Single Camera Worker Detection, Tracking and Action Recognition in Construction Site

Hiroaki Ishioka[a], Xinshuo Weng[b], Yunze Man[b], and Kris Kitani[b]
[a] Institute of Technology, Shimizu Corporation, Japan
[b] Robotics Institute, Carnegie Mellon University, United States
E-mail: h.ishioka@shimz.co.jp, xinshuow@cs.cmu.edu, yman@andrew.cmu.edu, kkitani@cs.cmu.edu

Abstract –

In Japan, the construction industry strongly be needed productivity improvement and increasing the number of new hires due to improvement of working environment. Site manager needs to grasp whether the daily progress is as planned and updates the schedule appropriately for improve site's productivity and safety. In image-based data acquisition approach in japan, there is a problem that learning is insufficient with only global public data, since construction worker in Japan has originality in image feature compare with other countries. In this study, we make original dataset for additional learning firstly. Then we proposed domain-specific algorithms specific to the Japan construction site, including a worker detection and tracking algorithm and a worker action recognition algorithm. As a result, our worker detection showed 87.9% accuracy in same-site evaluation and 77.5% accuracy in cross-site evaluation. Our worker action recognition showed 60.2% mean accuracy. Finally, the method of translation into activity element based on the output value of worker detection was indicated.

Keywords –

Construction Process; Worker Detection and Tracking; Action Recognition

1 Introduction

In construction management study area, many researchers try to improve productivity and safety by acquiring various data from construction work in construction site. Tarak et al [1] reviewed various paper and classified these research technologies into 3 types; Enhanced IT technologies, Geospatial technologies, and Imaging technologies. Amin et al [2] also reviewed various papers focusing on BIM and Computer Vision (CV) and describe the development.

In Japan, decreasing the number of construction workers is pointed out based on annual statistical data changes [3]. Therefore, the construction industry strongly be needed productivity improvement and increasing the number of new hires due to improvement of working environment. In construction management in Japan, site manager who belongs to general contractor company is assigned to the construction site to manage the progress of construction. Site manager needs to grasp whether the daily progress is as planned and updates the schedule appropriately. In the case of typical Japanese construction sites, site manager saves the construction time by subdividing the work space so that multiple contractors can perform their work on the same day. In order to appropriately update the schedule, it is necessary to understand the daily construction process. In data capturing, site manager wants to reduce management effort by reducing the number of capturing devices. It is better to use a camera that can sense the area with one unit than wearable sensors that be required the same number of workers.

In image-based approach in japan, there is a problem that learning is insufficient with only global public data, since construction worker in Japan has originality in image feature compare with other countries. From the safety awareness of Japanese construction workers, it is common in the construction industry to wear long-sleeve workwear that has both resistance and protectiveness. It also has color-variation. Due to the protective properties of workwear, safety vests in vivid color common on construction sites around the world are not usually worn on construction sites in Japan. The rules for wearing hard hat and safety belts have been generalized, and there are various product variations for tool bags attached to safety belts. In order to build image-based data acquisition technology for images with these unique characteristics of Japan, it is necessary to build a unique Japanese dataset.

In this study, we make original dataset for additional learning firstly. Then we proposed domain-specific algorithms specific to the Japan construction site, including a worker detection and tracking algorithm and a worker action recognition algorithm. Additionally, the method of translation into activity element which has effect to grasp whether the daily progress is as planned and updates the schedule appropriately based on the output value from the algorithms.

2 Literature review

2.1 Dataset in construction site

In study about image-based data acquisition technologies, there are many example of dataset for each case study. Jun et al [4] made 11 actions' movie clips for action recognition. Kaijian and Mani [5] described comparison of efficiency on various annotation task condition for out sourcing. Mohammad et al [6] made dataset of Excavator, Loader, and Truck for comparison of accuracy using various CV technologies. Recently, Mingzhu et al [7] made dataset of construction equipment and workers. Each of them, the images included dataset made from each country's construction site. These are not Japan. Still, Japanese construction site datasets need to be uniquely constructed.

2.2 Object Detection and Tracking

Object detection. There has been tremendous advancement in object detection in the last decades. Before the era of the deep learning, methods with hand-crafted feature descriptors for detecting specific types of objects were dominating in the literature. For example, various feature descriptors such as HoG (Histogram of Gradient) [10], SIFT (Scale-Invariant Feature Transform) [11] and DPM (Deformable Part Model) [12] have been proposed, which were customized for detecting pedestrians. While these methods made significant improvement in object detection, manual engineering of these feature descriptors requires significant efforts from the researchers for each individual type of object, making it difficult to generalize to other object types.

With deep learning technique becomes favorable in recent years, many modern data-driven object detectors have been proposed which can learn universal feature descriptor for many object types jointly from the data, removing the need of feature engineering. Among different approaches, region proposal-based object detectors [13, 14, 15, 16, 17, 18] are most popular. These approaches first define a set of anchors with different scales and size and covering the entire image, and then a box regression and classification network is applied to classify object class and meanwhile refine the box position and size. In this work, we choose one popular universal object detector, Faster-RCNN [15], and adapt it to specifically work on the construction workers in the Japan construction site, in order to achieve the best possible performance.

Multi-Object Tracking. Beyond classifying the object class and detecting the object location in the image, multi-object tracking aims to associate the detected objects in video and output movement of objects. To that end, recent multi-object tracking methods often employ a tracking-by-detection pipeline. Specifically, given detected objects in all frames, the tracker assigns the identity to each object where the same object receives the same identity. For example, SORT [19] proposed to use the motion information and used the Kalman filter with constant velocity as the motion model in the tracking-by-detection pipeline. To leverage the appearance cue in addition to the motion cue, deep learning-based methods such as Deep-SORT [20] and MOTS [21] incorporated an object re-identification branch to learn the object appearance feature for discriminative feature learning. While these data-driven methods with deep learning technique often achieves better performance than Kalman filter-based methods when the appearance information is sufficient, data-driven methods lack the ability to run in the real-time speed. In this work, we employ the Kalman-filter based tracking methods due to their favorable speed and the lack of appearance differences in construction workers who often wear similar uniforms and helmets.

2.3 Action recognition

Different from learning the location, orientation and trajectory of the object, action recognition aims to learn the internal logic of the object motions and its interaction with the environment. With the development of deep convolution networks in the computer vision area, a large number of CNN based methods have significantly improve the performance of traditional action recognition methods [25, 26]. We broadly categorize video action recognition methods into two genres: 2D and 3D CNN approaches. Methods of the first type make use of the recent advances in 2D single image CNNs by applying a CNN to each individual video frame and aggregating the prediction along the time axis [27]. In order to further consider the temporal dynamics of a motion rather than treat them individually, two-stream method [28, 29] are proposed to model appearance and dynamics separately and allow their interaction by early or late fusion. Among these methods, Simonyan et al. [30] first proposed the two-stream ConvNet architecture by introducing the temporal stream which takes optical flow frames as input. Wang et al. [31] proposed Temporal Segment Networks – a sparse temporal sampling strategy for the two-stream structure and fused the two streams by a weighted average. Other methods have taken different approaches to help better incorporate temporal information into single-frame feature extraction backbones such as CRF and LSTM [32, 33].

The other genre seeks to learn spatiotemporal features from videos directly with 3D CNN [34, 35, 36, 37]. Among them, C3D [34] is the first to leverage 3D

kernel on video data to learn spatiotemporal features, making it able to capture long range temporal information. Following C3D, Joao Carreira et al. [35] proposed i3D, in which they inflated the ImageNet pre-trained 2D kernel into 3D and take advantage of optical flow information with a two-stream architecture. They also proposed a new large-scale action recognition dataset named Kinetics and achieved competitive performance in other benchmark datasets. STCNet [36] inserted its STC block into 3D ResNet to capture the channel-wise correlation of both spatial and temporal features. Slowfast [37] proposed a slow-path to capture spatial semantics and a fast path the capture motion at fine temporal resolution.

Other than CNN, there is also method that model action recognition as a graph-based problem and use graph neural network to solve this problem [38]. In this work, we leverage two-stream 3D-ConvNet as our backbone [35] due to its stability and its strong feature extraction ability in complex scenario. We adapt the method to work in the Japan construction site scenario which contains high occlusion and complex actions and achieve the best possible performance.

2.4 Data usage

Output data from various data acquisition technologies reshape and use for each purpose. For example, Eirini and Ioannis [8] got the time when the worker stopping at workspace as production time from 4D trajectory data. Ye et al [9] showed time-space heatmap from workers' location data from BLE beacon for grasp usability of workspace. The former lack the method to grasp works on moving time, and the latter has difficulty to understand the meaning of location.

3 Method

3.1 Dataset creation

Dataset creation will be done by process below.

1. Selection of sites
2. Negotiation about permission
3. Capture movies and pickup movies
4. Annotate worker location and parameter

In annotation, we define three data; bounding box for worker detection, workers' ID for worker tracking, and tag of action categories for worker action recognition, into each frame. Action categories should be defined on common words for easily understand by annotator who not familiar with construction work and image of inside of construction site. In this study, we defined action as 6 actions; Walk, Crouch, Stand-up, Carry, Place, and Pick-up. Table 1 shows work definition for annotator and Figure 1 shows additional category definition about stable action Standing and Crouching which were needed because "other" tag was too large amount.

Table 1. Work definition for annotator

[Human Bounding Boxes]
●Description: A minimum enclosing box for a person
●Start: When all parts above shoulder (include face/head) is visible or when more than half of the person's body is visible
●End: When the above conditions no longer hold
●Notes:
○Must cover all the visible parts of the person, occluded parts do not need to be covered in the bounding box
○If the person is interacting with an object, all the contact points between the person and the object must be included in the bounding box, but it is not necessary to cover the whole object
○The bounding box should be as tight as possible. Small misalignment due to Imperfect interpolation is allowed, but it should be reasonable and calibrated at least every 10 frames.
[Actions]
Notes for all actions:
·There are 6 actions. All of them are single human actions. Walk, Crouch, Stand-up are atomic actions that do not have any object interactions. Pick-up, Place, Carry are interactive actions that depend on object interactions.
·One person can belong to one of the above 6 actions or none of them at a time.
·The 6 actions are not complete. A person can be doing "other" action, which does not need to be labeled.
·Only need to label when visually satisfies the definitions. For example, a man might start crouching when he is 90% occluded. We don't need to label this action. Another example, a man might pick-up something while the contact points between his hands and the object are totally occluded, we don't need to label this action.
[Action Definitions:]
●Walk
○Description: A person walking without carrying any object.
○Start: When the person starts moving and his first foot leaves the ground
○End: When the person stops walking.
○Note: Walking must be obvious and at least two steps. If a person is just changing the position by a little bit or moving just one step, that is not walking
●Crouch
○Description: A person starts to bend, sit, squat, crouch, knee down, etc. from a non-crouching status.
○Start: When the person's knees, waist, or back starts to bend
○End: When the person's knees, waist, or back stops moving or is fully bent
○Note: The actions must be obvious, for example a person looking down and bending only a little bit doesn't belong to crouch.
●Pick-up
○Description: A person makes contact, lifts, and starts to support an object against gravity.
○Start: When the person starts to lift the object up and begins supporting its gravity.
○End: When the person finishes lifting the object and reaches a stable gesture.
●Carry
○Description: A person starts walking/moving with a wide, long, large, or heavy object.
○Start: When the person starts walking/moving with the object
○End: When the person stops moving, or stops supporting the object's gravity
○Note: The object should be wide, long, large, or heavy. Carrying a small bag or backpack doesn't belong to carry.
●Place
○Description: A person lose contact, putting down, and stops supporting an object against gravity.
○Start: When the person starts lowering the object or starts to put it down
○End: When the person lose contact with the object, no longer supports its gravity, or body stops moving.
●Stand-up
○Description: A person stands up from a crouching or other non-standing status.
○Start: When the person's knees, waist, or back starts to recover from bent status
○End: When the person's knees, waist, or back is fully recovered from bent status

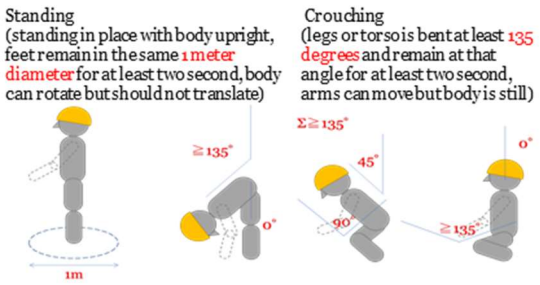


Figure 1. Additional category definition

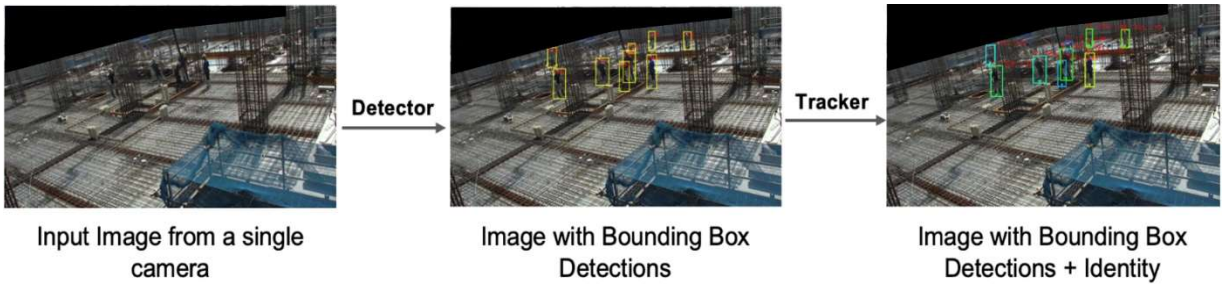


Figure 2. Illustration of our worker detection and tracking pipeline

3.2 Worker detection and tracking

We show the pipeline of our worker detection and tracking algorithm in Figure 2 Given a single image from video, we first run our worker detector to obtain a list of bounding boxes representing detected workers. Every worker bounding box has the same yellow color as we have not assigned an identity to each worker yet. Then, the outputs of the worker detector will be fed to the worker tracker, where we assign the identity to each worker based on temporal information. In the tracking outputs, each box is painted with a different color in the image as we know their identities now.

As we have mentioned in the related work, we use FasterRCNN [15] network as our worker detector, which was originally designed for universal object detection. To adapt the network for our task, which is to differentiate foreground workers and background, we change the dimension and parameters in the last few layers of the FasterRCNN so that the network only outputs for two classes instead of near 100 classes defined in COCO [22]. Also, to output more accurate worker detection, we change the size and aspect ratio of the default anchors so that the anchors are more aligned with the size and aspect ratio of our workers. To achieve the best possibly performance, we first train the original network on ImageNet [23] and COCO to learn universal features, and then we fine-tune the network on our own datasets with construction workers. We will show in the experiments that our pre-training help improve the performance and generalization to new videos compared to directly train on our dataset from scratch.

Besides customizing our worker detector, we also track the detected workers over time. Specifically, we use Kalman filter-based tracking method SORT [19] as our worker tracker. In this way, we remove the need of training for tracking algorithm and can also achieve real time worker safety monitoring. Beyond tracking in the images, we also develop a 2.5D tracking method by using homography transformation technique. As a result, we can visualize the resulting worker motion trajectory in the top down view so that it is helpful to site manager for better understanding the worker's activities. The final output worker trajectories from our tracker will be used as inputs to our action recognition system to obtain detailed action category information for each worker.

3.3 Worker action recognition

Following detection and tracking of the objects in the video, we further conduct action recognition with a two-stream 3D-ConvNet architecture. Our architecture follows i3D with an RGB appearance stream and an optical flow dynamics stream. As stated in the previous section, our tracker will aggregate bounding boxes in continuous frames of the same person as video clips, these videos are treated as input of our RGB stream. We use an off-the-shelf optical flow extraction algorithm from OpenCV toolbox. In order to ignore the background shift caused by bounding box shift, we calculate the flow directly from the video, and clip the patch from the generated flow videos.

We use 3D Inception-v1 as our backbone for feature extraction as used in i3D [35]. The image stream takes RGB frames as input, and flow stream takes optical flow frames as input. Each video clip is stacked by 60 continuous frames. After backbone network extracts appearance and dynamic feature maps, we leverage the late fusion strategy by averaging the two feature maps. Finally, a classification head is used to get the final the prediction. Horizontal flip and random translation of bounding box is used as data augmentation, and the same augmentation is conducted on all frames of a video sample.

3.4 Data usage

Site manager needs to grasp whether the daily prog-

Table 2. Method to get activity element

Element	Data type	Method
When	Datetime	Read from video metadata
Who	WorkerID	Use output from worker tracking
Where	XYZ in real	Calculate from 2D coordination in image and camera parameter
What	Work name	Search from worklist by When and Where
How	Process of action	Use output from worker action recognition

ress is as planned and updates the schedule appropriately. Daily progress can be grasped from some elements; when the work has done, who done the work, where is the worker done the work, what is the work, how the worker has progressed the work, as shown in Table 2.

4 Result and Discussion

4.1 Dataset creation

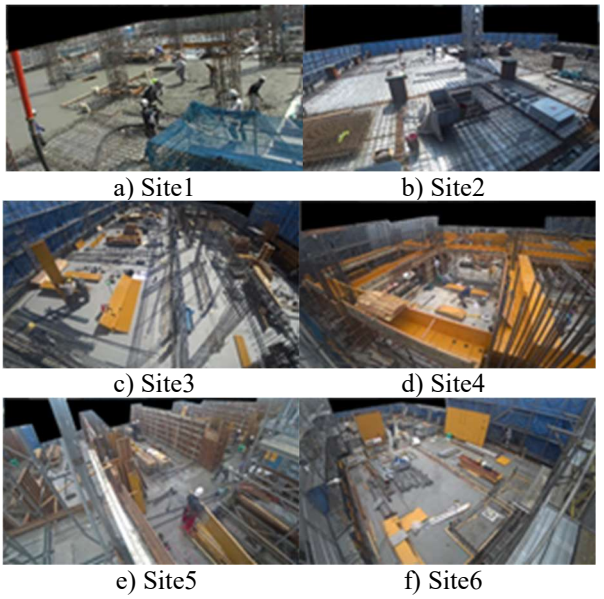
Selection of sites. We selected 6 sites in Japan managed by Shimizu Corporation which were under construction of main structures. Example image of 6 sites are shown in Figure 3.

Negotiation about permission. By promising that the image of worker not to link the workers' personal data (i.e. name), the legal department of Shimizu Corp and CMU permitted this study. Then we got permission about capturing movie from onsite manager and the owner of the building which the site construct. Then we displayed notification of movie capturing and its purpose for the workers in the site.

Capture movies and pickup movies. We set 2 cameras (Sony-FDRX3000, JVC-GZRY980) onto outer scaffolding frame using metal clip on little look-down direction. After capturing, we checked all movies and selected or cut into 52 movies on 17 scenes. As shown in Figure 3, site 1 was concrete placing work on sunny day, site2 was placing rebar of floor work in sunny day, site3 was formwork carrying in sunny day, site4 and site5 were formwork related work in cloudy day, site6 was formwork and scaffolding work in rainy day. Table 3 shows total length of movies. Total frame count was about 270k, and total length was about 2.5h.

Annotate worker location and parameter. Annotation was done by over 10 annotators hired by outsourcing company and spent over 2 months. The total count of data is shown in Figure 4. Since site1 and site4 had over 10 workers in each frame, total data count was be huge. Figure 4 colored by size category of height of bounding box (i.e. count of pixel), by this, we can understand the workers image size almost in Easy (>=40pixel) category. Action category ratio is shown in Figure 5. It can be seen "other" category occupied more then half of all and can be seen basic 6 categories has mis-balanced.

In the future data set creation, it will be necessary to set the video length according to the number of workers included in the video, and to define the action category as avoiding a mismatch in the number of tags.



a) Site1 b) Site2

c) Site3 d) Site4

e) Site5 f) Site6

Figure 3. Example of each 6 sites' image

Table 3. Total length of movies each 6 sites

Name	Weather	Movie Count	Frame Count	FPS	Total length
Site1	Sunny	9	50535	29.97	28'06"
Site2	Sunny	5	30591	29.97	17'01"
Site3	Sunny	7	61695	29.97	34'19"
Site4	Cloudy	9	64637	29.97	35'57"
Site5	Cloudy	4	24825	29.97	13'48"
Site6	Rainy	18	38565	29.97	21'27"
Total	-	52	270848	-	150'37"

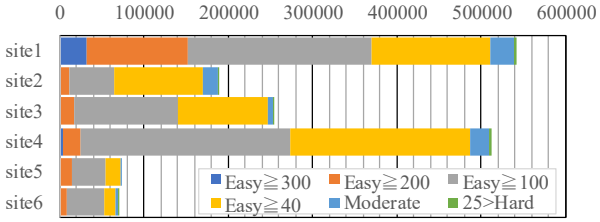


Figure 4. Data amount and size ratio

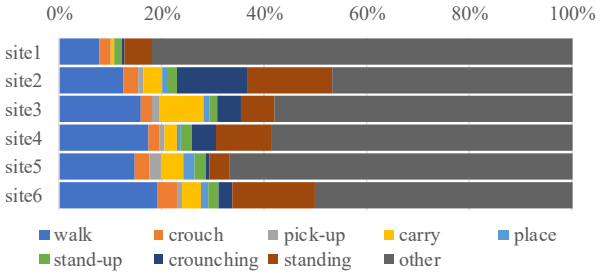


Figure 5. Action category ratio

4.2 Worker detection evaluation

We evaluate our worker detector on our dataset collected from the Shimizu Corp. construction site. We use the standard metric of average precision (AP) defined with an IoU (intersection of union) threshold of 0.5 for worker detection evaluation. Also, we use three difficulty levels for evaluation (easy, moderate and hard) where each difficulty level has a different threshold on the number of pixels (40, 25, 1 respectively) in the ground truth worker's bounding box height. For example, for the easy case, we filter out ground truth workers that have height less than 40 pixels and only evaluate the rest of workers which are relatively closer to the camera and easy to be detected. We compare our customized worker detector trained on our Shimizu data with generic detectors such as Faster RCNN [15] and Mask RCNN [24] pre-trained on the COCO dataset.

We use two types of data split during evaluation: (1) same site evaluation; (2) cross-site evaluation. For each construction video in the same site evaluation, we use first 70% frames for training, middle 15% for validation and the last 15% for testing. We summarize our results for same site evaluation in Table 4. We can see that, without any customization, the generic object detectors do not perform well on our construction dataset, and have lower performance than our customized detector, e.g., 33.5 AP of the Mask RCNN v.s. 66.8 AP of our detector in the easy case. Also, when we pretrain our customized detector on the COCO dataset before fine-tuning on our construction data, the final performance can be even higher, e.g., 87.9 AP v.s. 66.8 AP in the easy case. This proves that pretraining on the COCO dataset with many generic objects does help improve final detection performance on our construction dataset.

In addition to same site evaluation, we also perform the cross-site evaluation, i.e., evaluation across different construction site videos. This means that we select one site (e.g., 5 videos from the site 2) as the testing data, while using the rest of other data from other sites for training and validation. In this way, the detection is more difficult than in the same-site evaluation as the data in the testing set is not seen during training. We summarize the results in Table 5. We can see that our customized detector still achieves reasonable performance when evaluating on the construction sites that are not seen during training. This can be useful to construction manager as our customized detector can be potentially applied to other new site videos collected in the future. Also, same as the same site evaluation, our customized detector works better than the generic object detector Faster RCNN in the cross-site evaluation (e.g., 77.5 AP of our customized detector v.s. 50.7 AP of the generic detector in the easy case).

Table 4. Quantitative performance of worker detection in same-site evaluation

Method / AP	Easy (>=40)	Moderate (>=25)	Hard (>=1)
Mask RCNN (COCO)	31.7	31.3	31.3
Faster RCNN (COCO)	33.5	32.1	31.9
Our Detector (Shimizu)	66.8	66.5	66.4
Our Detector (COCO+Shimizu)	**87.9**	**87.4**	**86.2**

Table 5. Quantitative performance of worker detection in cross-site evaluation

Method / AP	Easy (>=40)	Moderate (>=25)	Hard (>=1)
Faster RCNN (COCO)	50.7	48.7	48.3
Our Detector (COCO+Shimizu)	**77.5**	**72.8**	**72.7**

4.3 Worker action recognition

Similar to detection and tracking task, we evaluate our action recognition model on our dataset collected from the Shimizu Corp. construction site. Specifically, we first pre-train the model on Kinetics and fine-tune on our dataset. The dataset is randomly split into 80% training set and 20% validation set. We evaluate the performance of each class by their recall rate, and the overall performance is measured by mean class accuracy. In order to get more information about surrounding environment, we expand the bounding box from detector by 10%. The learning rate for our model is set to 0.0005 with a weight decay 0.0001. Batch size is set to 4. We show some qualitative results of our model in figure 6. Note that the real data is in video format, and we show a representative frame of each video clip.

We summarize our quantitative results in Table 6. We can see that our method performs better than generic ST-GCN [38] and i3D [35]. Note that ST-GCN was trained on an early labelling strategy where standing and crouching are not specifically labeled, making the dataset simpler. As we can see, even with a simpler task, generic ST-GCN performs worse than our method. And generic i3D only performs marginally better on walk and carry class. We can see that for Other class, out method performs the worst, this is because we are using recall as our metric. A high recall of un-defined class is usually the result of many false positive samples in other classes. Also, we notice that the performance on Place class is very low. This is because of the internal imbalance of our dataset in number and

difficulty of each class. In real-world construction site, “Place” happens far less frequent than other classes of actions like walk or standing. Moreover, it usually lasts very short, sometimes less than 1 second. These results in the high relative difficulty of class “Place”.

Figure 6. Qualitative performance of our worker action recognition model

Table 6. Quantitative performance of worker action recognition

Method	Mean Class Accuracy	Walk	Crouch	Stand-Up	Pick-Up	Carry	Place	Standing	Crouching	Other
ST-GCN	33.1	34.7	49.2	49.5	14.7	8.1	**18.7**	-	-	**57.2**
i3D	54.0	**80.7**	37.2	85.9	38.2	**79.6**	5.6	73.8	30.3	55.0
Our Method	**60.2**	75.5	**84.1**	**90.5**	**59.0**	61.1	6.1	**82.2**	**34.3**	50.0

Table 7. Ablation study on our worker action recognition model

Method	Mean Class Accuracy	Walk	Crouch	Stand-Up	Pick-Up	Carry	Place	Standing	Crouching	Other
RGB	51.6	44.0	82.0	86.0	56.0	77.0	0.0	71.0	16.0	28.0
Flow	56.2	68.0	82.0	83.0	49.0	40.0	**18.0**	77.0	**39.0**	48.0
RGB + Flow	**60.2**	**75.5**	**84.1**	**90.5**	**59.0**	**61.1**	6.1	**82.2**	34.3	**50.0**

We also conduct ablation study as shown in Table 7. The first two rows are using only RGB stream or optical flow stream without late fusion. We see that using only one stream will result in a performance drop in mean class accuracy and most classes. This experiment also proves that RGB is better at extracting appearance feature like “Pick-up” and “Carry” which involves interaction with other objects that might not be visible in optical flow. On the other hand, optical flow steam helps the model by providing better dynamic feature, and thus improves classes like “Walk” and “Place”.

4.4 Data usage

The data acquisition technology was implemented except for the "How" element. A result of data acquisition using our method is shown in Figure 7. The rectangle frame was output from detector and the “id001” was output from tracker. The line in bounding box indicated a vertical line calculated from 2D coordination of bounding box output from worker detector and camera parameter. The XYZ and height of worker were calculated by regarding the center of the frame as the midpoint of the worker's height. It also can be seen the work name was successfully searched from list by datetime and XYZ coordination. The activity elements were acquired as planned.

Figure 7. Translation to activity element

5 Conclusion

We first built a dataset for a Japanese construction site. Then, the domain specific algorithms of worker detection, tracking, and worker action recognition were customized. As a result, our worker detection showed 87.9% accuracy in same-site evaluation and 77.5% accuracy in cross-site evaluation. Our worker action recognition showed 60.2% mean accuracy. Finally, the method of translation into activity element based on the output value of worker detection was indicated. Whether these accuracies are actually sufficient needs to be verified in the future.

As a limitation, this method cannot deal with occlusion because it uses a single camera. It is possible to deal with this by installing multiple units on site so that occlusion is reduced.

As future work, it is essential to improve accuracy by increasing the data set, and it will be necessary to build technology to ensure data consistency when multiple cameras are installed.

6 Reference

[1] Tarek O. and Moncef N., Data acquisition technologies for construction progress tracking, Automation in Construction 70 (2016) 143-155

[2] Amin A. et al., Sensor-based safety management, Automation in Construction 113 (2020) 103128

[3] Construction industry handbook 2019 [in Japan ese], Japan federation of construction contractors, https://www.nikkenren.com/publication/pdf/handbook/2019/2019_04.pdf, Accessed: Jun./06/2020

[4] Jun Y. et al, Automatic Recognition of Construction Worker Activities Using Dense Traj ectories, 2015 Proc. of the 32nd ISARC, Pages 1-7

[5] Kaijian L. and Mani G., Crowdsourcing Video-based Workface Assessment for Construction Activity Analysis, 2015 Proc. of the 32nd ISARC

[6] Mohammad S. et al, Evaluating the Performance of Convolutional Neural Network for Classifying Equipment on Construction Sites, 2017 Proc. of the 34rd ISARC, pages 509-516

[7] Mingzhu W. et al, Predicting Safety Hazards Among Construction Workers and Equipment Using Computer Vision and Deep Learning Techniques, 2019 Proceedings of the 36th ISARC, Pages 399-406

[8] Eirini K. and Ioannis B., Trajectory-Based Worker Task Productivity Monitoring, 2018 Proc. of the 35th ISARC, Pages 1145-1151

[9] Ye H. et al, A Visualization System for Improving Managerial Capacity of Construction Site, 2017 Proceedings of the 34rd ISARC, Pages 388-395

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

[12] Felzenszwalb et al. A Discriminatively Trained, Multiscaled, Deformable Part Model. CVPR 2008.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object dete ction and semantic segmentation. In CVPR 2014

[14] R. Girshick. Fast R-CNN. In ICCV, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: To- wards real-time object detection with region proposal net- works. In NIPS, 2015.

[16] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In CVPR, 2017.

[17] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. IJCV, 2013.

[18] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In NIPS, 2016.

[19] A. Bewley, G. Zongyuan, F. Ramos, and B. Upcroft, Simple online and realtime tracking. in ICIP, 2016.

[20] N. Wojke, A. Bewley, D. Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. ICIP 2017.

[21] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-Object Tracking and Segmentation. CVPR, 2019.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Doll´ar, and C. Zitnick. Microsoft COCO: Com- mon objects in context. In ECCV 2014.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR 2009.

[24] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask RCNN. In ICCV 2017.

[25] Wang, Heng, et al. "Action recognition by dense trajectories." CVPR 2011.

[26] Wang, Heng, et al. "Action recognition with improved trajectories." In ICCV 2013.

[27] Karpathy, Andrej, et al. "Large-scale video classifyation with convolutional neural networks." In CVPR 2014.

[28] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fu sion for video action recognition." In CVPR 2016.

[29] Chen, Yunpeng, et al. "A^ 2-nets: Double attention networks." Advances in Neural Information Processing Systems. 2018.

[30] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems. 2014.

[31] Wang, Limin, et al. "Temporal segment networks: Towards good practices for deep action recognition." In ECCV 2016.

[32] Sigurdsson, Gunnar A., et al. "Asynchronous tem poral fields for action recognition." In CVPR 2017.

[33] Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." In ICCV 2015.

[34] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." In ICCV 2015.

[35] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In CVPR 2017.

[36] Diba, Ali, et al. "Spatio-temporal channel correlation networks for action classification." In ECCV 2018.

[37] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." In ICCV 2019.

[38] Yan, Sijie, Yuanjun X., and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In AAAI 2018.