A Method for Detecting Hands Moving Objects from Videos

Rikuto Konishi[1], Toru Abe[2,1][a] and Takuo Suganuma[2,1][b]
[1]*Graduate School of Information Sciences, Tohoku University, Aoba-ku, Sendai 980-8578, Japan*
[2]*Cyberscience Center, Tohoku University, Aoba-ku, Sendai 980-8578, Japan*
rikuto.konishi.t6@dc.tohoku.ac.jp, {beto, suganuma}@tohoku.ac.jp

Keywords: Human Activity, Hand-Object Interaction, Skeleton Information, Motion Information.

Abstract: In this paper, we propose a novel method to recognize human actions of moving objects with their hands from video. Hand-object interaction plays a central role in human-object interaction, and the action of moving an object with the hand is also important as a reliable clue that a person is touching and affecting the object. To detect such specific actions, it is expected that detection model training and model-based detection can be made more efficient by using features designed to appropriately integrate different types of information obtained from the video. The proposed method focuses on the knowledge that an object moved by a hand shows movements similar to those of the forearm. Using this knowledge, our method integrates skeleton and motion information of the person obtained from the video to evaluate the difference in movement between the forearm region and the surrounding region of the hand, and detects the hand moving an object by determining whether the similar movements as the forearm occur around the hand from these differences.

1 INTRODUCTION

Human-object interaction recognition from video is a fundamental issue in many computer vision applications, including security, VR, and human-machine interface (Antoun and Asmar, 2023; Wang et al., 2023). There are many different types of human-object interaction, but one of the most significant is human actions of moving objects with their hands. Hand-object interaction plays a central role in human-object interaction (Kim et al., 2019; Fan et al., 2022), and moving an object with the hand is also important as a reliable clue that a person is touching and affecting the object.

Existing approaches for human-object interaction recognition can be roughly divided into the two-stage approach and the one-stage approach (Antoun and Asmar, 2023; Luo et al., 2023). Currently, the one-stage approach, which simultaneously performs the person-object pair association and interaction recognition according to the features acquired from the video, is widely used due to its efficiency.

Generally, there are two approaches to determine features for human-object interaction recognition: learning-based approach and handcrafted approach (Zhu et al., 2016; Sargano et al., 2017). The learning-based approach, which implicitly determines features from samples through neural network-based machine learning and makes recognition models, can be applied to various recognition targets. For this reason, most current methods for human-object interaction recognition determine features and make recognition models based on this approach. However, when using different types of information obtained from the video, a multi-stream framework is used that processes each type of information in a different stream and integrates the results. Therefore, for using not only the image information itself but also different types of information, such as a person's skeleton and motion information, in order to effectively recognize human actions, the neural network becomes large and complex, which increases the resources required for processing (Haroon et al., 2022; Shafizadegan et al., 2024). In contrast, the handcrafted approach that explicitly designs features based on knowledge of the recognition target is limited in its applicability. However, for specific target actions, such as the action of moving an object with the hand, it is easy to acquire knowledge about the target action. Based on the acquired knowledge, by designing features that appropriately integrate different types of information and processing those features in a single stream, it will be possible to efficiently recognize the target action.

In this paper, we propose a novel method to detect human actions of moving objects with their hands

[a] https://orcid.org/0000-0002-3786-0122
[b] https://orcid.org/0000-0002-5798-5125

Konishi, R., Abe, T. and Suganuma, T.
A Method for Detecting Hands Moving Objects from Videos.
DOI: 10.5220/0013167800003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 392-399
ISBN: 978-989-758-728-3; ISSN: 2184-4321

from videos. The proposed method adopts the one-stage approach, and uses the features designed by the handcrafted approach for detecting the target action. This method focuses on the knowledge that an object moved by a hand shows movements similar to those of the forearm. Based on this knowledge, our method integrates the skeleton and the motion information of the person obtained from the video to evaluate the difference in movement between the forearm region and the surrounding region of the hand, and then detects the hand moving an object by determining whether the similar movements as the forearm occur around the hand from these differences. In our method, the skeleton and the motion information obtained from the video are integrated more effectively than the existing method as the difference in movement between the forearm region and the surrounding region of the hand. By using these differences as features for target detection, it is possible to perform processing on a single stream even when using neural network-based machine learning, and it is expected that the hand moving an object can be detected efficiently.

2 RELATED WORK

2.1 Human Action Recognition Using Learning-Based Features

Many methods have been proposed to recognize various types of human actions, including human-object interaction and hand-object interaction. Most of the recent methods are based on the one-stage approach, and use features determined by the learning-based approach (Zhu et al., 2016; Sargano et al., 2017).

Recently, several methods have been proposed that use different types of information obtained from the video to achieve effective human action recognition. In the method of (Simonyan and Zisserman, 2014), image and motion information obtained from the video are input into different CNNs to extract features, and actions are recognized by integrating the outputs from the two streams. In the method proposed of (Haroon et al., 2022), sequences of image and person's skeleton are processed using different LSTMs, and recognition is performed by integrating their outputs. The methods in (Gu et al., 2020; Khaire and Kumar, 2022) use sequences of image, skeleton, and depth information for human action recognition, but each type of information is processed by a separate neural network, and recognition is performed by integrating the results processed by the different streams.

As described above, when utilizing different types of information obtained from the video by learning-based approach, each type of information is processed in a different stream, which makes the neural network configuration large and complex, and increases the resources required for processing.

2.2 Human Action Recognition Using Handcrafted Features

Based on the observation that important interaction between persons and objects are made mainly by their hands, several methods have been proposed for detecting a person's hand which moves an object by designing handcrafted features from the surrounding states of the hand and using them.

In the method of (Tsukamoto et al., 2020), based on the knowledge that when an object is moved by the hand, similar movements occur around the hand as with the forearm, features that integrate skeleton and motion information is designed, and the hand moving an object is detected by these features without extracting the object region. This makes it possible to efficiently use different types of information obtained from the video. However, there are problems with this method, such as some forearm movements (movements toward or away from the camera) are not considered, and the features that integrate skeleton and motion information being unable to express the state of movement around the hand in detail.

3 PROPOSED METHOD

An overview of the proposed method is shown in Figure 1. The processing flow of the proposed method is similar to that of the existing methods in (Tsukamoto et al., 2020). First, (a) skeleton information of each person is extracted from every frame image of the input video. Based on the extracted skeleton information, (b) a region FR is determined for each forearm and the forearm motion i modeled in it. Using the forearm motion model, (c) the differences between the movements expected to occur when an object is moved by the hand and the movements actually observed from the video are evaluated in the surrounding region SR of the hand. According to the movement differences, (d) the hand moving an object is detected.

3.1 Modeling Forearm Motion

Skeleton (a set of keypoints) of each person is extracted for every frame image, and a region FR is determined for each forearm using the skeleton. The motion of the forearm at a pixel $p = (x, y)$ in the image

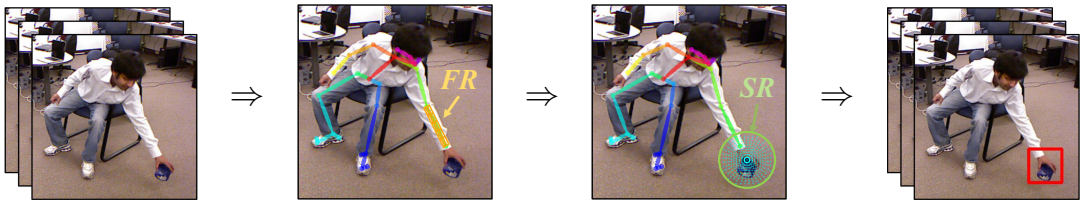


Figure 1: An overview of the proposed action detection method: (a) extracting skeleton information from the input video, (b) modeling the forearm motion in the forearm region FR, (c) evaluating the difference in movement between the forearm region and the hand surrounding region SR, and (d) detecting the hand moving an object.

is modeled as $v_{FR}(p)$, and the parameters in the forearm motion model are determined to minimize dif_{of}^2, which is the sum of squares of the difference between the optical flow $of(p)$ at p actually observed from the video and $v_{FR}(p)$ in FR, defined by Eq. (1)

$$dif_{of}^2 = \sum_{p \in FR} \|of(p) - v_{FR}(p)\|^2. \quad (1)$$

Several methods have been developed to extract human skeleton information as a set of keypoints on the human body (Cao et al., 2021; Fang et al., 2023). In the proposed method, skeleton information of each person is extracted by applying one of these methods. As shown in Figure 2 (a), using the extracted keypoint positions of the elbow P_E and wrist P_W, the point where $P_E - P_W$ is extended by $\Delta L = \alpha \times L$ toward the tip of the forearm is set as the center $O = (x_O, y_O)$ of the hand. Here, L represents the distance between $P_E - P_W$. The forearm region FR is determined as a rectangular region of size $l \times w$ along $P_E - O$, where its length l and width w are set to $l = L + \Delta L = (1+\alpha)L$ and $w = \beta L$, respectively.

In the existing method of (Tsukamoto et al., 2020), it is assumed that the forearm moves in the image with a rotational component ω and a translational component (t_x, t_y), and the movement $v_{FR}(p)$ at a pixel $p = (x, y)$ in FR is modeled by Eq. (2)

$$v_{FR}(p) = (-\omega y + t_x,\ \omega x + t_y), \quad (2)$$

where ω, t_x, and t_y are determined to minimize dif_{of}^2. Because the forearm motion is modeled only by rotation and translation in the image, when the forearm moves back and forth relative to the camera, this method cannot accurately represent its movements. On the other hand, in the proposed method, $v_{FR}(p)$ is modeled by Eq. (3) based on affine transformation

$$v_{FR}(p) = (c_1 x + c_2 y + c_3,\ c_4 x + c_5 y + c_6), \quad (3)$$

where $c_1, c_2, \ldots, c_6$ are determined to minimize dif_{of}^2. As a result, the proposed method is able to approximately represent the motion of the forearm not only when it rotates or translates in the image, but also when it moves back and forth relative to the camera.

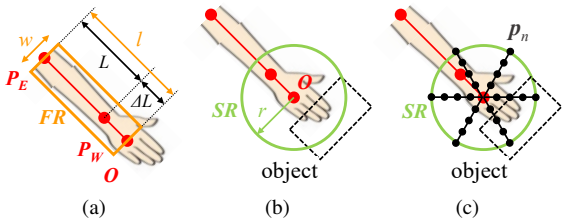


Figure 2: (a) Forearm region FR, (b) hand surrounding region SR, and (c) sampled pixels p_n in FR.

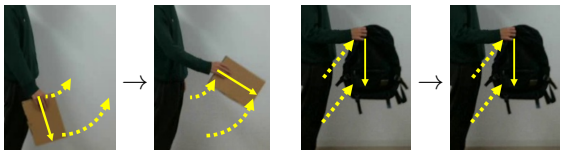
3.2 Difference in Movement Between Forearm and Around Hand

When an object is moved by a hand, the object shows the movements similar to those of the forearm. Consequently, if there are areas around the hand that show movements similar to those of the forearm, it can be determined that these areas are highly likely to correspond to the object moved by the hand, even without object recognition. From this, the degree to which a pixel p in the region SR set surrounding the hand does not correspond to an object moved with the hand can be evaluated by the difference between the movement $v_{eo}(p)$ expected to occur on the object moved by the hand and the movement (optical flow) $of(p)$ actually observed from the video.

In the proposed method, as shown in Figure 2 (b), the surrounding region SR is determined for each hand as a circle with radius r centered at the center O of the hand, where r is set as $r = \gamma L$ from $L = P_E - P_W$. The difference at p in SR between the expected movement on an object moved by the hand and the optical flow actually observed from the video is evaluated as $ndv(p)$, normalized by $v_{eo}(p)$ using Eq. (4) to reduce the effect of the hand movement speed

$$ndv(p) = \|v_{eo}(p) - of(p)\| \ / \ \|v_{eo}(p)\|. \quad (4)$$

As shown in Figure 3 (a), when an rigid object is tightly held and moved by the hand, the object moves as an extension of the forearm in the same way as other parts of the forearm, and the movement $v_{eo}(p)$



(a) object hold tightly (b) object hold loosely

Figure 3: Movements of an object moved by a hand.

on the object is expected to be represented by Eq. (5)

$$v_{eo}(p) = v_{FR}(p). \quad (5)$$

On the other hand, as shown in Figure 3 (b), when an object is loosely held and moved by the hand, the object translates and the movement $v_{eo}(p)$ on the object shows the same movement at the center O of the hand, regardless of the position of p, and is expected to be represented by Eq. (6)

$$v_{eo}(p) = v_{FR}(O). \quad (6)$$

These two situations are extreme examples, and in reality, a mixture of both is thought to occur on an object moved by the hand. Accordingly, in the proposed method, $v_{eo}(p)$ is computed by Eq. (7)

$$v_{eo}(p) = \eta \cdot v_{FR}(p) + (1-\eta) \cdot v_{FR}(O). \quad (7)$$

where η is set to minimize $\|v_{eo}(p) - of(p)\|$ for p.

3.3 Features for Action Detection

In the existing method of (Tsukamoto et al., 2020), the number N_{eo} of pixels p in SR for which the normalized difference $ndv(p)$ is less than a threshold T_{eo} is counted, and the ratio N_{eo}/N_{SR} of these pixels to the total number N_{SR} of pixels in SR is determined. If N_{eo}/N_{SR} is greater than a threshold T_{SR}, the existing method determines that movements similar to those of the forearm occur at many places in SR and that the hand moves an object. Since this method represents the state of movements in SR by a single index N_{eo}/N_{SR} and performs detection by heuristic thresholding on that index, it is difficult to detect based on the detailed state in SR.

On the other hand, in the proposed method, as shown in Figure 2 (c), SR is equally divided in the circumferential direction and the radial direction, respectively, and sampled pixels p_n are determined. A feature vector FV_{SR} is constructed from $ndv(p_n)$ computed at all p_n, and it is determined whether an object moved by the hand is in SR, i.e., whether the hand moves an object, by applying a machine learning-based classifier to FV_{SR}. In this way, the proposed method uses a feature vector constructed by ndv, which integrates skeleton information and motion information obtained from the video, to detect the target actions. This will enable processing in a single stream, even when using neural network-based machine learning, and is expected to enable more efficient detection of target actions. Besides, by using this feature vector, the proposed method is able to perform detection based on more detailed states in SR than the existing method.

4 EXPERIMENTS

To evaluate the effectiveness of our proposed method, we conducted experiments to detect hands moving objects in videos.

4.1 Experiment Environments

In the experiments, we used videos from the Cornell Activities Dataset (CAD-120) (Koppula et al., 2013), a publicly available dataset for human daily activity recognition experiments. This dataset consists of a total of 124 videos across 10 activity categories (picking objects, arranging objects, unstacking objects, taking food, stacking objects, microwaving food, taking medicine, cleaning objects, having meal, and making cereal). Each activity category contains 12 videos of four subjects, as each subject's similar activity was captured three times (only "making cereal" category contains 16 videos of four subjects, as each subject's similar activity was captured four times).

For each frame image of every video, human body keypoints were detected by applying OpenPose (Cao et al., 2021), and forearms with detection confidence of elbow and wrist keypoints greater than 0.5 were extracted as visible hands. Each extracted hand was visually inspected by referring to the next frame image to see if it was holding and moving an object, and was manually labeled as being the hand moving an object or not. These labeling results were used as ground truth for evaluating the detection experiment results.

Table 1 shows the total number of videos for each activity category, the total number of frame images, the cumulative number of extracted hands, and the cumulative number of extracted hands moving objects.

4.2 Experiment Methods

Detection methods were applied to each visible hand extracted in each frame image of every video, and it was determined whether the hand was moving an object or not. The results were compared to the ground truth, and the number of True Positives (TP, detected "hand moving object") TP, False Positives (FP, detected "hand not moving object") FP, and False Negatives (FN, undetected "hand moving object") FN

Table 1: Videos used in the experiments: total number of videos (frame images) and cumulative number of extracted hands (hands moving objects) for each activity category.

Activity Category	# of videos (frame images)		# of ext. hands (moving obj.)	
picking objects	12	(2501)	4818	(669)
arranging objects	12	(3781)	6629	(1413)
unstacking objects	12	(5586)	10986	(2751)
taking food	12	(5614)	8616	(2129)
stacking objects	12	(5813)	11472	(2972)
microwaving food	12	(6350)	9946	(2870)
taking medicine	12	(6394)	12887	(3461)
cleaning objects	12	(7406)	11539	(3799)
having meal	12	(9829)	19066	(4933)
making cereal	16	(11647)	22500	(7711)
Total	124	(64921)	118459	(32708)

were counted. From these, Precision P, Recall R, and F1 score $F1$ were computed for evaluation.

The experiments were conducted using the following four methods:

- the existing method (Tsukamoto et al., 2020) with the translation / rotation forearm motion model and the heuristic thresholding based classifier,
- the method using only the affine transformation forearm motion model,
- the method using only the feature vector based classifier,
- our proposed method with the affine transformation forearm motion model and the feature vector based classifier.

The forearm region FR and the hand surrounding region SR are set in the same way for all methods. For the length L of each forearm (the distance from the elbow keypoint P_E to the wrist keypoint P_W), based on the average body shape (Drillis et al., 1964), the center O of the hand is set along $P_E - P_W$ at a distance $\Delta L = 0.35 \times L$ from P_W. A rectangle is set as FR along $P_E - O$, and its length l and width w are set to $l = L + \Delta L = 1.35 \times L$ and $w = 0.25 \times L$, respectively. A circle is set as SR with radius $r = 1.1 \times L$ centered at O. Also, the optical flow in the t th frame is computed using $t-1$ th and t th frames.

For the existing method, the movement of a forearm is modeled by translation and rotation in the image using Eq. (2). For each forearm, the number N_{eo} of pixels for which ndv is less than a threshold T_{eo} is counted, and the ratio of N_{eo} to the total number N_{SR} of pixels in SR is computed. If N_{eo}/N_{SR} is greater than a threshold T_{SR}, it is determined that the hand is moving an object. The thresholds were set to $T_{eo} = 0.55$ and $T_{SR} = 0.15$ based on preliminary experiments.

For the proposed method, the movement of a forearm is modeled by affine transformation using Eq. (3). The feature vector FV_{SR} (36×10 dimensions) is constructed from ndv computed where SR is sampled at 36 locations along the circumference and 10 locations in the radial direction. Whether the forearm moves an object or not is determined by classifying FV_{SR} using SVM. When detecting hands moving object in each frame image of every video, SVM model is trained using feature vectors and ground truth labels obtained from each frame image of all 123 videos other than the target video.

4.3 Experiment Results

The results of the experiments to detect hands moving objects are listed in Table 2, and examples of detection results by the proposed method are shown in Figure 4, where red, green, and blue squares indicate TP, FP, and FN, respectively.

The following can be confirmed from the results:

- Compared to using the existing method, when using the method with the affine transformation forearm motion model, although FP increased for all activity categories, TP increased even more for most categories. As a result, P and especially R improved, and $F1$ improved for most categories.
- By using the method with the feature vector based classifier, compared to using the existing method, TP increased for all categories, and FP decreased for many categories, resulting in substantial improvements in P, R, and $F1$ for most categories.
- By using the proposed method with the affine transformation forearm motion model and the feature vector based classifier, TP increased further for most categories, and FP decreased further for more categories. The P, R, and $F1$ results improved significantly for most categories.

The reason TP increased with the introduction of the affine transformation forearm motion model is thought to be that it became possible to better capture forearm motion other than translation / rotation with respect to the image plane. On the other hand, the reason for the increase in FP is thought to be that optical flow estimation errors were more often regarded as forearm motion. By introducing the feature vector based classifier, which can perform detailed classification, it was able to increase TP while suppressing the increase in FP. Furthermore, by introducing the affine transformation forearm motion model simultaneously with the feature vector based classifier, as in the proposed method, it became possible to further increase TP and decrease FP.

Table 2: The results of the experiments to detect hands moving objects.

Activity category	Method	Affine transformation motion model	Feature vector based classifier	TP	FP	FN	P	R	$F1$
picking objects	proposed	✓	✓	591	828	78	0.42	0.88	**0.57**
		-	✓	**599**	930	**70**	0.39	**0.90**	0.55
		✓	-	461	570	208	**0.45**	0.69	0.54
	existing	-	-	425	**543**	244	0.44	0.64	0.52
arranging objects	proposed	✓	✓	**1183**	581	**230**	0.67	**0.84**	**0.74**
		-	✓	1107	672	306	0.62	0.78	0.69
		✓	-	1056	487	357	0.68	0.75	0.71
	existing	-	-	963	**398**	450	**0.71**	0.68	0.69
unstacking objects	proposed	✓	✓	**2225**	**383**	**526**	**0.85**	**0.81**	**0.83**
		-	✓	1982	518	769	0.79	0.72	0.75
		✓	-	1579	950	1172	0.62	0.57	0.60
	existing	-	-	1195	829	1556	0.59	0.43	0.50
taking food	proposed	✓	✓	**835**	**208**	**1294**	**0.80**	**0.39**	**0.53**
		-	✓	711	276	1418	0.72	0.33	0.46
		✓	-	302	461	1827	0.40	0.14	0.21
	existing	-	-	229	404	1900	0.36	0.11	0.17
stacking objects	proposed	✓	✓	**2512**	**488**	**460**	**0.84**	**0.85**	**0.84**
		-	✓	2235	619	737	0.78	0.75	0.77
		✓	-	1891	1033	1081	0.65	0.64	0.64
	existing	-	-	1538	923	1434	0.62	0.52	0.57
microwaving food	proposed	✓	✓	**1609**	**329**	**1261**	**0.83**	**0.56**	**0.67**
		-	✓	1399	490	1471	0.74	0.49	0.59
		✓	-	1111	732	1759	0.60	0.39	0.47
	existing	-	-	845	611	2025	0.58	0.29	0.39
taking medicine	proposed	✓	✓	**1923**	**626**	**1538**	**0.75**	**0.56**	**0.64**
		-	✓	1742	840	1719	0.67	0.50	0.58
		✓	-	1561	785	1900	0.67	0.45	0.54
	existing	-	-	1252	659	2209	0.66	0.36	0.47
cleaning objects	proposed	✓	✓	**1506**	**271**	**2293**	**0.85**	**0.40**	**0.54**
		-	✓	1266	330	2533	0.79	0.33	0.47
		✓	-	588	446	3211	0.57	0.15	0.24
	existing	-	-	410	345	3389	0.54	0.11	0.18
having meal	proposed	✓	✓	3408	386	1525	**0.90**	0.69	0.78
		-	✓	**3629**	438	**1304**	0.89	**0.74**	**0.81**
		✓	-	2608	447	2325	0.85	0.53	0.65
	existing	-	-	2618	**358**	2315	0.88	0.53	0.66
making cereal	proposed	✓	✓	**5265**	**1537**	**2446**	**0.77**	**0.68**	**0.73**
		-	✓	5015	1762	2696	0.74	0.65	0.69
		✓	-	4542	1833	3169	0.71	0.59	0.64
	existing	-	-	4035	1716	3676	0.70	0.52	0.60
Total	proposed	✓	✓	**21057**	**5637**	**11651**	**0.79**	**0.64**	**0.71**
		-	✓	19685	6875	13023	0.74	0.60	0.66
		✓	-	15699	7744	17009	0.67	0.48	0.56
	existing	-	-	13510	6786	19198	0.67	0.41	0.51

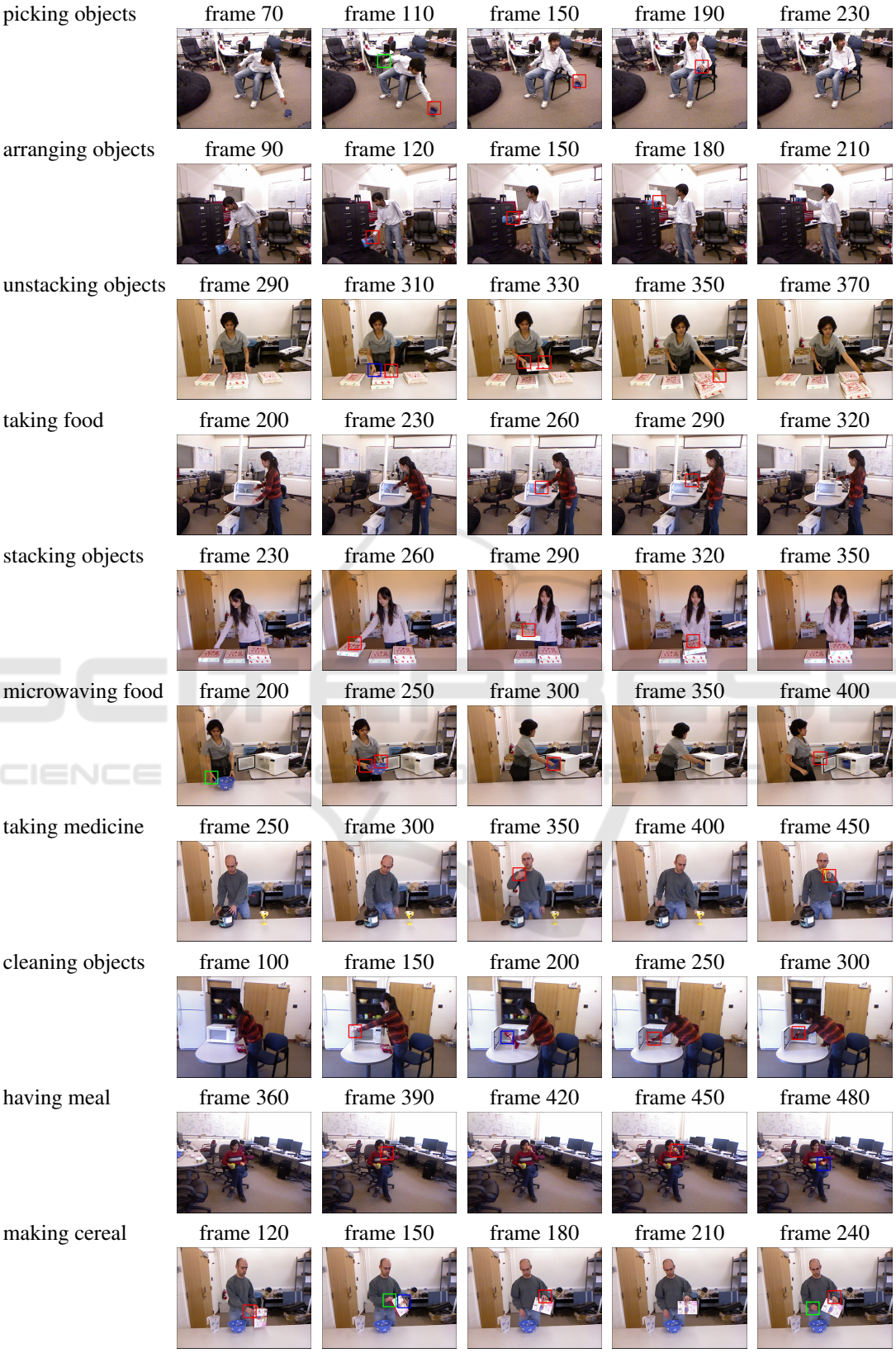


Figure 4: Examples of detection results by the proposed method (□ TP, □ FP, □ FN).

These experiment results show the effectiveness of the proposed method in detecting hands moving objects. However, $F1$ was still low for some activity categories. For "picking objects," large optical flows were often observed in areas other than an object moved by a hand, and an increase in FP caused a decrease in P, resulting in a decrease in $F1$. For "taking food" and "cleaning objects," forearm movements were often slow or small, and a decrease in TP caused a decrease in R, which in turn led to a decrease in $F1$. Therefore, our future task is to improve the proposed method so that it can handle such situations.

5 CONCLUSION

In this paper, we focused on the action of people moving objects with their hands, and proposed a method to detect hands moving objects from video.

The proposed method integrates skeleton and motion information obtained from video into a single type of features by using prior knowledge about the detection target, and performs detection processing based on those features. Since this approach performs detection based on a single type of features, it is expected to improve the efficiency of the necessary processing, including training the detection model.

Compared to the existing method based on a similar approach, our method deals with various hand movements by introducing the affine transformation forearm motion model, and discriminates hand states in detail by introducing the feature vector based classifier. Through the experiments on the video dataset of human daily activities, we demonstrated that the proposed method can improve the accuracy of detecting hands moving objects from video (compared to existing method, $F1$ improved from 0.51 to 0.71).

As future work, we plan to:

- implement the proposed method using more powerful classifier, such as deep learning based classifier, instead of the current SVM based classifier,
- conduct comparative experiments with methods that process different information, such as skeleton and motion information, in separate streams.

REFERENCES

Antoun, M. and Asmar, D. (2023). Human object interaction detection: Design and survey. *Image Vision Comput.*, 130:104617.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):172–186.

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

Fan, H., Zhuo, T., Yu, X., Yang, Y., and Kankanhalli, M. (2022). Understanding atomic hand-object interaction with human intention. *IEEE Trans. Circuits Syst. Video Technol.*, 32(1):275–285.

Fang, H.-S., Li, J., Tang, H., Xu, C., Haoyi Zhu and, Y. X., Li, Y.-L., and Lu, C. (2023). Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):7157–7173.

Gu, Y., Ye, X., Sheng, W., Ou, Y., and Li, Y. (2020). Multiple stream deep learning model for human action recognition. *Image Vision Comput.*, 93:103818.

Haroon, U., Ullah, A., Hussain, T., Ullah, W., Sajjad, M., Muhammad, K., Lee, M. Y., and Baik, S. W. (2022). A multi-stream sequence learning framework for human interaction recognition. *IEEE Trans. Human-Mach. Syst.*, 52(3):435–444.

Khaire, P. and Kumar, P. (2022). Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey. *J. Visual Commun. Image Represent.*, 86:103531.

Kim, S., Yun, K., Park, J., and Choi, J. Y. (2019). Skeleton-based action recognition of people handling objects. In *Proc. IEEE Winter Conf. Appl. Comput. Vision*, pages 61–70.

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.

Luo, T., Guan, S., Yang, R., and Smith, J. (2023). From detection to understanding: A survey on representation learning for human-object interaction. *Neurocomput.*, 543:126243.

Sargano, A. B., Angelov, P., and Habib, Z. (2017). A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Appl. Sci.*, 7(1):110.

Shafizadegan, F., Naghsh-Nilchi, A. R., and Shabaninia, E. (2024). Multimodal vision-based human action recognition using deep learning: A review. *Artif. Intell. Rev.*, 57(178):85.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Proc. Neural Inf. Process. Syst. Conf.*, volume 27, pages 1–11.

Tsukamoto, T., Abe, T., and Suganuma, T. (2020). A method for detecting human-object interaction based on motion distribution around hand. In *Proc. 15th Int Joint Conf. Comput. Vision, Imaging Comput. Graphics Theory Appl.*, volume 5, pages 462–469.

Wang, J., Shuai, H.-H., Li, Y.-H., and Cheng, W.-H. (2023). Human-object interaction detection: An overview. *IEEE Consum. Electron. Mag.*, pages 1–14.

Zhu, F., Shao, L., Xie, J., and Fanga, Y. (2016). From handcrafted to learned representations for human action recognition: A survey. *Image Vision Comput.*, 55:42–52.