gap between the simulator and the actual construction site: the models here are not able to learn effective features for action recognition in actual construction sites from simulator-generated data. The proposed method shows a slightly lower performance than the other two methods when tested on simulator-generated data. This is certainly due to the introduction of background images from actual construction sites, which are irrelevant for the test data generated in the simulator and have the opposite effect of hindering performance. This is in line with the findings of [15)]. However, the benefits of our proposed method are visible when tested on the data collected in actual construction sites: while there is still some performance drop, the model still managed to perform at over 70% accuracy.

The study published in [18)], also using a camera-based Deep Learning approach but with training conducted with real-world training data, reported performances of about 75% accuracy. This is similar to the performance of our proposed method, which has the merit of only using easy-to-obtain simulator training data. Human action recognition is a more flourished field and several approaches with varying levels of performance have been proposed over the recent years [19)]. Among them approaches such as [20)] used simulator-generated training data to augment the real-world training data and reported performances up to 81% accuracy.

In **Fig. 10** are reported the confusion matrices obtained by the method of [10)] and the proposed method in the same conditions as those in Fig. 9. Looking at the diagonal of the confusion matrix of the method of [10)] shown in Fig. 10(a), it can be seen that only the class ”turning” has been somewhat correctly recognized: about 80% of ”turning” samples were correctly classified. The performance is much more lacking for the other two classes of ”digging” and ”loading”, with correct classification rates of about 40% and 20%, respectively. The proposed method, for which the confusion matrix is shown in Fig. 10(b), shows a much better performance equally across action classes: looking at the diagonal of the confusion matrix, correct classification rates remain between 60 to 80%. While keeping the classification performance of the ”turning” class at the same level, our proposed method managed to significantly increase performance for the two other classes. The ”digging” and ”loading” classes are characterized mainly by the movements of the boom and bucket, as seen in Fig. 2. Those are more subtle movements compared to ”turning” that involves rotation of the body. We surmise that the inclusion of real background images allowed the proposed method to grasp such subtle movements for action recognition, which was not possible for the method of [10)].

It is worth noting that in our study only a single model of excavator was considered. Furthermore, this model was a generic one,

only roughly matching the excavator contained in our test data collected in an actual construction site. Since except for some extreme cases, excavators across different manufacturers roughly possess the same configuration, we surmise that our results are still valid even with other extractors from different manufacturers.

In rare cases where an excavator from a manufacturer exhibits particular characteristics, i.e., has large disparities with the excavator model used in the simulations generating the training data, the simple counter-measure of changing the model in the aforementioned simulator could be easily taken.

Additionally, in our study, the excavator model in the simulator was operated by a single individual. In the field, the actual excavator was operated by a different individual. Therefore, the issues related to individual differences of operators were not investigated. According to [21], the differences in excavator movements across operator skill levels are apparent in small movements parameters, namely the speed of movement, the range of motion, and the angle of bucket. Those small contributions from each individual small differences through several work cycles ultimately result in visible differences between excavator operation by different individuals. Since our proposed system is based on short videos segments, the influence of individual differences between operators is thought of having only limited effects.

6. Conclusion

In this paper was proposed a method to augment simulator-generated training data using background images collected in actual construction sites to increase action recognition performance. By replacing the background in video segments generated in a simulator with real-world background images, the model showed a significantly better ability to learn the features characteristic of each of the considered excavator action classes without the use of real-world training data.

In the future, we would like to pursue efforts to further increase the action recognition performance of our system, including focusing on the issue of viewpoint changes. In our experiments, training data was generated to roughly match the viewpoints in actual constructions sites and therefore, if the viewpoint information is not available at the training data generation phase, lower performance can be expected since camera-based approaches are known to suffer from changes in viewpoints [22].

References

1) C. Cheng, A. Rashidi, M.A. Davenport and D.V. Anderson: Activity analysis of construction equipment using audio signals and support vector machines, Automation in Construction, **81**, (2017) 240.
2) N. Pradhananga and J. Teizer: Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Automation in Construction, **29**, (2013) 107.
3) R. Akhavian and A.H. Behzadan: Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers, Advanced Engineering Informatics, **24**, 4, (2015) 867.
4) Q.V. Le, W.Y. Zou, S.Y. Yeung and A.Y. Ng: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2011) 3361.
5) S. Herath, M. Harandi and F. Porikli: Going deeper into action recognition: A survey, Image and vision computing, **60**, (2017) 4.
6) J. Kim and S. Chi: Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles, Automation in Construction, **104**, (2019) 255.
7) C. Chen, Z. Zhu and A. Hammad: Automated excavators activity recognition and productivity analysis from construction site surveillance videos, Automation in Construction, **110**, (2020).
8) H. Nakamura, Y. Tsukada, T. Tamaki, B. Raytchev and K. Kaneda: Pose estimation of excavators, International Workshop on Advanced Imaging Technology, (2020).
9) J. Sim, J.Y. Louhi Kasahara, S. Chikushi, H. Yamakawa, Y. Tamura, K. Nagatani, T. Chiba, S. Yamamoto, K. Chayama, A. Yamashita and H. Asama: Action Recognition of Construction Machineryfrom Simulated Training Data Using Video Filters, Proceedings of the International Symposium on Automation and Robotics in Construction, 2020.
10) J. Sim, J.Y. Louhi Kasahara, S. Chikushi, H. Yamakawa, Y. Tamura, K. Nagatani, T. Chiba, S. Yamamoto, K. Chayama, A. Yamashita and H. Asama: Effects of Video Filters for Learning an Action Recognition Model for Construction Machinery from Simulated Training Data, Proceedings of the IEEE/SICE International Symposium on System Integration, (2021) 12.
11) S. Han, S. Lee, T. Hong and H. Chang: Simulation analysis of productivity variation by global positioning system (GPS) implementation in earthmoving operations, Canadian Journal of Civil Engineering, **33**, 9, (2006) 1105.
12) M.M. Soltani, Z. Zhu and A. Hammad: Skeleton estimation of excavator by detecting its parts, Automation in Construction, **82**, (2017) 1.
13) Vortex Studio, `https://www.cm-labs.com/vortex-studio/` accessed 2020.06.27.
14) A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers and T. Brox: Flownet: Learning optical flow with convolutional networks, Proceedings of the IEEE International Conference on Computer Vision, (2015) 2758.
15) N. Dvornik, J. Mairal and C. Schmid: On the importance of visual context for data augmentation in scene understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2019).
16) J. Donahue, H.L. Anne, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko and T. Darrell: Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2015) 2625.
17) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna: Rethinking the inception architecture for computer vision, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016) 2818.
18) J. Zhang, L. Zi, Y. Hou, M. Wang, W. Jiang and D. Deng: A Deep Learning-Based Approach to Enable Action Recognition for Construction Equipment, Advances in Civil Engineering, (2020).
19) H. Zhang, Y. Zhang, B. Zhong, Q. Lei, L. Yang, J. Du, and D. Chen: A comprehensive survey of vision-based human action recognition methods, Sensors, **19**, 5, (2019).
20) H. Hwang, C. Jang, G. Park, J. Cho and I. Kim: ElderSim: A Synthetic Data Generation Platform for Human Action Recognition in Eldercare Applications, arXiv, (2020).
21) Y. Sakaida, D. Chugo, H. Yamamoto and H. Asama: The analysis of excavator operation by skillful operator-extraction of common skills, Proceedings of the SICE Annual Conference, (2008) 538.
22) D. Weinland, M. Özuysal and P. Fua: Making action recognition robust to occlusions and viewpoint changes, Proceedings of the European Conference on Computer Vision, (2010) 635.