ACTION RECOGNITION BASED ON 3D OBJECT DETECTION AND NORMALIZED POSE ESTIMATION

Satsuki Maeda, Bismark Kweku Asiedu Asante, and Hiroki Imamura

Graduate School of Science and Engineering, Information System Science, Soka University

ABSTRACT

Action recognition has many practical applications, but the task still faces significant challenges. A major challenge is the variation in human action poses across different viewpoints, which complicates determining the ideal pose for action recognition. To address these viewpoint-invariant issues, we propose a pose normalization approach combined with object-based action recognition to classify actions in videos. In this method, the normalized pose is compared with a reference pose to identify the action being performed. The objective of this research is to develop a three-dimensional (3D) object-associated action recognition framework that leverages the stereo camera's ability to capture accurate distance information. This approach offers three main advantages: (1) action recognition that incorporates object context, (2) resolving occlusion problems, and (3) improving recognition accuracy through precise distance information. Experimental results show that our proposed approach achieves 70% classification accuracy across ten selected action categories, independent of viewpoint or camera angle.

KEYWORDS

Object Recognition, Behavior Recognition, AI, Stereo Camera, Active Detection

1. INTRODUCTION

In recent years, Artificial Intelligence (AI) has been considered in various fields, with significant research and development focused on its integration into image recognition. Among them, behavior recognition has attracted attention. Human action recognition has not only been studied to understand human behaviors for safety and surveillance applications but is also important for a number of other applications such as video retrieval [1,2], [3], human–robot interaction [3], sports [4], and entertainment [5].

The behavior recognition technology Actlyzer [6] Fig.1, developed by Fujitsu Ltd., is one of the well-known behavior recognition systems. Actlyzer captures the characteristic points of body joints from an estimated skeleton, enabling the recognition of approximately 100 basic actions, such as ”bending an arm” or ”stretching a knee”. Furthermore, by combining these basic actions, it is possible to recognize more complex human behaviors. As a practical example, Actlyzer has been implemented in a surveillance camera in front of a door, as shown. From the camera's footage system, the system can detect a person crouching in front of the door and peering into the door hole, recognizing this as suspicious behavior. Such technology can be used not only for crime prevention but also for various other purposes.

David C. Wyld et al. (Eds): DPPR, CNDC, AIAA, DSA, WeST, ACITY, IoTE – 2025
pp. 01-12, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.152401

However, existing behaviors/action recognition approaches that utilize human skeletal information have been developed primarily based on a large amount of 2D image data. Consequently, extensive image data are required for re-training to account for variations in the position and orientation of a person captured by the camera. Furthermore, determining the position of objects and whether a person is interacting with or holding an object is inherently challenging when relying solely on conventional 2D image-based approaches. By normalizing the position and orientation of human pose data and object data obtained from 3D information, we aim to achieve behavior recognition that is robust to positional and orientational variations. In particular, we hypothesize that by considering not only human

(a) Standing in front of the door (b) Sitting in front of the door

(c) Looking at the doorknob (d) Touching the doorknob

Fig.1: Four stages in the Actlyzer suspicious person detection process. The suspicious person depicted through various stages from standing, sitting, looking and touching a door knob.

behavior but also object behavior, it is possible to achieve more accurate human behavior recognition. Many actions can look similar based purely on pose for example reaching vs pushing), but the object (its location, affordance, size, etc.) and how it is associated with body parts (hands, grasp, etc.) helps disambiguate. Knowing which object(s) is involved helps narrow down where and when in the video the action occurs, supports detecting active object interactions vs mere presence. Therefore, in this study, we propose a behavior recognition method that is robust to position and orientation and determines behaviors based on object associations with specific critical poses. our main contributions in this research work are summarized as follows:

1. develop a normalized pose which is invariant to the viewpoints

2. pose and object-based association for action recognition.

1.2. Related Works

Action recognition is gaining some much popularity in diverse fields due to its versatility in distinguish activities that are being carried by people and animals [7]. One of the reason for this popularity Every human action, no matter how trivial, is done for some purpose [8]. For example, in order to prepare meal, a chef is interacting with and responding to the environment using his/her fingers for grasping utensils, hands and arms lifting utensils,legs for moving to and fro in the kitchen, etc.

The growth trend of action recognition using human pose or skeletons have been phenomenal and the influences that depth sensing have also provides cannot be ignored. Earlier developments focusing on handcrafted features in controlled lab environments using approaches such as silhouettes as motion templates, optical flows and 2D joint estimation using geometric model. This approaches are sensitive to background clutter, lighting and occlusion. The next phase saw emergence of depth sensing for action recognition techniques such as 3D skeleton extraction, skeleton-based action recognition via hand crafted features such as joint angles and velocities, and template matching with the hidden Markov models (HMM) [9] for temporal modeling.This also suffers from view dependencies and required fixed setups as well. Even though the era of deep learning brought a lot of excitements and successes using CNN and RNN for spatial feature extraction and temporal dynamics respectively. Large computation demands and performance drops in unconstrained environments makes it difficult to attained high precisions in recognition. Combining SpatialTemporal Deep Networks and Transformers [10] are recent approaches to tackle issues such as cross-view, cross-dataset generalization.

There are several data modalities for performing action recognition: RGB color image [11], 3D Skeleton (Human Pose) [12], Depth data [13], Infrared Sequence [14], Pointcloud [15], Event Stream [16], Audio [17], Acceleration [18], Radar [19] and WiFi [19]. Difference in modalities can be attributed various sensing devices and the data captured. This data provide information from which features that can be extracted, features then classify or identify certain actions being performed. Different modalities provide advantages and disadvantages based on the sensing mode and the feature extraction, classification techniques.

Before, deep learning approaches offer end-to-end solution to recognition and detections tasks.

In our research, we focus on the 3D human pose and the objects that are being used in the performance of the action. We believe that normalized pose with object-association can provide rich information for classifying actions into various categories. In Fig. 2 we illustrated the steps in obtaining the dictionary information of the reference pose for various poses considered for the actions.

2. THE PROPOSED METHOD

2.1. Problem in the Conventional Method

In action recognition, changes in camera viewpoint can significantly alter the apparent shape of a detected human pose, even when the underlying action remains the same. Consider an observed person performing an action, where an initial image I_1 captures the person at position (x_1,y_1,z_1). A subsequent image I_2 is taken after the camera has moved and rotated by an angle θ relative to the original viewpoint. Due to this viewpoint change, the projected human pose in I_2 may differ

substantially from that in I_1, despite both depicting the same action. The view-invariant action recognition task seeks to develop representations or transformations of pose features that remain consistent across different viewpoints, thereby enabling robust action recognition regardless of camera position or orientation.

2.2. Our Approach

In order to solve the problem described in 2.1, we propose a viewpoint invariant approach where the trajectory of human pose for an action is normalized to a key skeleton pose (front view facing the camera) and an association with an object to determine the action being performed by a person. The proposed method is divided into two sections, where the first part creates a dictionary data based on the 3D skeletal data on the recognized person in action and the object of reference. The object of reference is determined by how they are attached or closed to the person in action. For the second part, the determination of the behavioral data based on the dictionary information created.

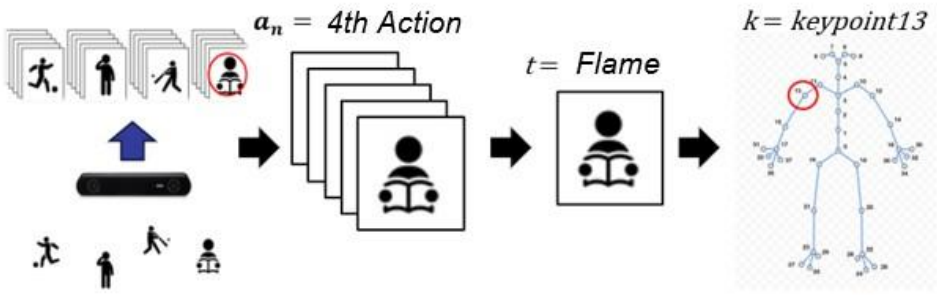

Fig.2: Illustrated data processing pipeline for obtaining the skeleton keypoints and making the reference keypoints for the action recognition

Data Processing For our action recognition, we process the data for ten (10) action categories with reference human pose. We obtain the trajectory of 3D joint coordinates using a The human skeletal coordinate acquisition method uses the Body Tracking module in the ZED SDK (Software Development Kit)Fig.3. It consists of 38 joint coordinate points and the bones connecting them. The trajectory of 3D joint coordinate points is acquired as data for 100 frames, and recorded to a file for each joint coordinate point. The recorded data is represented by frame t, human skeletal coordinates k, and action data a, where $a_1 = action1, a_2 = action2, ..., a_n = action$ where n is the number of actions being considered ($n = 0,1,2,...,n$) . Let $\mathbf{p}$ be known or reference pose the data to be prepared in advance given by:

$$\mathbf{p}(a_n,k,t) = (x_p(a_n,k,t), y_p(a_n,k,t), z_p(a_n,k,t)) \quad (1)$$

and i be the actual data, and acquire during the action recognition as follows:

$$\mathbf{i}(a_n,k,t) = (x_i(a_n,k,t), y_i(a_n,k,t), z_i(a_n,k,t)) \quad (2)$$

Using (1), we obtained the reference data which is the dictionary data. The behavioral data is captured during the action recognition and it represented in (2).

Fig.3: ZED camera

Pose Normalization To ensure that the human pose is detection is not affected by the viewpoint or the angular movement of the camera. The pose normalization requires both position normalization and orientation normalization. First, position normalization is achieved by replacing the chest coordinate points with local coordinates based on the reference point. Next, orientation normalization is performed by calculating the cross product. This allows us to obtain a normal vector that specifies the orientation of the human body, and then calculate a rotation matrix to align it with the reference vector and it is illustrated in Fig. 4

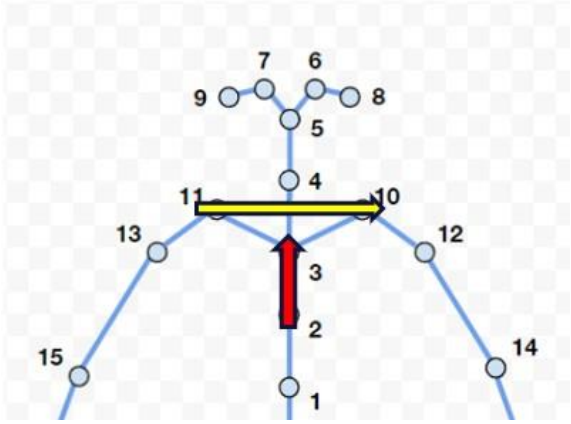

Fig.4: An illustration of the shoulder keypoints used in the cross vector rotation for the pose normalization

$$\mathbf{v}(a_n,t) = (x(a_n,k_{11},t),y(a_n,k_{11},t),z(a_n,k_{11},t)) - (x(a_n,k_{10},t),y(a_n,k_{10},t),z(a_n,k_{10},t)) \quad (3)$$

$$\mathbf{w}(a_n,t) = (x(a_n,k_2,t),y(a_n,k_2,t),z(a_n,k_2,t)) - (x(a_n,k_3,t),y(a_n,k_3,t),z(a_n,k_3,t)) \quad (4)$$

The actual vectors obtained are shown in Fig 2. Let $\mathbf{v}_n$ be the vector from the left clavicle to the right clavicle which is obtained in (3), and $\mathbf{w}_n$ be the vector from the lower spine to the upper spine in (4). The calculation is based on the difference between each coordinate points. The rotation matrix R is calculated using the obtained normal vector and the reference vector $\mathbf{c}_0$ where $\mathbf{c}_0 = (0,0,1)$. Normalization is performed by applying the calculated rotation matrix R to the data. The rotation matrix $\mathbf{R}$ is calculated using the obtained normal vector and the reference vector $\mathbf{c}_0$ where $\mathbf{c}_0 = (0,0,1)$. Normalization is performed by applying the calculated rotation matrix R to the data. Given a reference

$$\mathbf{c}_0 = \mathbf{R} \cdot \mathbf{c}(a_n,k,t) \quad (5)$$

$$\mathbf{c}(a_n,t) = \mathbf{v}(a_n,t) \times \mathbf{w}(a_n,t) \quad (6)$$

$$\mathbf{p}'(a_n,k,t) = \mathbf{R} \cdot \mathbf{p}(a_n,k,t) \tag{7}$$

$$\mathbf{i}'(a_n,k,t) = \mathbf{R} \cdot \mathbf{i}(a_n,k,t) \tag{8}$$

Object Recognition Object recognition is important for our action recognition for differentiating actions with similar poses, knowledge of the objects offers information about the spatial and temporal locality of the actions being performed. For the detection of objects being considered for our action classification, we trained a YOLOv8 [20] model for detecting objects that are involved in the actions we do want to detect. The objects that are associated with the actions we are trying to recognize. The object present in the action being provide vital information for association with actions being performed. Therefore an accurate recognition of objects will improve the recognition of the actions being performed. The closes of the center of the detected objects to the pose point of attachment is determine from the center coordinates of the detected object to be considered as the object involved in the action.

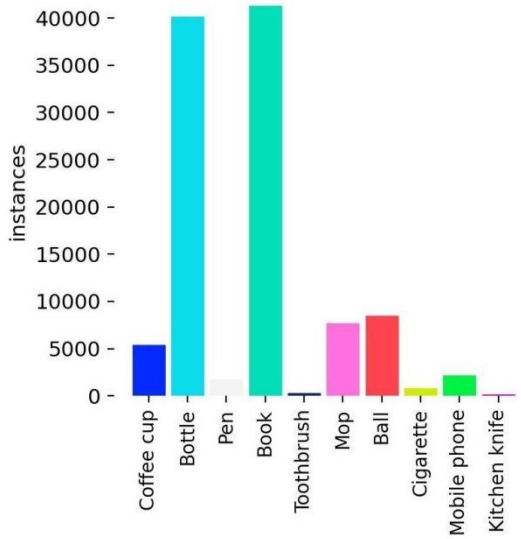

Fig.5: A bar graph representation of the data with objects instance

To determine object retention, the determine threshold euclidean distance th_1 is the between object's center coordinates and the closest skeleton coordinate point **k** to the object are used. Let the object's center coordinate point be **o**. The eight (8) vertex of the 3D bounding box are denoted as $\mathbf{b}_m = (x_m,y_m,z_m)$, where m represents each vertex m = 1,2,...,8 . These are obtained by applying the following calculation formula.

$$\mathbf{k}(a_n,t) = (x_n(a_n,t),y_n(a_n,t),z_n(a_n,t)) \tag{9}$$

The threshold euclidean distance th_1 is obtained by;

$$th_1 = \|\mathbf{b}_m - \mathbf{o}\| \tag{10}$$

For the object retention to be considered in the calculated distance between the object center coordinates and the closest skeleton coordinate point should be less than the threshold euclidean distance Fig.6.

$$\|\mathbf{o}(a_n,t) - \mathbf{k}(a_n,t)\| < th_1 \tag{11}$$

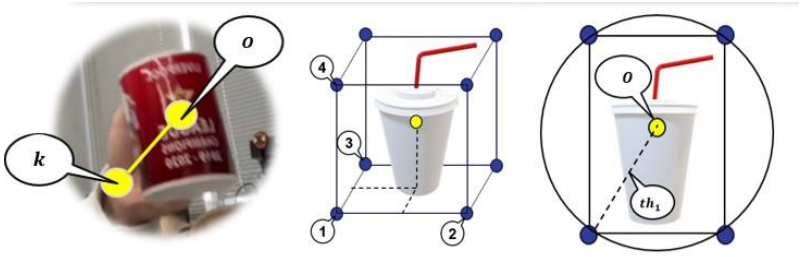

Fig.6: Object Holding Calculation Method

Action Recognition This sub section describes the action recognition process. The action recognition approach minimizes error distance between the reference pose and the normalized pose of the action being performed. To determine, the minimal error distance between the reference pose and action being performed, the error in classifying the action $E(a_n)$ is giving by:

$$E(a_n) = \sum_{k=0}^{M} \sum_{t=0}^{100} ||\mathbf{i}'(a_n,k,t) - \overline{\mathbf{p}'(a_n,k,t)}|| \tag{12}$$

where t is the frame, k is the keypoint, a_n is the *nth* action, $\mathbf{i}$ is the normalized posed

$'$ data and the $\mathbf{p}$ is the normalized . averaged

prior data.

$$argminE(a_n)(a_n \epsilon A) \tag{13}$$

With the *argminE* determined, the actions class with least error is considered the action that is being performed.

3. EXPERIMENTS

3.1. Datasets

This section describes the datasets used to evaluate action recognition methods involving specific objects. To assess posture- and object-based action detection methods, two datasets were utilized. The posture dataset is a custom dataset collected from five subjects performing ten specific actions. Human skeleton coordinates were captured using a stereo camera, and a pose estimation neural network was employed to extract the poses. The ten actions were selected with an emphasis on daily activities, sports, and security related behaviors. For the object recognition dataset, we selected objects associated with these ten actions and constructed a corresponding dataset from the Open Images Dataset (OID) [21]. This dataset comprises approximately 30,000 images drawn from OID, which were used to train the object detection model. For action recognition, ten (10) short clips of the actions was recorded to recognize the actions being performed.

3.2. Experimental Setup

The experimental setup for this research is illustrated in Fig. 8. The setup is performed in two sets. Step 1 involves the reference data (dictionary data) which is using the stereo camera to obtain the

front view data of all the actions selected in the ten (10) categories. The second step is performed by obtaining the data from different angles such as the sides of the subjects while the same action is being performed. The steps is illustrated in Fig. 5 where the step one which is the collation of the dictionary data or information is

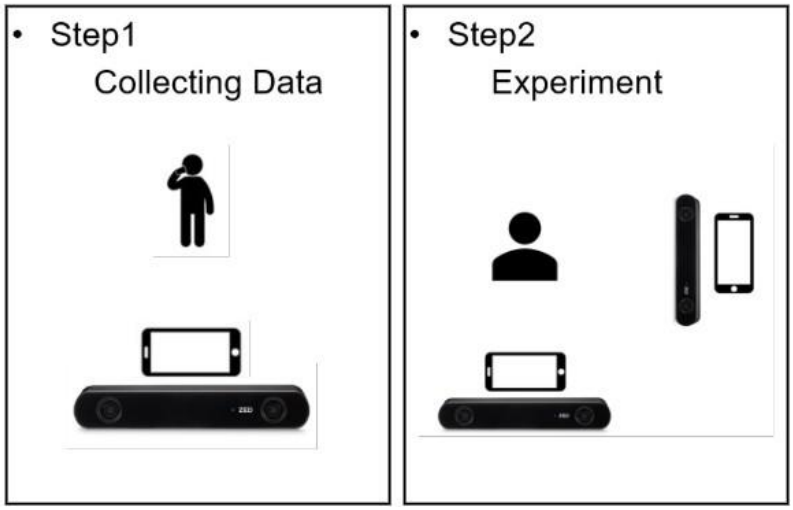

Fig.7: Experimental Setup

captured and second step is capturing data from different angles or views other than the one use in the dictionary data. Using our approach, we perform poses normalization and object recognition to obtain the key-pose with its association object. To determine which action being performed by comparing them with the reference key-pose and select the best category using the minimal error distance. For the action classification, we use precision, accuracy, recall and F1-score as the metrics.

In another setup, we also evaluated the performance of the approach and its robustness to occlusion. In the setup as shown in Fig 8. We considered the drinking action for the robustness test. There are four scenarios that were considered to determine the robustness of the approach to occlusion in scenario 1 the part of the body or human pose that is not involves in the action performance. So we occluded the arm that is not involved in the action. In scenario 2, the object is partially occluded with the arm that is not involved in the action. For scenario 3, we occluded the arm involved in the action only while scenario for we occluded the arm involved in action and then partially occluding the object.

4. RESULTS AND DISCUSSIONS

From our experimentation, we obtained the following results which we presented in three folds; the results on our novel pose normalization strategy, the results on the custom training of our YOLOv8 models and the classification of the actions using the combined pose and object detection approach.

4.1. Pose Normalization

The result of the pose normalization process is presented in the figures below. We demonstrate that with a given image Fig. 9 with a detected pose. Fig. 10 the detected pose is presented before the transformation of the pose into the front view in Fig. 11.

First, position normalization is achieved by replacing the chest coordinate points with local coordinates based on the chest. The normalization process transform any viewpoints to the reference viewpoint making it easier to compare the keypoints for a particular or relevant pose for the action being performed.

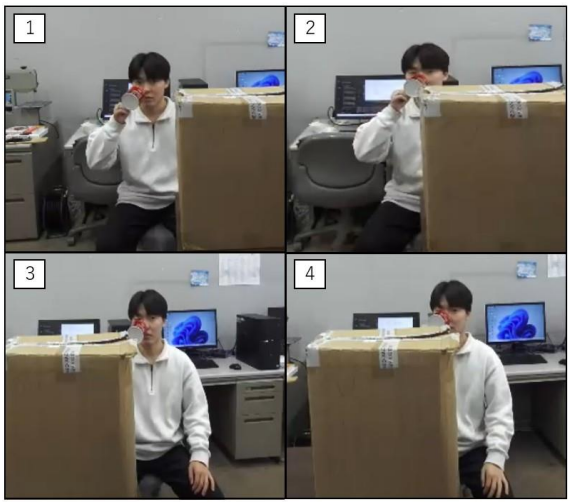

Fig.8: Occlusion Test Setup. The images demonstrates four scenarios. In (1) the occlusion is on the left side and covering the hand but not the associated object

Fig.9: A captured frame showing a human posed detected in an image with the human skeleton keypoints

4.2. Object Detection

We presented the results of our customized YOLOv8 model which we trained in this section. Fig 12 shows the performance of the model to classify objects in the dataset we curated for training the model.

The results shows the model was well trained to detect the objects however there was an imbalance in the data result in a few of the object not well classified. Mobile phone and knife

were mostly misclassified. In order to fix this, we intend to acquire more data of this objects to ensure a better classification in the future

4.3. Action Recognition

Table Fig. 13 reports the precision, recall and f1-score for the ten (10) action categories we selected. Fig 14 show that most of the classifications were correctly predicted with a

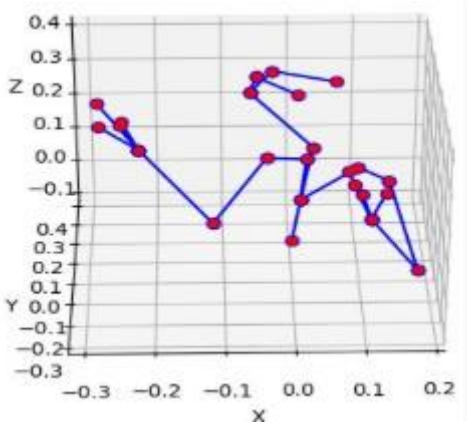

Fig.10: An illustrated keypoints extracted from 9 for a better presented visualization.

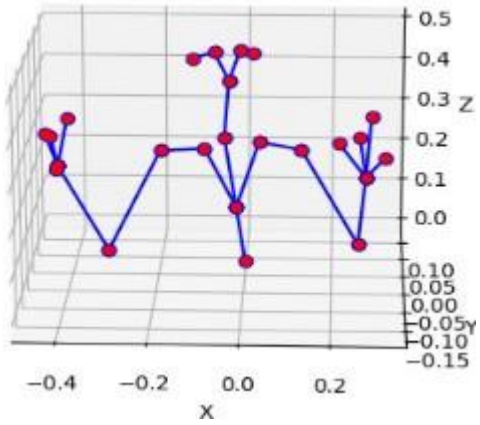

Fig.11: An illustrated normalized keypoints being visualized

few that were misclassified. Drinking, opening bottle and reading were misclassified.

The accuracies for the actions shows that our approach for detecting the following six (6) actions brushing teeth, cleaning floor, opening bottle, throwing, walking with knife, and walking with smartphone. However, the drinking action and reading action are misclassified. There is also partly misclassified opening bottle and drinking.

The results demonstrated that the approach could be used for human pose with objectbased action recognition or complement the deep learning based approach to effectively recognize various actions that involves object association. We will compare these approach with existing state of the art method for action recognition as well.

Table 1: Prediction results for four *Drinking* test scenarios

Sample	True Label	Predicted Label	Error Distance
Scenario 1	Drinking	Drinking	912.64
Scenario 2	Drinking	Drinking	960.41
Scenario 3	Drinking	BrushingTeeth	935.84
Scenario 4	Drinking	BrushingTeeth	1166.46

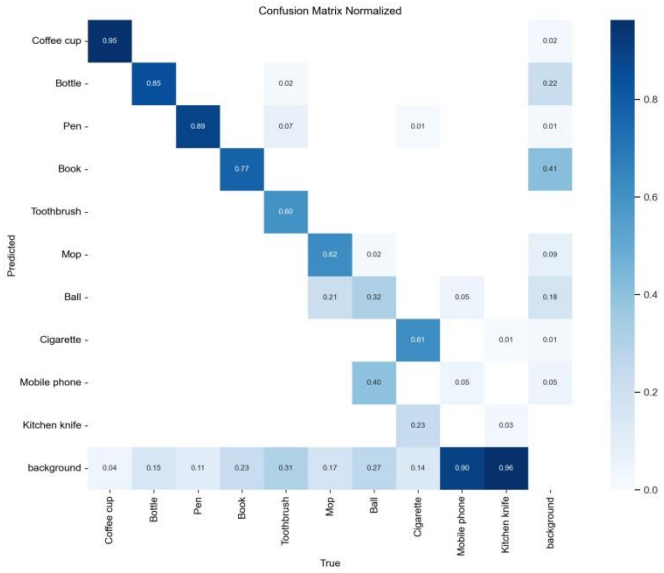

Fig.12: The confusion matrix for the object detection which are associated with the actions we want to recognize

Actions	precision	recall	f1-score	support
BrushingTeeth	1.00	1.00	1.00	2.00
CleaningFloor	1.00	1.00	1.00	2.00
Drinking	0	0	0	2.00
OpeningBottle	1.00	0.5	0.67	2.00
Reading	0	0	0	2.00
Smoking	0.5	0.5	0.5	2.00
Throwing	1.00	1.00	1.00	2.00
WalkingWithKnife	1.00	1.00	1.00	2.00
WalkingWithSmartphone	1.00	1.00	1.00	2.00
Writing	0.33	1.00	0.5	2.00
accuracy	0.7	0.7	0.7	0.7
macro avg	0.68	0.7	0.68	20.00
weighted avg	0.68	0.7	0.68	20.00

Fig.13: Evaluations of the performance our approach to classify 10 categories action classes

On the test for the occlusion, Table 1 presents the results of the action recognition that was performed with four scenarios to demonstrate the occlusion. In the scenarios 1, the drinking action was recognized correctly with the object showing with one arm hidden, Also with scenario 2 the object is also partially covered with one arm was correctly recognized. The remaining two

scenarios 3 and 4 which involves covering the arm performing the action and the covering the object partially. With the action performing covered the approach

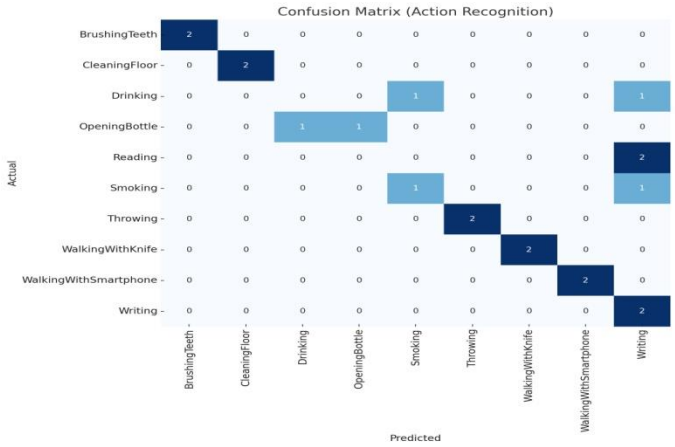

Fig.14: The confusion matrix for the classification of ten (10) actions categories failed to recognized which indicates the importance of the part of the posed involved in the action.

Table 2: Results of the occlusion test performed with the drinking action labels

Action	Precision	Recall	F1-Score
Brushing Teeth	0.00	0.00	0.00
Drinking	1.00	0.50	0.67

Table 2 presents the results in terms of precision, recall, and F1-score for the occlusion test.

The accuracy under occlusion was 50% across the four scenarios. The reported precision for the drinking action is 1.00, whereas that for brushing teeth—which was misclassified—is 0. The approach is not affected when occlusions occur on body parts not directly involved in performing the action, or when the object is only partially occluded. However, it misclassifies the drinking action as brushing teeth when such occlusions occur. Pose normalization is affected only when the occluded body parts are directly involved in the action being recognized.

5. CONCLUSION

In this research, we proposed action recognition approach with object based and 3D information. The approach relates the object information and the normalized pose information to detect the action being performed. Our approach shows robustness in the viewpoint invariant challenges and well adapted to detecting the ten (10) action categories that we selected for experimenting. Our results shows an accuracy of 70% in classifying the action classes. In the future, we plan to integrate it with a deep learning method to make it an end-to-end deep learning approach.

REFERENCES

[1] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: recent updates, datasets, challenges, and applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021.

[2] M. Ramezani and F. Yaghmaee, "A review on human action analysis in videos for retrieval applications," *Artificial Intelligence Review*, vol. 46, pp. 485–514, 2016.

[3] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2702–2706.

[4] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*. Springer, 2015, pp. 181–208.

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[6] S. TAKAHASHI, T. SAITOU, Y. TAKAHASHI, and Y. IKAI, "Technology for recognizing complex human behavior from video and its recognition rule," *JSAI Technical Report, Type 2 SIG*, vol. 2022, no. SWO-056, p. 11, 2022.

[7] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *2008 IEEE international symposium on circuits and systems (ISCAS)*. IEEE, 2008, pp. 2737–2740.

[8] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[9] T. Lu, L. Peng, and S. Miao, "Human action recognition of hidden markov model based on depth information," in *2016 15th International Symposium on Parallel and Distributed Computing (ISPDC)*. IEEE, 2016, pp. 354–357.

[10] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563. 12. *VSCC '17: Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*. New York, NY, USA: Association for Computing Machinery, 2017.

[12] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 1147–1153.

[13] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "Infar dataset: Infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36–47, 2016.

[14] H. Cheng and S. M. Chung, "Orthogonal moment-based descriptors for pose shape query on 3d point cloud patches," *Pattern Recognition*, vol. 52, pp. 397–409, 2016.

[15] E. Calabrese, G. Taverni, C. Awai Easthope, S. Skriabine, F. Corradi, L. Longinotti, K. Eng, and T. Delbruck, "Dhp19: Dynamic vision sensor 3d human pose dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[16] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE workshop on applications of computer vision (WACV)*. IEEE, 2013, pp. 53–60.

[17] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[18] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st annual international conference on mobile computing and networking*, 2015, pp. 65–76.

[19] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https:

[20] //github.com/ultralytics/ultralytics

[21] Y. Niitani, T. Akiba, T. Kerola, T. Ogawa, S. Sano, and S. Suzuki, "Sampling techniques for largescale object detection from sparsely annotated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6510–6518.

AUTHORS

S. Maeda received his B.S. degree from Soka University, Japan, where he is currently pursuing his Master's course in Information Systems Engineering. His research interests include image information processing, such as action recognition and object detection.

B.K. Asiedu Asante received his B.S. degree in Computer Science and Physics in 2012, he furthered to obtain Master of Philosophy (MPhil) in Computer Science in 2017 from the University of Ghana. In 2024, he earned his doctoral degree in Information System Science Engineering from Soka University. Subsequently, in the same year, he assumed the position of Assistant Professor at Soka University, specifically in the Information System Science Engineering department. His research interests lie in artificial intelligence and deep learning, with a focus on applying these technologies to address human and environmental challenges.

H. Imamura received the B.S. degree in engineering from Soka University, Japan, in 1997 and the M.S. degree in information science from JAIST, Japan, in 1999 and the Ph.D. degree in information science from JAIST, Japan in 2003. From 2003 to 2009, he was an Assistant Professor at Nagasaki University, Japan. From 2009 to 2020, he was an Associate Professor at Soka University, Japan. From 2020, he has been a Professor at Soka University, Japan. His research Interest includes image processing, artificial intelligence, and XR.