Development of Action-Recognition Technology Using LSTM Based on Skeleton Data

Hechen Yun^a, Etsuro Nakamura^a, Yoichi Kageyama^{a,*}, Chikako Ishizawa^a
Nobuhiko Kato^b, Ken Igarashi^b, Mamoru Suzuki^b

^aAkita University, 1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan

^bADK Fuji System, Inc., 110-3 Tegata Yamazaki, Akita-shi, Akita 010-0851, Japan

*Corresponding Author: kageyama@ie.akita-u.ac.jp

Abstract

Recently, the population of workers of adults aged over 55 years is growing at work sites. Middle-aged adult workers have a higher occurrence rate of work accidents than younger workers. Therefore, it is necessary to develop a safety-management system to ensure safety. This paper proposes an approach for the recognition of human actions based on human-skeleton data as a part of the system in construction industries. The proposed approach consists of four processes: i) extraction of skeleton data from captured video data, ii) interpolation of skeleton joint-points that were missed, iii) calculation of features using interpolated skeleton data, and iv) construction of action-recognition model using interpolated data and calculated features. We evaluated the action-recognition accuracy performance for 5 types of actions from 6 subjects. The evaluation result achieved a high recognition accuracy of 93.1% on average. The results reveal that the proposed approach can be used to recognize actions from video data, and interpolation methods can significantly improve the action-recognition accuracy of the proposed approach.

Keywords: Deep learning, Long short-term memory, Action recognition, Skeleton data.

1. Introduction

Recently, with the development and improvement of deep learning, more industrial and research communities have focused on deep learning for practical applications, such as work assistants, video-based monitoring, and surveillance. In the last decade, the population of middle-aged adults, those aged 55–65 years, is growing at

work sites, especially in the construction industry, where the proportion of older workers is higher than in all industries⁽¹⁾. In general, older workers have a wealth of knowledge and experience that equip them with the ability to make correct decisions and lead groups to success in business⁽²⁾. However, as physical and mental functioning declines with age, older workers have a higher occurrence rate of work accidents compared with younger workers. Therefore, it is necessary to develop a safety-management system to ensure the safety of elderly workers at construction sites. Fig. 1 shows the safety-management system working in the workplace.

At present, many researchers have proposed methods using 3D skeleton data or RGB-D sensor to recognizing workers' actions then improving productivity and safety in construction site^(3,4). However, depending on the construction site, it may become difficult to setting multiple devices in a limited space. On the other hand, image-based action recognition approaches have got significant results that can recognize the basic work actions of construction workers in workplace^(5,6). Although there is a problem that training data is insufficient due to the worker's body is hidden and action features cannot be acquired correctly when detect object carrying large goods and interacting

Fig. 1. System operation schematic.

Fig. 2. Data-acquisition environment.

with construction equipment. Since the problem on action recognition has not examined the above issue in detail, an interpolation processing approach was necessitated to complement action data from construction workers.

In this paper, we propose an approach for recognition of human actions using long short-term memory⁽⁷⁾ (LSTM) based on human skeleton data as a part of the management system. In this approach, the human-skeleton data extracted from the classified images data with OpenPose⁽⁸⁾ and interpolated using three methods if existing missing points exist. Furthermore, features such as joint-joint orientation, distance, and frame-frame trajectory are calculated from interpolated skeleton data. Finally, a time-sequence dataset was created using skeleton data and features for the learning classification model. The performance-evaluation experiment determined the usability based on the calculated time-series features for action recognition. Finally, the precision and recall were demonstrated by using the best result of pattern in the highest recognition accuracy rate of the cross-validation.

2. Data

To verify that the proposed approach has the ability to recognize human action, 5 kind of basic workplace actions were simulated for our video data and are defined below.

- (1) Walking: a person starts walking at a constant speed while waving hands.
- (2) Squatting up: a person starts standing up from squatting position until straighten the back and the legs.
- (3) Squatting down: a person starts squatting from standing position and the knees are deeply bent until stop moving.
- (4) Climbing up: a person starts climbing a stepladder

Fig. 3. Overview of the proposed approach.

to move upwards step by step with hands.

- (5) Climbing down: a person starts climbing down from a stepladder step by step with hands.

The video data was recorded in 1920×1080 pixels at 60 fps. A total of 6 subjects, comprising 4 men and 2 women, participated in this experiment. The actions of one subject were recorded 3 times with a monocular camera (FDR-AX60, SONY) under general fluorescent lights (500–700 lx). In particular, walking actions were captured from left and right when subjects walked between left and right. In addition, walking actions were captured when the subjects approached the camera and moved away from the camera. Moreover, squatting and climbing actions are captured from the front, side, and back. Fig. 2 shows the data-acquisition environment in the experiment. The data used in this investigation were acquired in accordance with the ethical regulations concerning studies involving humans at Akita University, Japan.

3. Proposed Approach

3.1 Overview

As Fig. 3 shows, first, the video data of the subjects are manually classified into each action in frame. Then, the skeleton data is extracted. Furthermore, row skeleton data are interpolated, as there exist some missing points. In addition, we calculate three time-sequential features to quantize the actional information from the skeleton data. Finally, the interpolated skeleton data and features are combined to train the LSTM model, and a trained model was used to recognize actions from each subject.

3.2 Human-Skeleton Data Extraction

In this research, OpenPose and BODY_25⁽⁸⁾ models were used to obtain 25 skeletal joint points and detection confidences, which are the accuracy scores (0.0–1.0) of the currently detected joint points from each frame of the video data. Owing to the camera angle, the body of the subject is hidden and undetectable, and row skeleton data thus contain many missing points. For this reason, we define the missing value as the coordinate of the joint location where the detection confidence is lower than 0.3.

3.3 Interpolation of Skeleton Data

Owing to the camera angles, such as from the side or back, some joint locations of the subject are hidden by the body or stepladder. To reduce the influence of missing points, we proposed four interpolation methods to process row skeleton data. The interpolation procedures using the four methods are listed below.

- (1) To interpolate symmetric joint points (e.g., hands and feet), the coordinate values of the missing point are interpolated by the corresponding point, which could be detectable.
- (2) To interpolate asymmetric missing points, in which limited frames (n) exist between two frames of detectable points (P), the coordinate values of missing points (P_m) are interpolated using linear interpolation, which adds increasing values calculated from the coordinates of the points that were detected first and second. Formula (1) shows the calculation method.

$$P_m(t + i) = P(t) + i \frac{P(t + n) - P(t)}{n}, \quad (1)$$

where t is the position number of frames, which was detected first, and $t+n$ is the position number of frames that was detected after. where i is the position number of frames that were counted from the first detectable frame of the joint point.

- (3) To interpolate asymmetric missing points, the coordinate values of the missing points (P_m) are interpolated using linear interpolation by adding increasing values calculated from the coordinate of points (P) in 5 frames before the missing points or subtracting decreasing values calculated from the points in 5 frames after missing points. Formulas (2) and (3) show the calculation methods.

$$P_m(t + i) = P(t) + i \frac{P(t + 4) - P(t)}{4}, \quad (2)$$

$$P_m(t - i) = P(t) + i \frac{P(t) - P(t - 4)}{4}, \quad (3)$$

where t is the 5th position number of the frames before or after the first frame of the missing point in the video data, and i is the position number of the frames counted from the first detectable frame of the joint point.

- (4) For interpolation when the missing points are both eyes and the nose, based on practical observation of the human face structure, the positions of the nose, ears, and eyes have a triangular relation for both the left and right sides of the face. Therefore, the coordinates of the eyes can be estimated using trigonometric functions. The coordinate of the nasal point is obtained by calculating the middle coordinate of point between the ear locations. Then, using the trigonometric function, the coordinates of the missing eye points are obtained.

3.4 Feature Extraction

In this research, 3 kinds of features were calculated to quantize the action information using interpolated skeleton data. i) A total of 21 orientation features were extracted from 2 joint points, in which 15 orientation features were based on the neck point with arbitrary points within points 2 to 14⁽⁸⁾, and the other 10 orientation features were extracted between each point within the same range of joint points. ii) The distance features were extracted from two arbitrary points within the range from points 0 to 14 in the same frame. iii) The trajectory features were extracted from 2 consecutive frames of the same points.

The interpolated skeleton coordinates and obtained features were used to create a time-sequence dataset to train the LSTM model and test its performance.

3.5 Training process of LSTM model

To recognize the actions captured from subjects using coordinates and features, we designed an LSTM model to learn the patterns of actions and train the recognition ability using time-series variable features from an arbitrary set of 15 consecutive frames. Structurally, the LSTM model has 6 layers, in which 15 neurons in the input layer are a fully connected layer, which was inputted dataset with setting shape, 128, 64, 64, and 32 neurons in the middle layers of the LSTM, and 5 neurons in the fully connected layer as the

Table 1. Total samples of dataset with 6 subjects.

Name	Video count	Frame count	FPS	Sample count
Walking	149	101,714	29.97	99,628
Squatting up	54	2,904	29.97	2,148
Squatting down	54	3,324	29.97	2,568
Climbing up	54	6,624	29.97	5,868
Climbing down	54	6,572	29.97	5,816
Total	365	121,138	—	116,028

output layer with the Softmax function. Adam⁽⁹⁾ was used for the optimization function in the processing of the model.

3.6 Dataset creation

In this study, image data were obtained from the video data processed using FFmpeg⁽¹⁰⁾ that can convert video stream to images with 29.97 fps rate. Then action-image data was manually selected only images in action are included. Samples of dataset was obtained by scanning 15 consecutive frames and shifting 1 frame in the order of the time-series stream, totally gain 116,028 samples after processing, are shown in Table 1. In addition, each sample include 15 consecutive frames and extracted then calculated 306 features (e.g., coordinates, orientation, distance, trajectory) from that. Furthermore, the samples of dataset was used for input to the trained LSTM model, which then recognized the types of actions (e.g., walking, squatting up, climbing up) based on each sample.

4. Experiment and Discussion

4.1 Contents of Experiment

In this research, we evaluated the proposed approach with the dataset that was made by the time-series coordinate and calculated features from the captured video data of 6 subjects. All the video data from 5 subjects were used as training data, and the rest of the video data from 1 subject were used to evaluate the performance of the proposed approach as testing data. Furthermore, cross-validation was performed on six different combinations (Patterns 1–6) of training and testing data subjects. Successful recognition was defined if the time-series features loaded into the trained LSTM model predicted the same action as that manually classified for the same video data. The comparison of recognition success rate was obtained for each predicted video data label, with the LSTM model

Table 2. The highest recognition accuracy rate of the cross-validation.

Pattern No.	Test data	Interpolated data (E: 81)	Uninterpolated data (E: 85)
No.1	Subj. A	92.4%	85.8%
No.2	Subj. B	89.4%	76.9%
No.3	Subj. C	97.0%	89.8%
No.4	Subj. D	94.7%	87.9%
No.5	Subj. E	89.5%	86.7%
No.6	Subj. F	95.3%	86.4%
Average Acc.		93.1%	85.6%

Table 3. The confusion metrics of precision and recall with Pattern 3.

		Predicted Class					
		W.	S.U.	S.D.	C.U.	C.D.	Recall
True Class	W.	4753	17	73	62	7	96.8%
	S.U.	0	463	87	14	0	82.1%
	S.D.	0	7	690	7	0	98.0%
	C.U.	0	0	23	1152	50	94.0%
	C.D.	0	1	1	556	706	55.9%
Precision		100.0%	94.9%	78.9%	64.3%	92.5%	

(E: Epochs; W.: Walking; S.U.: Squatting up;
S.D.: Squatting down; C.U.: Climbing up;
C.D.: Climbing down)

trained in epochs from 10 to 100 times in increments of 1.

4.2 Experiment Results

In this experiment, two performance results were obtained using the same LSTM model settings with different datasets. One result was obtained by using datasets without interpolated coordinates and time-series features calculated without interpolated joint points. In addition, another result was obtained by using datasets containing interpolated coordinates, where the time-series features were calculated from the interpolated joint points. A comparison of the two results, which have the highest average recognition accuracies and recognition accuracies for each pattern in specific epochs, are shown in Table 2.

In contrast, the recognition accuracy significantly improved when datasets were interpolated using the proposed approach. The approach using interpolation methods obtained varied recognition accuracies according to the times of epochs, with a maximum of 94.5% and a minimum of 87.7% after 20 epochs. Furthermore, the average recognition accuracy was highest in the 81st epoch,

with all subjects achieving a recognition accuracy of 89.0% or higher. In addition, when subject C was used as the test data with interpolation, the maximum recognition accuracy was 97.0%.

Furthermore, the confusion metrics with precision and recall were calculated from the highest recognition rate of video data, which is Pattern 3 in 81th epoch, are shown in Table 3. The maximum values of 100.0% and 98.0% for precision and recall were obtained using the proposed method. Therefore, it is clear that the proposed approach can obtain high action-recognition accuracy. However, it incorrectly predicted climbing up as climbing down. We considered that this action is difficult to learn because of the similarity of the climbing movements in each direction. Moreover, these action data only comprised a few limited frames, from which it was difficult to recognize the difference between two similar actions. Accordingly, the proposed approach needs further discussion with the appropriate setting of limited frames in the action data. Additionally, more kinds of actions and subjects need to be captured to test the robustness of the proposed approach.

5. Conclusions

In this paper, we proposed an action-recognition system using four interpolation methods to process human-skeleton data to train the LSTM model. Based on the experiment, it was determined that i) the proposed approach can be used to correctly recognize basic workplace actions from video data and ii) the interpolation methods can improve the performance of the proposed approach. Currently, we did not have any opportunity to obtain actual-situation data from construction site. The versatility and the usefulness of proposed approach will be examined using simulated action data of construction site in the future. In addition, the level of accuracy is required in actual situation will be considered for further work.

References

- (1) Ministry of Land, Infrastructure, Transport and Tourism :
“Status and Issues in Construction Industry”,
<https://www.mlit.go.jp/common/001187379.pdf>
(Accessed 2021/01/13)
- (2) Ministry of Health, Labour, and Welfare :
“About the elder workers”,
https://anzeninfo.mhlw.go.jp/yougo/yougo58_1.html
(Accessed 2021/01/13)
- (3) SangUk Han, Madhav Achar, SangHyun Lee and Feniosky Peña-Mora : “Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring”, *Visualization in Engineering*, Vol. 1, No. 6, 2013
- (4) Hainan Chen, Xiaowei Luo, Zhuang Zheng and Jinjing Ke : “A proactive workers' safety risk evaluation framework based on position and posture data fusion”, *Automation in Construction*, Vol. 98, pp.275–288, 2019
- (5) Hiroaki Ishioka, Xinshuo Weng, Yunze Man and Kris Kitani : “Single Camera Worker Detection, Tracking and Action Recognition in Construction Site”, *ISARC 2020*, pp.653–660, 2020
- (6) Jun Yang, Zhongke Shi and Ziyang Wu : “Vision-based action recognition of construction workers using dense trajectories”, *Advanced Engineering Informatics*, Vol. 30, No. 3, pp.327–336, 2016
- (7) Sepp Hochreiter and Jürgen Schmidhuber : “Long short-term memory”, *Neural computation*, Vol. 9, No. 8, pp.1735–1780, 1997
- (8) Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh : “Realtime Multi-Person 2D pose Estimation using Part Affinity Fields”, In *Proc. Of CVPR*, pp.7201–7299, 2017
- (9) Diederik P. Kingma and Jimmy Ba : “Adam: A Method for Stochastic Optimization”, *arXiv preprint arXiv:1412.6980*, 2014
- (10) “FFmpeg”, <https://ffmpeg.org/documentation.html>
(Accessed 2021/01/13)