A Method for Detecting Human-object Interaction based on Motion Distribution around Hand

Tatsuhiro Tsukamoto[1], Toru Abe[2][a] and Takuo Suganuma[2][b]

[1] *Graduate School of Information Sciences, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan*

[2] *Cyberscience Center, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan*

Keywords: Human-object Interaction Detection, Human Skeleton, Forearm Movement, Motion around Hand.

Abstract: Detecting human-object interaction in video images is an important issue in many computer vision applications. Among various types of human-object interaction, especially the type of interaction where a person is in the middle of moving an object with his/her hand is a key to observing several critical events such as stealing luggage and abandoning suspicious substances in public spaces. This paper proposes a novel method for detecting such type of human-object interaction. In the proposed method, an area surrounding each hand is set in input video frames, and the motion distribution in every surrounding area is analyzed. Whether or not each hand moves an object is decided by whether or not its surrounding area contains regions where movements similar to those of the hand are concentrated. Since the proposed method needs not explicitly extract object regions and recognize their relations to person regions, the effectiveness in detecting the human-object interaction, technically hands which are right in the middle of moving objects, is expected to be improved for diverse situations, e.g., several persons individually move unknown objects with their hands in a scene.

1 INTRODUCTION

Detecting human-object interaction in video images is an important issue in many computer vision applications, e.g., surveillance, human-computer interface, virtual reality, and sport image analysis. Among various types of human-object interaction, especially the type of interaction where a person is in the middle of moving an object with his/her hand is a key to observing several critical events such as stealing luggage and abandoning suspicious substances in public spaces. Most existing methods for detecting human-object interaction extract object regions from an input video image and recognize the relations of them to person regions (Le et al., 2014; Chao et al., 2015; Meng et al., 2018; Zhang et al., 2019). Consequently, when there are several persons and moving objects in a scene, it is hard for the existing methods to accurately detect the persons who move the objects with their hands, due to the difficulties in recognizing the relations of the object regions to the person regions and/or their body part regions.

We propose a novel method for detecting human-object interaction where persons are in the middle of moving objects with their hands. In our method, an area surrounding each hand is set in every input video frame, and then the motion distribution in each surrounding area is analyzed. Whether or not each hand moves an object is decided by whether or not its surrounding area contains regions where movements similar to those of the hand are concentrated. Since the proposed method needs not explicitly extract object regions and recognize their relations to person regions, the above-mentioned issues in the existing methods can be solved and the effectiveness in detecting the human-object interaction, technically hands which are right in the middle of moving objects, is expected to be improved for diverse situations, e.g., several persons individually move unknown objects with their hands in a scene.

The remainder of this paper is organized as follows: Section 2 presents the existing methods of human-object interaction detection, Section 3 explains the details of our proposed method for detecting human-object interaction based on the motion distribution around a hand, Section 4 presents the results of human-object interaction experiments on several different types of video images, and then Section 5 concludes this paper.

[a] https://orcid.org/0000-0002-3786-0122

[b] https://orcid.org/0000-0002-5798-5125

Tsukamoto, T., Abe, T. and Suganuma, T.
A Method for Detecting Human-object Interaction based on Motion Distribution around Hand.
DOI: 10.5220/0009092204620469
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 462-469
ISBN: 978-989-758-402-2; ISSN: 2184-4321

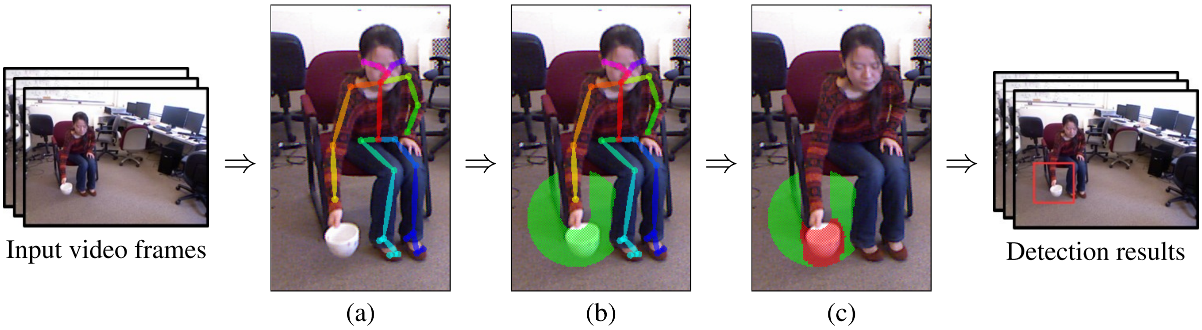


Figure 1: Overview of the proposed method. (a) extracting the skeleton of each person from every input video frame, (b) setting an individual area surrounding each visible and moving hand, (c) analyzing the motion distribution in each surrounding area. Red parts represent regions where movements similar to those of the hand are concentrated.

2 RELATED WORK

Various methods have been proposed for detecting human-object interaction in 2D images and/or depth images. Their aims vary widely, e.g., monitoring of daily activities in rooms (Meng et al., 2018), observation of customer behavior in retail environments (Liciotti et al., 2014), detection of stealing luggage (Roy and Chalavadi, 2018) and abandoning suspicious objects (Lin et al., 2015; Ghuge and Dhulekar, 2017) in public spaces, and analysis of sport images (Leo et al., 2008; Shih, 2018). To detect human-object interaction, most existing methods extract object and person regions from input images, and recognize the relations between the extracted regions. Since the region extraction of unknown objects and the relation recognition between object and person regions are still difficult, such human-object relation based methods cannot easily adapt to the situations where several persons individually move unknown objects in a scene.

Compared to those, based on the observation that important interaction between persons and objects are made mainly through their hands, several methods have been proposed for detecting a person's hand which moves an object by the states around the hand. The depth image based method in (Ubalde et al., 2014) detects such type of human-object interaction by the depth changes around a hand. The 2D image based method in (Mitsuhashi et al., 2014) detects human-object interaction by the motion (optical flow) distribution around a hand. Since these methods need not explicitly extract object regions and recognize their relations, the problems in the human-object relation based methods are expected to be solved. However, in these methods, procedures for setting the area around a hand, acquiring states from the area, and analyzing the acquired states have not been investigated adequately. Consequently, these methods cannot produce intended effects on detecting the human-object interaction in diverse situations.

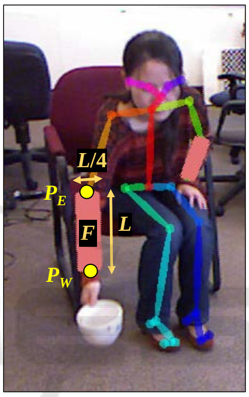


Figure 2: Forearm region F.

3 PROPOSED METHOD FOR DETECTING HUMAN-OBJECT INTERACTION

Figure 1 shows the overview of our proposed method for detecting human-object interaction. In the proposed method, (a) the skeleton of each person is extracted from every input video frame, (b) an individual area surrounding each visible and moving hand is set, and (c) the motion distribution in each surrounding area is analyzed. According to the analysis result for each hand, our method decides whether or not the hand moves an object by whether or not its surrounding area contains regions where movements similar to those of the hand are concentrated. In the following, we explain the proposed method details.

3.1 Extracting Skeleton of Each Person

In recent years, human detection technology has made great progress and it allows to extract accurately body parts from images. We obtain the skeleton of each person (a set of human body keypoints) from every input video frame (image) by OpenPose (Cao et al., 2017). Hands whose skeletons are not extracted from the image are ignored as invisible, and moreover, hands at rest are also ignored in subsequent processes.

To decide whether or not a hand moves, as shown in Figure 2, a forearm region F is set along the forearm skeleton (a line between the elbow P_E and the wrist P_W keypoints). The size of F is $L/4$ by L, where L is the length of the forearm skeleton in the image. At each pixel p_m in F, the observed movement $vo(p_m)$ is obtained as optical flow. If the average of movement norms $\|vo(p_m)\|$ in F is large, the proposed method decide that the hand moves. This condition is expressed by

$$\sum_{p_m \in F} \frac{\|vo(p_m)\|}{L \times N_F} \quad > \quad T_F, \tag{1}$$

where N_F is the number of pixels p_m in F and T_F is a given threshold. Since the values of optical flow are affected by the size of a target in the image, the movement norm $\|vo(p_m)\|$ in Eq. (1) is normalized by the forearm length L.

3.2 Setting Individual Area Surrounding Each Hand

Based on the extracted skeletons, an individual area is set for each visible and moving hand. Each area surrounds the corresponding hand and excludes the hand and the other body part region.

As shown in Figure 3 (a), for each forearm, a straight line from the elbow P_E to the wrist P_W is extended by ΔL, where the extended portion is set as $\Delta L = 0.35 \times L$ by referring to the standard body part proportion (Drillis et al., 1964). The end of the extended line is regarded as the center O of the hand. As shown in Figure 3 (b)-1, a circle centered at O with radius $R1$ is set as a neighborhood area S' of the hand. As shown in Figure 3 (b)-2, a circle with radius $R2$ is moved along the skeleton including the extended line from P_E to O, and its locus is determined as the person's body part region B. By excluding B from S', as shown in Figure 3 (c), the area S surrounding the hand is obtained. In these procedures, $R1$ and $R2$ are determined according to the forearm length L in the image by $R1 = \alpha_1 L$ and $R2 = \alpha_2 L$, where α_1 and α_2 are given parameters.

3.3 Analyzing Motion Distribution in Each Surrounding Area

The individual area S surrounding each hand excludes the hand and the other body part region B, thus if S contains regions where movements similar to those of the hand are concentrated, then these regions are highly likely to correspond with an object moved with the hand. To conduct such analysis, the proposed method observes the movement at every pixel in S, estimates the movement expected to be observed at the pixel when it corresponds with the moved object, and acquires the distribution of the normalized differences between the observed and the expected movements.

At each pixel $p_n = (x_n, y_n)$ in S, the observed movement $vo(p_n)$ and the expected movement $ve(p_n)$ are obtained as optical flow. The normalized difference $ndv(p_n)$ between $vo(p_n)$ and $ve(p_n)$ is computed by

$$ndv(p_n) \quad = \quad \frac{\|vo(p_n) - ve(p_n)\|}{\|ve(p_n)\|}. \tag{2}$$

To estimate ve, we consider two types of typical object movements.

Type 1 (ve_1) **:** As shown in Figure 4 (a), when an object is held tightly with a hand, it is not only translated but also rotated. In this case, the movement $ve_1(p_n)$ at each pixel p_n in the moved object region is modeled as shown in Figure 5, where $T = (t_x, t_y)$ and ω are the translation and the rotation components of a forearm movement, respectively. Thus, $ve_1(p_n)$ at p_n in S is represented by

$$ve_1(p_n) \quad = \quad (-\omega y_n + t_x,\ \omega x_n + t_y). \tag{3}$$

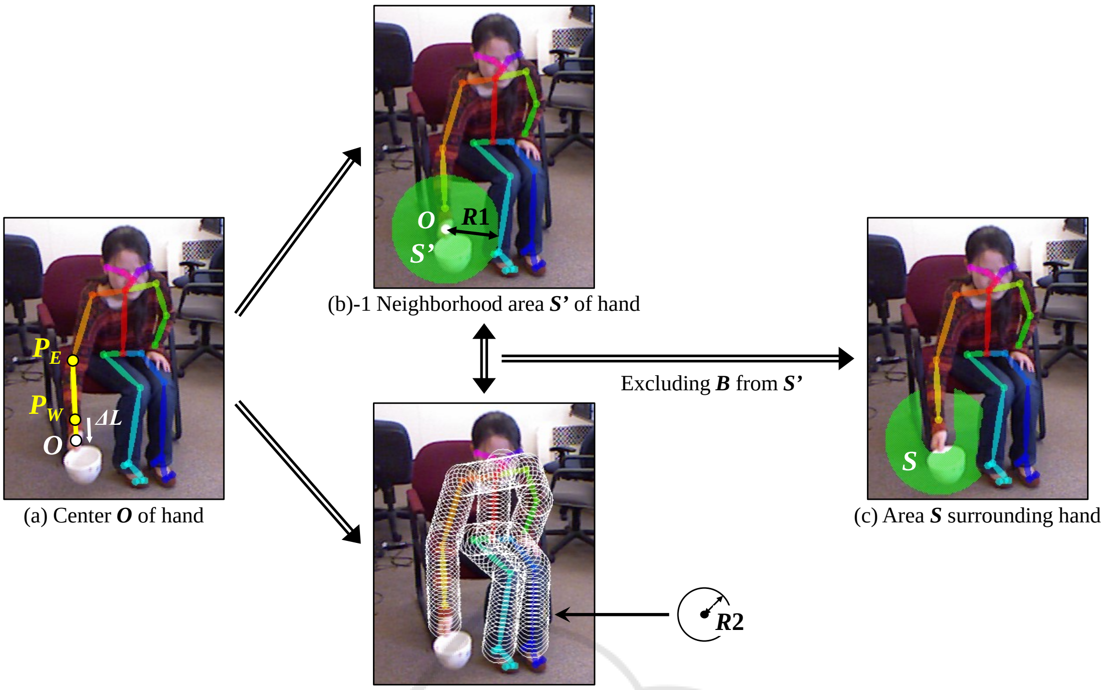
Type 2 (ve_2) **:** When an object is held loosely with a hand, as shown in Figure 4 (b), the object is almost entirely translated. In this case, the movements at all pixels in the moved object region can be assumed to be the same movement as ve_1 at the center O of the hand. Thus, $ve_2(p_n)$ at any pixel p_n in S is represented by

$$ve_2(p_n) \quad = \quad ve_1(O). \tag{4}$$

For computing ve_1 by Eq. (3), the translation component $T = (t_x, t_y)$ and the rotation component ω of a forearm movement are required. To determine these components, the optical flow $vo(p_m)$ is observed at each pixel p_m in a forearm region F, which is described in Section 3.1. The components T^* and ω^* minimizing the square sum $E^2(T,\omega)$ of the difference between $vo(p_m)$ and $ve_1(p_m)$ are computed by

$$E^2(T,\omega) \quad = \quad \sum_{p_m \in F} \|vo(p_m) - ve_1(p_m)\|^2, \tag{5}$$

$$E^{*2}(T^*,\omega^*) \quad = \quad \min_{T,\omega} E^2(T,\omega), \tag{6}$$



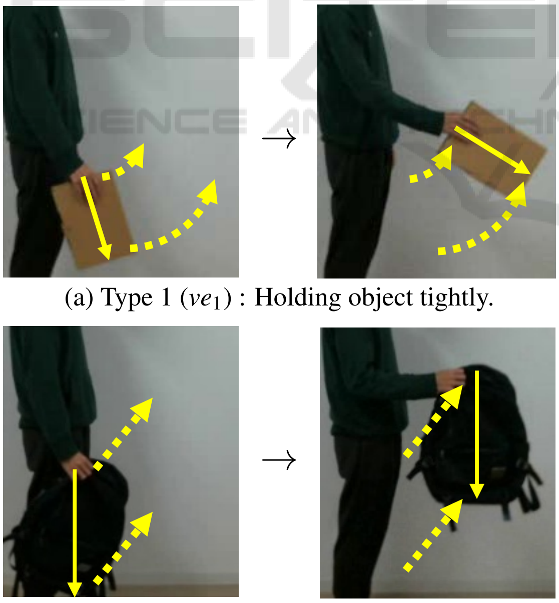
(a) Center $\boldsymbol{O}$ of hand

(b)-1 Neighborhood area $\boldsymbol{S'}$ of hand

(b)-2 Body part region $\boldsymbol{B}$

(c) Area $\boldsymbol{S}$ surrounding hand

Figure 3: An individual area surrounding each hand.



(a) Type 1 (ve_1) : Holding object tightly.

(b) Type 2 (ve_2) : Holding object loosely.

Figure 4: Two types of object movements.

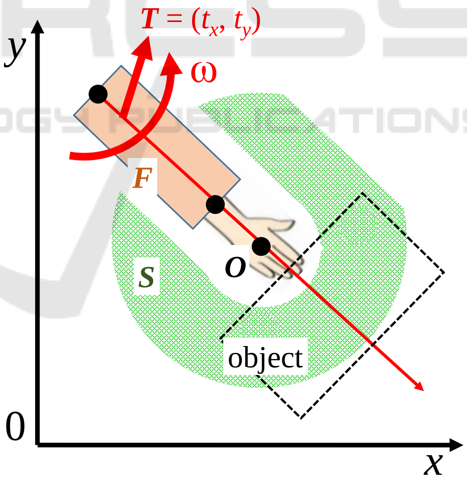


Figure 5: Object movement model (ve_1).

and used for the components T and ω in Eq. (3). An actual object movement is likely to be a combination of Type 1 and 2 movements. Consequently, for computing $ndv(p_n)$ at each p_n, an expected movement $ve(p_n)$ is expressed as a mixture of both type movements $ve_1(p_n)$ and $ve_2(p_n)$ by

$$ve(p_n) \quad = \quad \beta \times ve_1(p_n) + (1-\beta) \times ve_2(p_n), \quad (7)$$

where β ranges from 0 to 1. The value of β for each p_n is set as the one which minimizes the difference between $vo(p_m)$ and $ve(p_n)$ by

$$\min_{\beta} \|vo(p_n) - ve(p_n)\|^2. \quad (8)$$

In the distribution of $ndv(p_n)$, when $ndv(p_n)$ show low values at a sufficient number of pixels p_n, the hand is highly likely to be right in the middle of moving an object. Accordingly, to detect such type of human-object interaction, the proposed method counts the number N_{low} of pixels p_n whose $ndv(p_n)$ are lower than a given threshold T_{low} as

$$ndv(p_n) \quad < \quad T_{low}, \tag{9}$$

and decides that the hand moves an object when the following condition is met

$$N_{low} \quad > \quad \gamma \times L, \tag{10}$$

where γ is a given threshold. Eq. (10) means that the proposed method can detect human-object interaction if the area of an object moved with the hand is observed more than $\gamma \times L$ pixels in the image.

4 HUMAN-OBJECT INTERACTION DETECTION EXPERIMENTS

4.1 Experimental Environments

To illustrate the effectiveness of our proposed method, we have conducted experiments in detecting human-object interaction.

In the current implementation of our proposed method, the skeleton of each person are extracted from every input video frame by OpenPose (Cao et al., 2017), and the end keypoints (Eyes, Ears, Heels, SmallToes, BigToes) of a human body are ignored. Optical flow in images is obtained by DISOpticalFlow (Kroeger et al., 2016). The thresholds described in Sections 3.1 and 3.3 are set as $T_F = 0.02$, $T_{low} = 0.60$, and the parameters described in Section 3.2 are set as $\alpha_1 = 0.90$, $\alpha_2 = 0.37$, $\gamma = 10.0$.

As the input video images in the experiments, we use 24 video images (640×480pixels RGB image sequences), which are chosen from "picking objects" and "arranging objects" categories in CAD-120 (Koppula et al., 2013). This dataset is open to the public and widely used in human activity recognition experiments. In each video image, a person moves an object with his/her hand. Since the movement directions of hands and the sizes of objects vary among the video images, we classify the video images into four groups (a), (b), (c), and (d). As shown in Table 1, the directions of hand movements are roughly parallel to the image plane in Groups (a), (b) video images, roughly perpendicular to the image plane in Groups (c), (d) video images, the sizes of objects are relatively large in Groups (a), (c) video images, and relatively small in Groups (b), (d) video images.

Table 1: Groups of video images used in the experiments.

	movement direction to image plane	object size	# of videos (# of frames)
Group (a)	parallel	large	6 (1896)
Group (b)	parallel	small	6 (1163)
Group (c)	perpendicular	large	6 (1860)
Group (d)	perpendicular	small	6 (1321)

We determine visible hand (wrist) positions in each frame of every video image, and label manually whether or not each hand moves an object as the ground truth of human-object interaction. The detection results (visible and moving hands which hold objects) by the proposed method are compared with the ground truth, and the number of True Positives TP, False Negatives FN, and False Positives FP are calculated. Even if a hand which moves an object is detected correctly by the proposed method, the detection result is regarded as a False Positive when its forearm position (elbow and wrist keypoint positions) is extracted incorrectly. From these values, Recall R, Precision P, and F-measure F are computed by

$$R \quad = \quad \frac{TP}{TP+FN}, \tag{11}$$

$$P \quad = \quad \frac{TP}{TP+FP}, \tag{12}$$

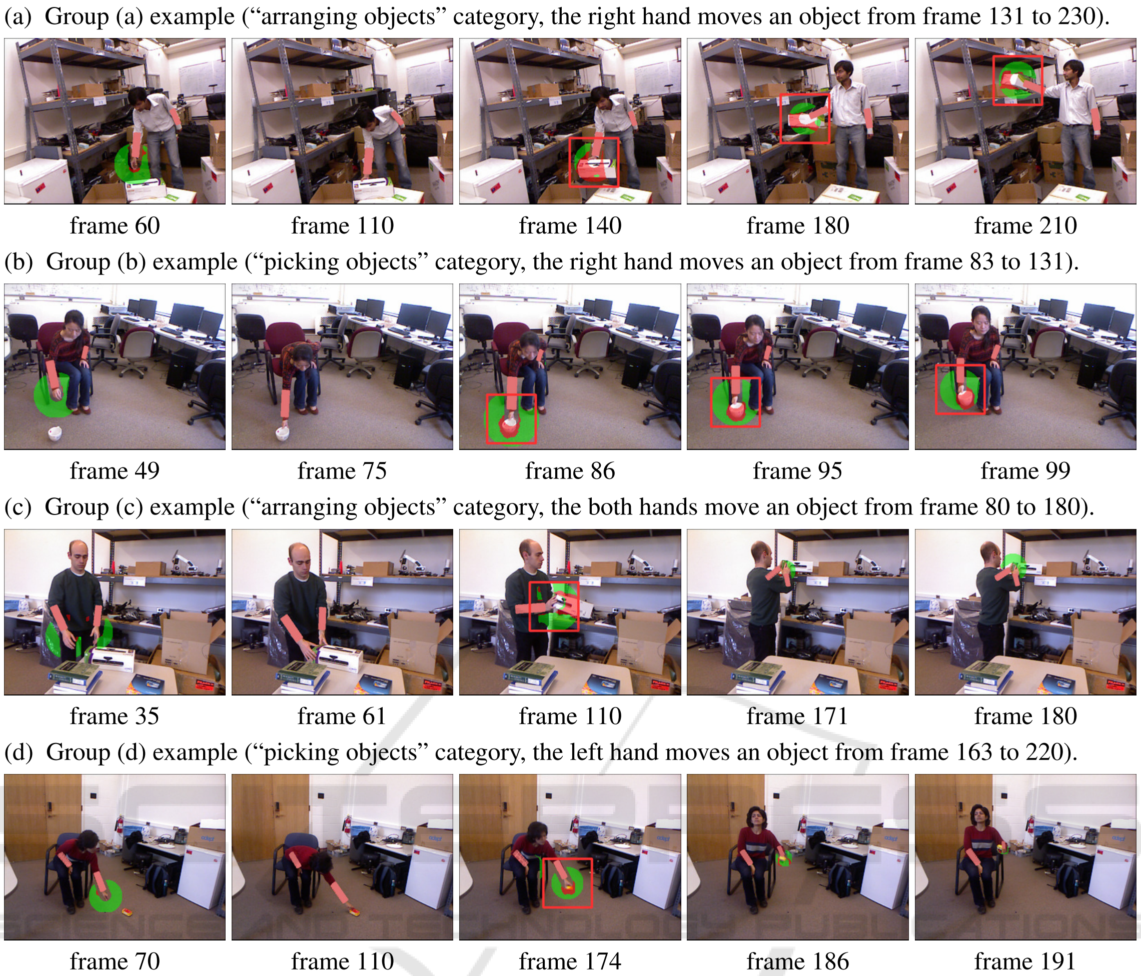
$$F \quad = \quad \frac{2 \times R \times P}{R+P}. \tag{13}$$

4.2 Experimental Results

Figure 6 shows examples of the experimental results. In those images, each flesh-colored square indicates the forearm region F set along a forearm which is extracted as visible one by OpenPose, each area composed of green and red parts indicates the area S surrounding a visible and moving hand, where green and red parts represent regions of movements dissimilar and similar to those of the hand, respectively, and each red square indicates detected human-object interaction (a hand which is in the middle of moving an object) by the proposed method.

The detection accuracy by the proposed method is listed in Table 2. In these results, False Negatives and Positives due to hand extraction failures by OpenPose are included. As can be seen from these results, the higher F-measure is attained for Groups (a), (b) (subtotal $F = 0.83$) than for Groups (c), (d) (subtotal $F = 0.54$). The higher F is attained for Group (a) ($F = 0.84$) than for Group (b) ($F = 0.81$), and the higher F is attained for Group (c) ($F = 0.57$) than for Group (d) ($F = 0.45$).

These results show that the proposed method is more affected by the movement direction of a hand



(a) Group (a) example (“arranging objects” category, the right hand moves an object from frame 131 to 230).

frame 60 frame 110 frame 140 frame 180 frame 210

(b) Group (b) example (“picking objects” category, the right hand moves an object from frame 83 to 131).

frame 49 frame 75 frame 86 frame 95 frame 99

(c) Group (c) example (“arranging objects” category, the both hands move an object from frame 80 to 180).

frame 35 frame 61 frame 110 frame 171 frame 180

(d) Group (d) example (“picking objects” category, the left hand moves an object from frame 163 to 220).

frame 70 frame 110 frame 174 frame 186 frame 191

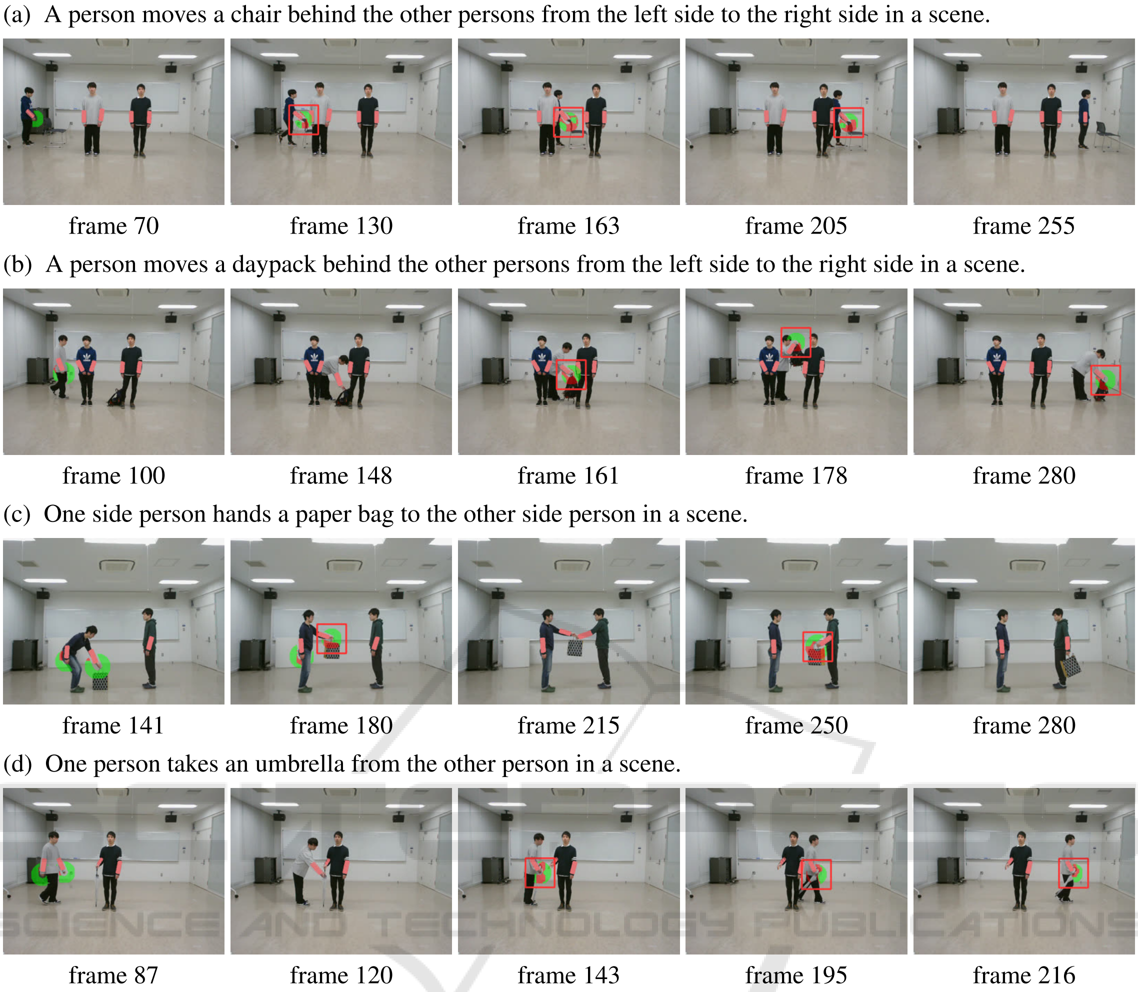
Figure 6: Examples of experimental results on each video image group.

than by the size of a moved object. The reason why the higher F-measure is attained for Groups (a), (b) than for Groups (c), (d) is because the optical flow of a movement parallel to the image plane can be obtained more accurately than that of a movement perpendicular to the image plane. The errors in estimating optical flow cause the decision errors by Eqs. (1), (9), and (10), which lead to failures in detecting human-object interaction. The examples of such detection failures can be seen in Figure 6 (c) and (d), i.e. False Negatives in frames 171, 180 of Group (c) example and frames 186, 191 of Group (d) example.

Table 2: Detection accuracy (including False Negatives and Positives due to hand extraction failures by OpenPose).

Group	TP	FN	FP	R	P	F
(a)	556	107	59	0.84	0.90	0.87
(b)	309	67	51	0.82	0.86	0.84
(a)+(b)	865	174	110	0.83	0.89	0.86
(c)	502	499	282	0.50	0.64	0.56
(d)	150	192	170	0.44	0.47	0.45
(c)+(d)	652	691	452	0.49	0.59	0.53
Total	1517	865	562	0.64	0.73	0.68

The detection accuracy excluding False Negatives and Positives due to hand extraction failures by OpenPose is listed in Table 3. From these results, the same tendency as Table 2 in F-measure can be seen. Compared to the results in Table 2, F-measure in Table 3 increases for each group, especially for Group (c) ($F = 0.57$ in Table 2 and $F = 0.64$ in Table 3). The reason for this is that the forearm skeletons cannot be extracted correctly by OpenPose in many frames in video images of Group (c). As shown in Figure 6 (c), a person moves an object roughly perpendicular to the image plane with his/her both hands in each video image of Group (c). Consequently, one forearm is likely to be occluded by the other forearm, and then the skeleton of the occluded forearm is not extracted by OpenPose. The example of such forearm extrac-



(a) A person moves a chair behind the other persons from the left side to the right side in a scene.

frame 70 frame 130 frame 163 frame 205 frame 255

(b) A person moves a daypack behind the other persons from the left side to the right side in a scene.

frame 100 frame 148 frame 161 frame 178 frame 280

(c) One side person hands a paper bag to the other side person in a scene.

frame 141 frame 180 frame 215 frame 250 frame 280

(d) One person takes an umbrella from the other person in a scene.

frame 87 frame 120 frame 143 frame 195 frame 216

Figure 7: Detection result examples of additional experiments on several different scenes.

Table 3: Detection accuracy (excluding False Negatives and Positives due to hand extraction failures by OpenPose).

Group	TP	FN	FP	R	P	F
(a)	556	76	49	0.88	0.92	0.90
(b)	309	56	51	0.85	0.86	0.85
(a)+(b)	865	132	100	0.87	0.90	0.88
(c)	502	429	154	0.54	0.77	0.63
(d)	150	166	166	0.47	0.47	0.47
(c)+(d)	652	595	320	0.52	0.67	0.59
Total	1517	727	420	0.68	0.78	0.73

tion failure can be seen in Figure 6 (c), i.e. the left forearm occluded by the right forearm is not extracted in frame 110 of Group (c) example.

These experimental results indicate that the proposed method is effective for detecting human-object interaction in the situation where a hand moves a relatively large object roughly parallel to the image plane.

4.3 Additional Experiments

We conducted additional experiments on video images of several different scenes. Figure 7 shows example frames in these video images (640×480pixels, 30fps) and detection results. In scenes (a) and (b), a person moves an object (chair or daypack) behind the other persons from the left side to the right side in the scene. In scenes (c) and (d), one person hands a paper bag to the other and takes an umbrella from the other, respectively.

As can be seen from scenes (a) and (b) in Figure 7, the proposed method detects a person hand which is in the middle of moving of an object even though the object or the person body is occluded partially (frames 130, 163, 205 of scene (a) and frames 161, 178 of scene (b)). This is because the proposed method can work if a moving forearm and a moving object are visible in a certain extent of each part (it is not necessarily that the entire regions of person body and object are visible). As can be seen from scenes (c) and (d) in Figure 7, when an object is hold by one person's hand after another, the proposed method detects such human-object interaction individually and correctly (frames 180, 250 of scene (c) and frames 143, 195 of scene (d)).

These experiment results shows the effectiveness of the proposed method in detecting human-object interaction for diverse situations.

5 CONCLUSIONS

In this paper, we have focused on the type of human-object interaction where a person is in the middle of moving an object with his/her hand, and proposed a novel method for detecting such type of human-object interaction by the motion distribution in an individual area surrounding each hand. Since our method needs not explicitly extract object regions from input images and recognize their correspondence to person regions, the effectiveness in detecting the human-object interaction is expected to be improved for diverse situations. Through the experiments on human activity video images, we confirmed the effectiveness of our proposed method in the situations where a person is right in the middle of moving a relatively large object roughly parallel to the image plane.

We will conduct further experiments on a variety of environments such as the different angles of cameras, the various types of objects, the different numbers of persons, and the diverse conditions of occlusion areas. Currently, our proposed method achieves several decision processes as thresholding procedures by Eqs. (1), (9), and (10). We would like to investigate approaches for achieving these processes as machine learning based procedures.

In future work, we plan to extend our proposed method to multiple camera environment. This is because, we can expect to deal with the decrease in interaction detection accuracy from unsuitable image condition by the following approach: several images of the same person are taken from different angles, unsuitable condition images, where his/her hand is hard to detect, overlaps considerably with other body part regions, or moves roughly perpendicular to the image plane, are excluded from the taken images, and human-object interaction is detected by using the remaining images.

REFERENCES

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conf. Comput. Vision Pattern Recognit.*, pages 1302–1310.

Chao, Y.-W., Wang, Z., He, Y., Wang, J., and Deng, J. (2015). HICO: A benchmark for recognizing human-object interactions in images. In *Int. Conf. Comput. Vision*, pages 1017–1025.

Drillis, R., Contini, R., and Bluestein, M. (1964). Body segment parameters: A survey of measurement techniques. *Artif. Limbs*, 8(1):44–66.

Ghuge, N. and Dhulekar, P. (2017). Abandoned object detection. *Int. J. Mod. Trends Sci.ence Technol.*, 3(6):215–218.

Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *Int. J. Rob. Res.*, 32(8):951–970.

Kroeger, T., Timofte, R., Dai, D., and Gool, L. V. (2016). Fast optical flow using dense inverse search. In *Eur. Conf. Comput. Vision*, pages 471–488.

Le, D.-T., Uijlings, J., and Bernardi, R. (2014). TUHOI: Trento universal human object interaction dataset. In *The 3rd Workshop Vision Lang.*, pages 17–24.

Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P. L., D'Orazio, T., and Distante, A. (2008). Real-time multiview analysis of soccer matches for understanding interactions between ball and players. In *Int. Conf. Content-Based Image Video Retrieval*, pages 525–534.

Liciotti, D., Contigiani, M., Frontoni, E., Mancini, A., Zingaretti, P., and Placidi, V. (2014). Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network. In *Int. Workshop Video Anal. Audience Meas. Retail Digital Signage*, pages 146–157.

Lin, K., Chen, S.-C., Chen, C.-S., Lin, D.-T., and Hung, Y.-P. (2015). Abandoned object detection via temporal consistency modeling and back-tracing verification for visual surveillance. *IEEE Trans. Inf. Forensics Security*, 10(7):1359–1370.

Meng, M., Drira, H., and Boonaert, J. (2018). Distances evolution analysis for online and off-line human object interaction recognition. *Image Vision Comput.*, 70:32–45.

Mitsuhashi, Y., Abe, T., and Suganuma, T. (2014). A detection method of human-object interactions in crowded environment based on hierarchical image analysis. *Tech. Rep. IEICE*, 114(PRMU2014-77):69–74.

Roy, D. and Chalavadi, K. M. (2018). Snatch theft detection in unconstrained surveillance videos using action attribute modelling. *Pattern Recognit. Lett.*, 108:56–61.

Shih, H.-C. (2018). A survey of content-aware video analysis for sports. *IEEE Trans. Circuits Syst. Video Technol.*, 28(5):1212–1231.

Ubalde, S., Liu, Z., and Mejail, M. (2014). Detecting subtle human-object interactions using Kinect. In *19th Iberoam. Congr. Pattern Recognit.*, pages 770–777.

Zhang, H.-B., Zhang, Y.-X., Zhong, B., Lei, Q., Yang, L., Du, J.-X., and Chen, D.-S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5):1005.