ICCEPM 2022
The 9th International Conference on Construction Engineering and Project Management
Jun. 20-23, 2022, Las Vegas, NV, USA

Vision-Based Activity Recognition Monitoring Based on Human-Object Interaction at Construction Sites

Yeon Chae[1]*, Hoonyong Lee[2] Changbum R. Ahn[3], Minhyuk Jung [4], Moonseo Park[5]

[1] *Department of Architecture and Architectural Engineering, Seoul National University, South Korea*, E-mail address: yeonchae62@snu.ac.kr
[2] *Department of Construction Science, College of Architecture, Texas A&M University, Texas, USA*, E-mail address: onarcher@ tamu.edu
[3] *Department of Architecture and Architectural Engineering, Seoul National University, South Korea*, E-mail address: cbahn@snu.ac.kr
[4] *Department of Architecture and Architectural Engineering, Seoul National University, South Korea*, E-mail address: archidea914@snu.ac.kr
[5] *Department of Architecture and Architectural Engineering, Seoul National University, South Korea*, E-mail address: mspark@snu.ac.kr

Abstract: Vision-based activity recognition has been widely attempted at construction sites to estimate productivity and enhance workers' health and safety. Previous studies have focused on extracting an individual worker's postural information from sequential image frames for activity recognition. However, various trades of workers perform different tasks with similar postural patterns, which degrades the performance of activity recognition based on postural information. To this end, this research exploited a concept of human-object interaction, the interaction between a worker and their surrounding objects, considering the fact that trade workers interact with a specific object (e.g., working tools or construction materials) relevant to their trades. This research developed an approach to understand the context from sequential image frames based on four features: posture, object, spatial features, and temporal feature. Both posture and object features were used to analyze the interaction between the worker and the target object, and the other two features were used to detect movements from the entire region of image frames in both temporal and spatial domains. The developed approach used convolutional neural networks (CNN) for feature extractors and activity classifiers and long short-term memory (LSTM) was also used as an activity classifier. The developed approach provided an average accuracy of 85.96% for classifying 12 target construction tasks performed by two trades of workers, which was higher than two benchmark models. This experimental result indicated that integrating a concept of the human-object interaction offers great benefits in activity recognition when various trade workers coexist in a scene.

Keywords: activity recognition, human-object interaction, vision-based automation in construction

1. INTRODUCTION

Activity recognition of construction workers is critical in monitoring construction performance and protecting workers' safety and health. Recognizing an individual worker's activity provides key data in tracking and analyzing labor productivity in construction tasks [1,2]. Also, recognized activity of a worker helps detect hazardous situations and identify at-risk workers on construction sites [3]. With the development of sensing technologies and activity recognition algorithms, wearable inertial measurement unit (IMU) sensors have been used for workers' activity recognition by monitoring an individual worker's body part movements [4]. However, wearing sensors would be intrusive while performing various tasks and individual monitoring systems are required for each worker, requiring high computational costs to simultaneously monitor multiple workers at the workplace [2]. Alternatively, vision-based approaches [1,2] provide a nonintrusive monitoring method and classify a worker's activities from continuous image frames recorded using surveillance cameras. As an image frame includes both workers and background, the recognition model that is trained with the entire image region is dependent on the background information. The model performance would be degraded when the model is used in different workplaces having different backgrounds [5]. In order to alleviate such dependencies, many studies [2,6] exploited workers' postural information in activity recognition. This approach extracts skeleton data of a worker from sequential image frames, tracks body joint positions across image frames, and then estimates activities per frame. However, such worker-oriented activity recognition would be vulnerable when different activities have similar patterns of sequential postures. For example, both *transporting rebar* and *transporting formboard* include similar postures, including walking and carrying.

In this context, this study aims to develop an approach to classify various activities across different trade workers that include similar postures. Given the fact that some objects (e.g., working tools) that trade workers interact with often provide information on their trades, this study exploits such human-object interaction information in recognizing activities in addition to postural information. The developed approach used convolutional neural network (CNN) to extract four different types of information, including postural features, object features, temporal features, and spatial features. These features were independently used to yield each activity score from CNN- and long short-term memory (LSTM)-based activity classifiers. Then, the average of three activity scores was used to generate the final estimation on performed activities.

2. BACKGROUND

2.1. Activity Recognition in Construction

With the development of sensing technologies and machine learning algorithms, many researchers have developed sensor-based activity recognition systems that collect sensory data and estimate target activities in real time [7–11]. As an activity is composed of several body part movements in some sequential order, wearable IMU sensors have been mainly used for collecting an individual worker's bodily movements by being attached to the target body parts. However, wearable IMU-based activity recognition requires workers to wear multiple sensors, which would be intrusive while performing various tasks [7]. Moreover, the sensors attached to the body part may collide with other obstacles or body parts, which changes the sensor location or alignment. This leads to low-quality data collection.

Previous studies [1,6] developed an approach to detect temporal changes from sequential image frames, such as the pattern of workers' bodily movements by considering both spatial and temporal information. Roberts et al. [2] extracted postural features from image frames that tracked movements of each body joint based on spatial-temporal understanding. These studies demonstrated the feasibility of vision-based activity recognition in construction, but their performance would be greatly hampered on real-world construction jobsites where workers with various trades coexist. Since these studies mainly relied on workers' postural information in activity recognition, they may not be able to correctly estimate tasks that include similar postural patterns across different trade workers.

2.2. Human-Object Interaction in Activity Recognition

In vision-based activity recognition, human-object interaction (HOI) has an aim of localizing both humans and objects and identifying their interactions at the same time [12]. As HOI itself provides additional contextual information, it has been exploited to recognize an activity that involves a specific interaction between human and object from a single image frame [13]. In the construction domain, HOI-based activity recognition has been exploited mainly for safety inspections by analyzing interactions between workers and target objects [14–16]. Xiong et al. [14] developed an approach to monitor whether workers properly wear their safety helmets by tracking both the movements of workers and the locations of safety helmets. Tang et al. [15] adopted HOI in their risk analysis after recognizing a single worker's activity to find out the surrounding hazards at the workplace. The HOI approach was expanded to detect a worker who performs a heavy manual task (e.g., excessive load-carrying) by analyzing the interaction between the worker and construction materials [16].

While these studies have shown the potential use of the HOI approach in construction, these studies only used such spatial-temporal information (i.e., relative locations of objects and a human) for their safety management without the awareness of how a change of these pairs affects a single activity. However, in order to classify fine-grained construction tasks, it would be necessary to consider the temporal information (i.e., how relative locations of objects change) of HOI from sequential images. Moreover, workers with different trades would have distinct sequential patterns of their postures and interactions with objects. In this context, this study developed an activity recognition approach that extracted the spatial-temporal interaction between workers and objects from sequent image frames.

3. METHODOLOGY

Figure 1 shows the overall framework of the developed approach that includes four streams. Each stream has different feature extractors and these features are fed into each activity classifier for the input data; five consecutive image frames are used as the input data. Pose stream and object stream are combined, recognizing activities based on the interaction between a worker and their surrounding objects. The other two streams extract spatial feature and temporal feature, respectively. The developed model yielded three activity scores from each stream that may contain prediction errors with mispredicted labels due to the imperfection of each frame in a specific case.

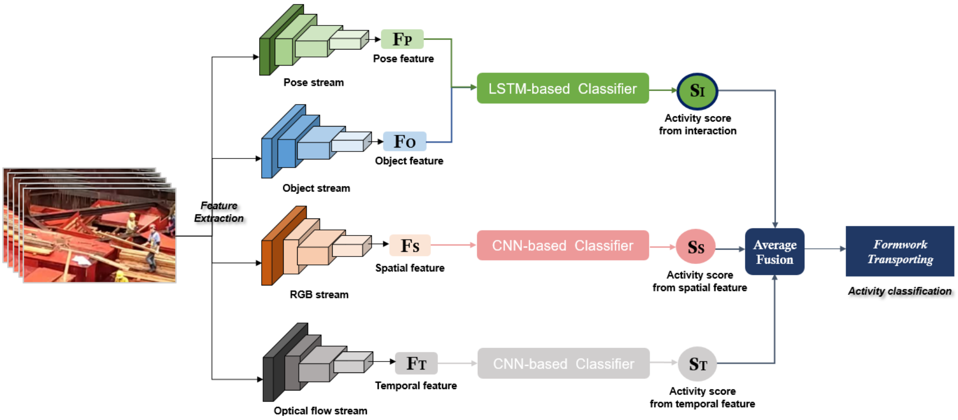


Figure 1. The overall framework of the developed approach

In order to alleviate such misclassification, the developed approach averaged all the activity scores and selected one that has the highest probability as the predicted activity. The detailed process of each stream is described in the following section.

3.1. Activity Recognition Based on HOI

The first stream includes two feature extractors to extract a worker's postural features and their surrounding object's locational features. Each feature extractor was constructed based on CNNs from previous studies [17,18]. The first extractor was built based on Fang et al. [17] that detected a worker's posture in construction. It extracted 18 coordinates of the worker's body joints, such as nose, shoulders, elbows, wrists, hips, knees, and ankles. The output shape of the extractor is set as $N \times 36 \times 1$, which respectively represents the number of samples, 18 coordinates in the *x*- and *y*-axis, and the number of channels. The second extractor is constructed using the you only look once (YOLO) model [18] to detect the type of the object and its central coordinate. In this study, six types of objects (two of them represented materials and four represented tools) were selected as the target objects: formboard, rebar, hammer, rebar machine, welding flame, and formboard machine. The output shape was as $N \times 12 \times 1$, which represents the number of samples, the combination of the object's type and its central location in the *x*- and *y*-axis, and the number of channels. For each image frame, the two outputs from the first and second feature extractors were combined to yield $N \times 48 \times 1$ shape. The combined feature represented the interaction between the worker and the target object. Figure 2 shows how extracted pose and object features are combined. The combined features are then inputted into an activity classifier constructed based on the LSTM networks that consists of two LSTM layers, one dropout layer with the ratio of 0.5, and one dense layer. Each layer of LSTMs has 64 memory cells, and this LSTM model used the softmax as the activation function to provide each probability for all the target activities, and the sum of these probabilities is 1.

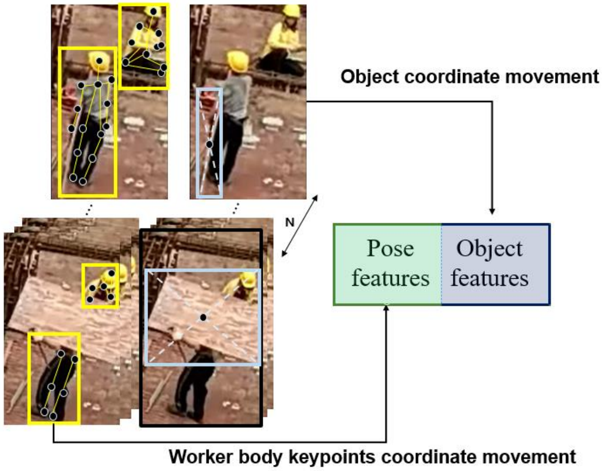

Object coordinate movement
N
Pose features
Object features
Worker body keypoints coordinate movement

Figure 2. Feature combination of workers' posture and their surrounding objects

3.2. Activity Recognition Based on Spatiotemporal Streams

The other two streams, the RGB and optical flow streams, detect spatiotemporal information from the entire image region. The RGB stream is constructed using Krizhevsky et al. [19] and interprets RGB values of all the pixels within the frame. The features extracted from this stream provide spatial information when it is fed on a pretrained CNN-based classifier [19]. The optical flow stream is constructed using Simonyan & Zisserman [20] and contains a stacked information of displacement vector [20]; this vector consists of direction (increment of *x*- and *y*- axis of pixel movement) and distance (size of their movement in *x*- and *y*-axis) between pairs of consecutive frames. For both streams, the input shape was as $340 \times 256 \times 3$, which represents the image size and the number of channels (RGB channels). During finetuning, through randomly jittering the image, the input is rescaled to 224×224. Then, rescaled images are then inputted into optical flow stream that consists of five convolutional layers with an activation function of relu. Here, the first and second convolutional layers used a kernel with stride size of 2, and padding, and each kernel size is 7 and 5 for each of such two convolutional layers. After each of these two layers, max pooling layers are added with pooling size of 2 that reduces the size of feature to the half. Afterward, additional three layers used a kernel size of 3, with stride 1 and padding. Then, three dense layers are added. Among the three dense layers, first two is used with the activation function of relu, while the last layer is activated with softmax function. On the other hand, the RGB stream extracted the features from the rescaled image based on Resnet 50 network [21]. Then, three dense layers with drop out layers are used. The spatial features from such RGB stream focused on the entire image scene of the individual frame, while the temporal features resulting from the optical flow stream traced the change in motion across the frames, extracting the movement of objects or workers. Both streams provided probabilities for all the target activities from the softmax activation. In fusing the probabilities with that from Section 3.1, these two streams would help recognize activities, especially when a worker or object is partially detected in the image frames, since these streams referenced the entire region of images.

4. Evaluation

This study used the data set available from the research community [1]. The data set was collected by a pan-tilt-zoom camera at 30 frames per second mounted on a scaffolding at the height of around 15 m to the working floor. The data set included 12 activities performed by workers in two trades as summarized in Table 1. For each activity, different numbers of video clips were collected and each clip included 90 consecutive image frames. For all activities, 70% of clips were randomly selected as the training data and the remaining 30% of clips were used as the testing data. A five-frame moving window with four frames of overlap was used for data sampling, which provided 34,744 training and 14,878 testing data samples.

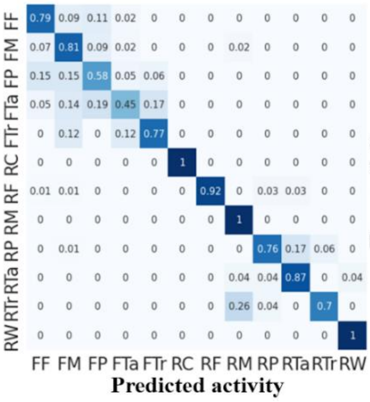
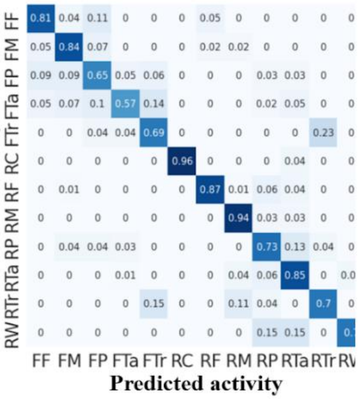
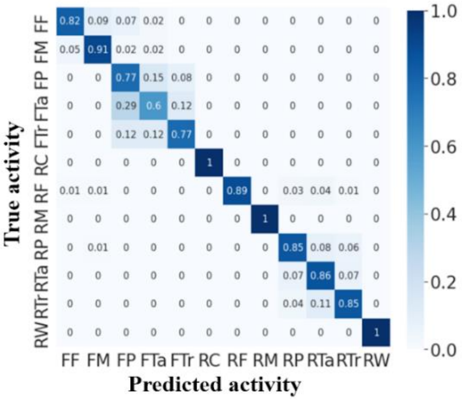
Table 1. Target activities with corresponding trades

Formwork		**Rebar**	
Activity (label)	**Number of clips**	**Activity (label)**	**Number of clips**
Fixing (FF)	57	Fixing (RF)	71
Machining (FM)	57	Machining (RM)	33
Placing (FP)	65	Placing (RP)	78
Taking (FTa)	42	Taking (RTa)	71
Transporting (FTr)	26	Transporting (RTr)	27
		Connecting (RC)	23
		Welding (RW)	27

The developed model was trained for 500 epochs with 10 batch size. The developed model provided an average of 85.96% classification accuracy, which had an average of 77.41% accuracy for the formwork trade and 92.06% accuracy for the rebar trade. In terms of precision and recall rate, each precision rate in the same order as Table 2 is, 0.79, 0.78, 0.78, with the recall rate of 0.80, 0.78, 0.86 each. The developed approach showed a higher activity classification accuracy on the rebar trade than the formwork trade. This performance difference originates from the characteristics of each trade. A rebar worker interacted with diverse objects (e.g., welding flame or rebar machine), and thereby recognizing objects that a rebar worker interacted with helped in recognizing her activity. However, a formwork worker interacted with only formboards, so HOI was not that effective in differentiating a formwork worker's activities.

The performance of the developed model is compared with two benchmark models [1,2]. The first benchmark model, Luo et al. [1] uses the spatial and temporal features from the entire image region, which are corresponding to the two streams that extract spatial and temporal features in our approach (see Section 3.2). This study uses the same data set in the validation, so their reported average accuracy and confusion matrix in the second split of cross validation are presented in Table 2. The second benchmark model is constructed based on Roberts et al. [2]. This model adds an individual worker's postural features to the first benchmark model. Table 2 summarizes accuracies of these models with the confusion matrix. Our model provides a higher performance as compared to the benchmark models. Since the first benchmark model did not consider a worker's posture, it provided lower accuracies to classify different activities of a single trade, as shown in the confusion matrix (see Table 2, false detection within a single trade) where false detection occurs within each trade. On the other hand, the second benchmark model provided lower accuracies to classify different trades because it mainly focused on an individual worker's posture for activity recognition (see Table 2, false detection between trades). The developed model, however, considered workers' postures, their interaction with objects, and spatiotemporal changes in the input image frames, and had less false detection within and between trades. This result demonstrates that the integration of HOI offers particular benefits in recognizing activities across multiple trades.

Table 2. Comparison with benchmark models

Luo et al. [1]	**Roberts et al. [2]**	**Ours**
Accuracy 80.5%	Accuracy 77.60%	Accuracy 85.96%
		

5. DISCUSSION

The development of vision-based activity recognition has many benefits, but there have been fragmented efforts [1,2,6] for construction-specific activity recognition. However, there is still a practical limitation from the absence of a construction-specific benchmark data set on activity recognition. A large amount of data is necessary for training, meaning that such cumbersome tasks should be repeated, not integrated and shared with the community. Especially on construction sites, various situations could occur; workers from multiple trades could cowork together, and a worker from a single trade could do various activities. In these situations, fine-grained activity classification is needed to recognize activities and their trades. In addressing this issue, this HOI-integrated activity classification approach could unify these fragmented efforts, as it demonstrates the potential of sharing each feature; transporting-like bodily movements could be paired with unknown tools (e.g., grinder, hawk). This HOI-based approach has the potential to generate a shared construction-specific knowledge engine, which could make full use of current practices and reduce additional annotating costs at the same time.

Furthermore, the results from this approach could enable the model to obtain explainability, through deeper scene understanding of its context. While previous studies [1,2,6] interpreted prediction results partially oriented to workers, this approach integrates HOI and shows how the triplet of worker, objects, and the entire image scene could complement each other. However, it still has a limitation in activity classification since the HOI only tracked the relative movements between an individual worker and the target object. Therefore, when the relative movements have similar patterns for different activities, the developed approach could not differentiate activities. For example, *taking rebar* is mispredicted as *placing rebar* and *transporting rebar*. Therefore, more contextual information needs to be exploited to classify fine-grained activities. As interactions between workers would provide additional contextual information, the future study will further consider a worker's interaction with coworkers by expanding the HOI approach to worker-object-worker interactions.

6. CONCLUSION

By integrating HOI in worker activity recognition on-site, the developed model provided higher performance compared to the state-of-the-art current practices, especially for differentiating similar patterns of workers' postures when they use a specific object for each activity. It showed 85.96% average accuracy, which is 5.53% and 8.36% higher than each benchmark model, respectively (see Table 2), for each activity. For each trade, the proposed model shows a 9.36% higher performance for formwork and 2.79% higher performance for rebar work compared to the first benchmark model from Luo et al. [1], and 6.23% and 9.89% higher performance for each trade, respectively, compared to the second benchmark model based on Roberts et al. [2]. Based on this result, the proposed model shows less prediction error in both differentiating workers' trades and recognizing worker's activities in a single trade. Further, the proposed method of HOI-integrated activity recognition approach could be generalized by adding different posture sets and object tools sets. This approach has great potential in integrating current individual studies through sharing each partial set. Through these shared sets, representative benchmark data sets will be generated, finding the key feature to recognize activities performed by different trades of workers.

ACKNOWLEGEMENTS

This work was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) Project in 2022 (No.4120200113771).

REFERENCES

[1] X. Luo, H. Li, D. Cao, Y. Yu, X. Yang, & T. Huang (2018). Towards efficient and objective work sampling: Recognizing workers' activities in site surveillance videos with two-stream convolutional networks. Automation in Construction, 94, pp. 360-370.

[2] D. Roberts, W. Torres Calderon, S. Tang & M. Golparvar-Fard (2020). Vision-based construction worker activity analysis informed by body posture. Journal of Computing in Civil Engineering, 34(4), 04020017.

[3] J. Seo, S. Han, S. Lee & H. Kim (2015). Computer vision techniques for construction safety and health monitoring. Advanced Engineering Informatics, 29(2), pp. 239-251.

[4] S. S. Bangaru, C. Wang, S. A. Busam & F. Aghazadeh (2021). ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors. Automation in Construction, 126, 103653.

[5] G. Moon, H. Kwon, K.M. Lee, M. Cho, IntegralAction: Pose-driven Feature Integration for Robust Human Action Recognition in Videos, in: 2021 IEEECVF Conf. Comput. Vis. Pattern Recognit. Workshop CVPRW, IEEE, Nashville, TN, USA, 2021: pp. 3334–3343. https://doi.org/10.1109/CVPRW53098.2021.00372.

[6] J. Yang, Z. Shi, & Z. Wu (2016). Vision-based action recognition of construction workers using dense trajectories. Advanced Engineering Informatics, 30(3), pp. 327-336.

[7] K. Yang, C. R. Ahn, M. C. Vuran, & S. S. Aria (2016). Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit. Automation in Construction, 68, pp. 194-202.

[8] H. Lee, K. Yang, N. Kim & C. R. Ahn (2020). Detecting excessive load-carrying tasks using a deep learning network with a Gramian Angular Field. Automation in Construction, 120, 103390.

[9] J. Chen, J. Qiu & C. R. Ahn (2017). Construction worker's awkward posture recognition through supervised motion tensor decomposition. Automation in Construction, 77, 67-81.

[10] H. Jebelli, C. R. Ahn & T. L. Stentz (2016). Fall risk analysis of construction workers using inertial measurement units: Validating the usefulness of the postural stability metrics in construction. Safety science, 84, 161-170.

[11] C. Sun, S. Ahn & C. R. Ahn (2020). Identifying workers' safety behavior–related personality by sensing. Journal of construction engineering and management, 146(7), 04020078.

[12] B. Kim, J. Lee, J. Kang, E. S. Kim & H. J. Kim (2021). Hotr: End-to-end human-object interaction detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 74-83.

[13] Y. L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H. Fang, Z. Ma, M. Chen, C. Lu (2020). Pastanet: Toward human activity knowledge engine. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 382-391.

[14] R. Xiong, Y. Song, H. Li & Y. Wang 2019). Onsite video mining for construction hazards identification with visual relationships. Advanced Engineering Informatics, 42, 100966.

[15] S. Tang & M. Golparvar-Fard (2021). Machine Learning-Based Risk Analysis for Construction Worker Safety from Ubiquitous Site Photos and Videos. Journal of Computing in Civil Engineering, 35(6), 04021020.

[16] S. Chen, & K. Demachi (2021). Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. Automation in construction, 125, 103619.

[17] H. S. Fang, S. Xie, Y. W. Tai & C. Lu (2017). Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE international conference on computer vision, pp. 2334-2343.

[18] A. Bochkovskiy, C. Y. Wang & H. Y. M. Liao (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[19] A. Krizhevsky, I. Sutskever & G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

[20] K. Simonyan & A. Zisserman (2014). Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27.

[21] X. Tian & C. Chen (2019, September). Modulation pattern recognition based on Resnet50 neural network. In 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), pp. 34-38.