Hierarchical Model for Zero-shot Activity Recognition using Wearable Sensors

Mohammad Al-Naser[1,2,*], Hiroki Ohashi[3,*], Sheraz Ahmed[1], Katsuyuki Nakamura[4], Takayuki Akiyama[4], Takuto Sato[4], Phong Nguyen[4] and Andreas Dengel[1,2]

[1] *German Research Center for Artificial Intelligence (DFKI), Germany*

[2] *University of Kaiserslautern, Germany*

[3] *Hitachi Europe GmbH, Germany*

[4] *Hitachi Ltd., Japan*

Keywords: Zero-shot Learning, Activity Recognition, And Hierarchical Model.

Abstract: We present a hierarchical framework for zero-shot human-activity recognition that recognizes unseen activities by the combinations of preliminarily learned basic actions and involved objects. The presented framework consists of gaze-guided object recognition module, myo-armband based action recognition module, and the activity recognition module, which combines results from both action and object module to detect complex activities. Both object and action recognition modules are based on deep neural network. Unlike conventional models, the proposed framework does not need retraining for recognition of an unseen activity, if the activity can be represented by a combination of the predefined basic actions and objects. This framework brings competitive advantage to industry in terms of the service-deployment cost. The experimental results showed that the proposed model could recognize three types of activities with precision of 77% and recall rate of 82%, which is comparable to a baseline method based on supervised learning.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Human activity recognition is important technology for many applications such as video surveillance systems, patient monitoring systems, and work support systems. A large body of works have investigated this technology especially in computer vision field (Aggarwal, 1999; Turaga et al., 2008; Lavee et al., 2009; Aggarwal and Ryoo, 2011).

The target of this study is workers' activities in factories. The conventional systems are designed to recognize particular set of activities by using supervised learning methods. Such systems, however, are not suitable for practical deployment because of the diversity of the activities in a practical field. It is usual in industrial situation that different factories have different demand for the target activities. In addition, the way an activity is performed may be different from factory to factory even though the name of the activity is identical. In these cases, the conventional systems need costly customization since they require retraining of the whole model for a new activity.

The goal of this research is to design a framework for zero-shot human activity recognition that overcomes this drawback and realize an easy-to-deploy system. The key idea is to recognize complex activities based on the combinations of simpler components, like the actions and objects involved in the activities.

Two wearable sensors, namely eye-tracking glass (ETG) and armband sensor, are utilized to recognize the basic objects and basic actions, respectively (Figure 1). Although many conventional systems use fixed cameras as sensors, wearable sensors are more appropriate especially in complex industrial environment because fixed cameras often suffer from the occlusion problem.

This framework enables to recognize a new activity without time-consuming retraining process if a new activity can be represented by a combination of predefined basic actions and objects. Figure 1 shows the overview of the proposed model. (Here, "action" is defined as a simple motion of body parts such as "raise arm" and "bend down", while "activity" is defined as

[*] Both authors contributed equally to this work

Al-Naser, M., Ohashi, H., Ahmed, S., Nakamura, K., Akiyama, T., Sato, T., Nguyen, P. and Dengel, A.
Hierarchical Model for Zero-shot Activity Recognition using Wearable Sensors.
DOI: 10.5220/0006595204780485
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 478-485
ISBN: 978-989-758-275-2
Copyright © 2018 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

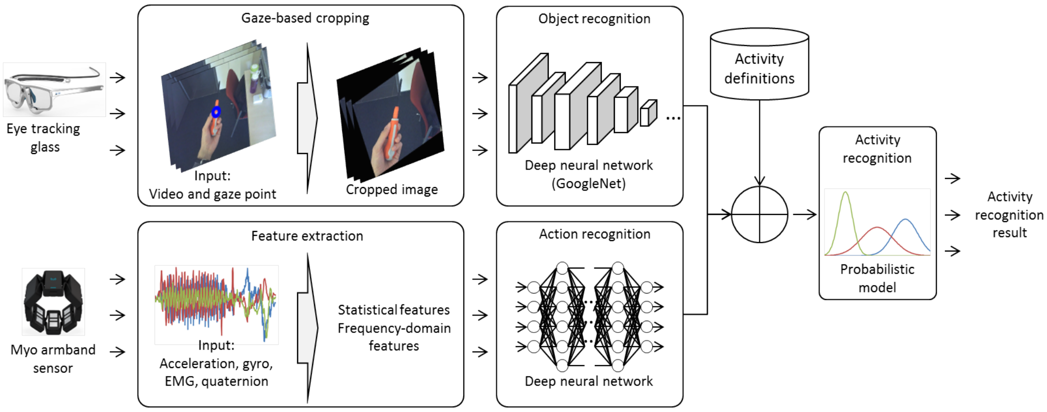


Figure 1: System overview. Object recognition module takes gaze-guided egocentric video and output the probabilities of basic objects. Action recognition module takes multi-modal armband signals and output the probabilities of basic actions. Activity recognition module process these probabilities to output the activity label.

a complex behavior such as “check manual” and “look for parts”.)

This study introduces a deep neural network (DNN) based action recognition method based on armband sensor and a gaze-guided object recognition method using ETG. The experimental result showed the accuracy of these two basic recognition methods are reasonably high. Moreover, the activity recognition method based on these two basic modules achieved about 80% both in precision and recall rate, which is comparable to the baseline method based on supervised learning.

SCIENCE AND TECHNOLOGY PUBLICATIONS

2 RELATED WORK AND CONTRIBUTION OF THIS STUDY

2.1 Related Work

Significant amount of studies has worked on the activity recognition problem. One of the most common and well-studied methods is the one based on video data obtained from a fixed camera (Wang et al., 2011; Wang and Schmid, 2013; Tran et al., 2015; Donahue et al., 2015). Especially a hierarchical model that uses two-stream DNN have achieved the state-of-the-art accuracy in various publicly available datasets (Simonyan and Zisserman, 2014; Wang et al., 2016; Peng and Schmid, 2016). The two-stream networks extract appearance based features (spatial features) and motion based features (temporal features) separately.

For example, Peng *et al.* (Peng and Schmid, 2016) introduced a method that extracts region of interest (ROI) by using a two-stream network that consists of RGB based faster R-CNN to extract appearance features and optical flow based faster R-CNN to extract motion features. Next to these region proposal networks, they added a multi-region generation layer to extract more detailed information. Their method achieved the best accuracy of 95.8% on the UCF sports dataset. However, the video data obtained from a fixed camera easily suffer from an occlusion problem especially in a practical environment such as in a factory.

To overcome this occlusion problem, the methods based on ego-centric video data have been studied. Ego-centric video data (or first-person-view video data) are obtained using a wearable camera devices such as Google glass. Recent surveys can be found in (Nguyen et al., 2016; Betancourt et al., 2015). Pioneering works (Ma et al., 2016; Li et al., 2015) showed the combination of motion and object cues computed from ego-centric video to infer the human activities. Ma *et al.* ’s method (Ma et al., 2016) is also based on two-stream network. One network was designed to detect objects by using hand location as a cue of ROI, and the other network recognize actions. Then the two networks are fine-tuned jointly to recognize objects, actions, and activities. This model outperformed state-of-the-are methods in average 6.6%.

Another way to overcome the occlusion problem of fixed-camera is to utilize data from other modalities. Spriggs *et al.* (Spriggs et al., 2009) used an egocentric camera, inertial measurement units (IMU) to classify kitchen activities. Maekawa *et al.* (Maekawa et al., 2010) used a wrist-mounted camera and sensors to detect activities in daily living (ADL). Fathi *et al.* (Fathi et al., 2012) and Li *et al.* (Li et al., 2015) use gaze information with egocentric video to recognize activities. It becomes easier to recognize certain activities by enriching the data source using multiple modalities, especially the data from different body parts such as head and arms.

Although the studies mentioned above have shown the very good performance in recognizing human activities, there is one more barrier for the practical deployment. It is the diversity of the activities and difficulty of collecting training data of those activities. It is usual in industrial situation that different factories have different demand for the target activities. In addition, the way an activity is performed may be different from factory to factory even though the name of the activity is identical. Those previous studies are designed to recognize activities that have been preliminarily learned. In other words, they require training-data collection and retraining of the model for recognizing a new activity.

Zero-shot learning (Palatucci et al., 2009; Socher et al., 2013) has a potential to address this challenge. Some works such as Lie *et al.*(Liu et al., 2011) and Cheng *et al.*(Cheng et al., 2013) have applied the concept of zero-shot learning to recognize a new activity on the basis of preliminarily learned attributes.

2.2 Contribution of This Study

As reviewed in section 2.1, there have been many researches on human activity recognition. This study takes the findings of all of those previous researches and tries to extend the previous researches toward practical deployment in the real world.

To do so, we decided to utilize ego-centric data to avoid the occlusion problem of fixed-cameras and zero-shot learning based approach to deal with the activities that are not preliminarily learned. Our activity recognition model is a hierarchical model. It recognizes the activity by a combination of objects involved in the activity and basic actions that compose the activity. It is known that the objects play an important role for activity recognition since it conveys contextual information (Jain et al., 2015; Yao et al., 2011; Ma et al., 2016). Although the basic components, namely, the object recognition module and the action recognition module need to be preliminarily trained, activities can be recognized without any training if the activity is represented as a combination of the predefined objects and actions.

We use SMI eye-tracking glass (ETG) and Myo armband sensor for our system. ETG is very useful to recognize an object that a target person is handling. The armband sensor measures IMU and electric myogenic data (EMG) and useful for recognizing an arm's movement, which is especially important for in an industrial situation,

To the best of our knowledge, this is the first study working on zero-shot activity recognition based on ego-centric video data and data from an armband. We will give a detailed description of the architecture to realize this concept as well as the quantitative evaluation results.

3 THE PROPOSED APPROACH FOR ACTIVITY RECOGNITION

The overview of the proposed model is shown in Figure 1. The model consists of three main components a) gaze-guided object detection module (Section 3.1), which is based on deep neural networks and is capable of recognizing objects, b) action recognition module (Section 3.2), which is also based on deep learning and uses Myo armband for detection of actions data, and c) activity recognition (Section 3.3), which can recognize complex activities, based on basic actions and objects detected by object and action detection modules.

3.1 Gaze-guided Object Recognition

The object recognition in the real world, especially in an industrial environment, is a challenging problem because of the complexity of the background. There are two ways of acquiring visual data by wearable sensors: attach a camera on head or on body, typically on chest. Since people sometimes look at something in the direction to which the body is not facing, on-head cameras capture more information on the wearer's view, or so-called 1st person view. In addition to the 1st person view video, eye-tracking glasses can capture which point the wearer is gazing at within the view ("gaze point"). This information is very useful because the gaze point is usually the point of interest for the user, and the system can easily focus of detecting the objects around the point of interest. In addition, SMI Eye-Tracking glass ETG 2 has 0.5 error degree and weighs 47 grams, which is the lightest on-head type vision sensors.

Since the gaze point usually indicates the point of interest of the wearer of the ETG, we assume that only the region around the gaze point is important and the other region in the image is not important. By cropping an image around the gaze point, a sub image that contains only a target object in reasonably large size can be obtained. Figure 2 shows an example of the cropped images.

The gaze-based cropping is applied not only in real-time recognition, but also in creating a training data set. Training data set is very important for building a good machine learning method. Especially for DNN models, which usually contain vast amount of parameters, acquiring enough amounts of data and diverse

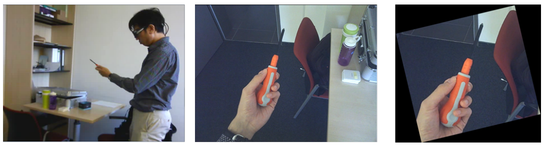


Figure 2: An example of cropped images. Left: The image from fixed camera. Center: The original image of the ETG. Right: The cropped image.

enough data is crucial to avoid the over-fitting problem. When cropping images around the gaze point, we randomly change the size of the cropping as well as the degree of rotation of the cropping area. By using this scheme, we can obtain $(60 * fps * Ns * Nr)$ data in 1 minute, where fps denotes the frame rate of the ETG video data, Ns denotes the number of different cropping sizes, and Nr denotes the number of different degrees of the rotation.

To deal with the case where no object is included in the cropped image, we defined “reject class” in addition to the target object categories. The reject class preferably includes all the possible objects and background scene other than the target objects. By training the model with this reject class, the model significantly becomes robust against the false positives.

We use GoogLeNet (Szegedy et al., 2015) as the initial model of the object recognition and fine-tune it with the collected training data using above-mentioned cropping method. GoogLeNet has 27 layers including Inception, CNN, and pooling layers. We fine-tuned the last two layers using our own training data.

3.2 Action Recognition

For industrial application, it is desireble to let the workers attach as less sensors as possible so as not to disturb them. We therefore decided to attach a sensor only on an arm, which is supposed to be one of the most important parts in many cases. We decided to use Myo armband because it is light weight (93 grams), has long battery life (more than 8 hours), and has good sensors (IMU sensors and electro-myogenic (EMG) sensors) that can be used for recognizing different types of arm actions.

DNN is used also for action recognition because of its overwhelming performance on most of the recognition tasks. As input, all the sensor data that can be collected with Myo armband, namely, quaternion, acceleration, gyro, and EMG data are utilized. The sensor data can be augmented if the number of training is not enough (Ohashi et al., 2017). Quaternion data are useful for representing the angle of an arm. Acceleration and gyro data are good for understanding the movement of the arm. EMG data well indicate the force of muscle contraction.

Features are firstly extracted from the raw sensor data by using a sliding window in order to deal with the time-sequential information of the actions. Statistical features such as maximum, minimum, mean and standard deviation as well as the features in frequency domain, namely, amplitude spectrum obtained by applying fast Fourier transform (FFT), are utilized in this research. The statistical features are good indicators for the intensity of the actions as well as how it changes within the sliding window, and the frequency-domain features are good for understanding the periodicity of the actions.

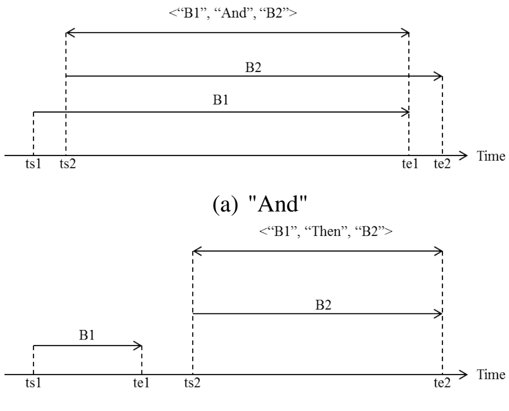
The “reject class” is defined in the action recognition model as well to deal with the case when no target action is performed.

3.3 Activity Recognition

One of the most common ways in hierarchical activity recognition model is to take the outputs of the action recognition module and the object recognition module as input, and build some classifier that automatically learns the mapping from the them to the target activity categories. However, reasonably big amount of training data is required in order to train a model that does this mapping well. In addition, if a new activity is added to the target activities, additional training data for the new activity needs to be collected.

In order to avoid this time-consuming data collection procedure, this study proposes a zero-shot recognition scheme. We define activities using name of objects, name of actions, and the conjunction words, which defines the relationship between the objects and actions. “And”, and “Then” are used as the concrete conjunction words. If an activity defined as <“B1”, “And”, “B2”>, its period is defined as the period when both B1 and B2 are observed. Here, B1 and B2 denote either object name or action name. Figure 3 (a) shows the image of the period that “And” represents. As shown in the figure, the time between $ts2$ and $te1$ are recognized as the period when the target activity occurred. Similarly, if an activity is defined as <“B1”, “Then”, “B2”>, its period is defined as the period when B2 is observed after B1 is observed. As shown in Figure 3 (b), the time between $ts2$ and $te2$ are recognized as the period when the target activity occurred. For example, the activity “Tightening a screw” can be defined as <”Screw driver”, “And”, “Twisting”>, and "Opening a lid of a bottle" can be defined as <"Bottle", "Then", "Twisting"> As explained in these examples, one basic class (in this case, "Twisting") can be used to represent multiple activities.

A probabilistic framework is employed for the activity recognition model to enhance and stabilize the



(a) "And"

(b) "Then"

Figure 3: Periods that the conjunction word "And" and "Then" represent.

performance. First, the activity recognition module receives the array of probabilities from basis recognition module, each of which represent the likelihood of each target object or action. Then the probability of the activity can be calculated as the conditional probabilities as follows.

$$\begin{aligned} & p(activity \mid s; def(activity)) \\ = \; & p(activity \mid object, action; def(activity)) \\ & p(object \mid s) p(action \mid s) \end{aligned} \tag{1}$$

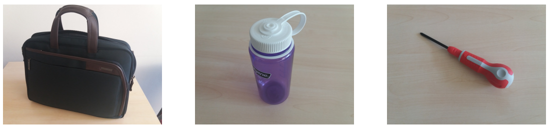
, where $def(activity)$ denotes the definition of the activity and s denotes the sensor data. This framework provides more robust and stable activity recognition results even if the basis recognition results are not very accurate.

4 EVALUATION

4.1 Experimental Data

The experimental data were collected in a laboratory. Table 1 shows the details of the data for evaluating the activity recognition method. The target activities are defined to be "Putting a bag on a table", "Opening a lid of a bottle", and "Tightening a screw". The definition of the activities is shown in Table 2. The data collection procedure was as follows. (1) Start recording data. (2) A subject performs one activity 3 to 5 times in a row with short interval between each performance. (3) Stop recording data. (4) Restart recording data. (5) The subject performs the 2nd activity 3 to 5 times in a row. (6) Stop recording data. (7) Iterate the same procedure for the last activity. (8) Iterate the same procedure for the other subjects.

Table 3 summarizes the data collected for evaluating the object recognition method. The basic objects involved in these activities are "a bag", "a bottle" and "a screw driver" (see Figure 4). In addition, "Reject class" is added to the target object classes in order to deal with the case of "no target object". An ETG was used to collect the training data. In order to acquire good amount of diverse data, a subject kept looking at an object from different angles and from different distances. Then a sub-image around the gaze point was cropped as explained in the section 3.1. Training data and test data were separately collected.



(a) Bag (b) Bottle (c) Screw driver

Figure 4: Target objects.

Table 4 summarizes the data collected for evaluating the action recognition method. The basic actions to compose the above-mentioned activities are "holding" and "twisting".

In addition, "Reject class" is included in order to deal with the case of "no target action". Training data and test data were separately collected. Only one subject participated in the both of data collection for action recognition and activity recognition.

4.2 Results

Table 5 shows the confusion matrix of the object recognition result. As shown in the figure, "bag" was often misclassified as "reject", and as a result, the recall rate of the "bag" class and the precision of the "reject" class was low. On the other hand, both precision and recall rate were high for "bottle" and "screw driver" class.

Table 6 shows the confusion matrix of the action recognition result. As shown in the figure, both precision and recall rate were high for all of the classes.

In order to compare the proposed zero-shot activity recognition model, a baseline method that utilizes normal supervised learning method (sometimes it's called "many-shots" learning as opposed to zero-shot learning) was implemented. The baseline method takes the output from the basis recognition modules, namely, the array of probabilities as an input and trained to output the corresponding activity. SVM was selected for the model. DNN was not selected simply because of the amount of available training data.

Table 7 shows the evaluation result of the activity recognition method. Intersection over union (IOU) based on ground truth and estimated results are calculated for the evaluation. Each estimation result was regarded as correct if the IOU is more than a threshold.

Figure 5 shows the precision and recall rate of the

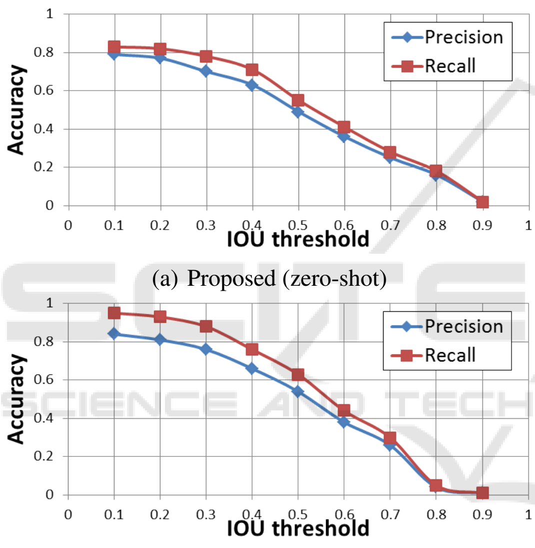
SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 1: Evaluation data for activity recognition method.

Number of subject	12
Number of target classes	4
Target classes	Putting a bag on a table, Opening a lid of a bottle, Tightening a screw, Others(reject class)
Number of data	Total: 131 Putting a bag on a table: 39, Opening a lid of a bottle: 50, Tightening a screw: 42

Table 2: Definition of the activities.

Putting a bag on a table	<"Bag", "Then", "Holding">
Opening a lid of a bottle	<"Bottle", "Then", "Twisting">
Tightening a screw	<"Screw driver", "And", "Twisting">



(a) Proposed (zero-shot)

(b) Baseline (many-shots)

Figure 5: Precision and recall rate for different IOU thresholds.

proposed method and the baseline method for different IOU thresholds. Figure 6 shows an example of the output from the proposed method.

5 DISCUSSION

In this section we will discuss the evaluation results, limitation, and also the future work.

The current model can perform a real time activity recognition of untrained activities using a combination of basic components. The impotence of real time recognition comes from the environment where the system can be deployed. In factories and maintenance sites, the real time recognition can prevent the wrong activities.

Table 7 shows that the model has a achieved a very good accuracy for unlearned activities, with 77% precision and 82% recall rate.

The "Putting bag" activity has lower recall rate comparing to the other activities. This is because of the low recall rate of the "bag" class in the object recognition module. Since the bag used in the experiment doesn't have much texture and also its size was significantly different from the other objects (Figure 4), the cropped images of the bag sometimes looked like just a black square. Another reason for this lower recall rate is the action recognition module. Even though the recall rate of the "Holding" action was very high as shown in table 6, sometimes the "Holding" action was not correctly recognized. The most dominant feature for recognizing "Holding" action is EMG data, which is more likely to be affected by subjects. As mentioned in section 4.1, only one subject's data out of 12 subjects was included in the dataset to train the action recognition module. To develop a more robust and subjects-independent action recognition method is one of our future work.

On the other hand, precision of the "Opening lid" activity and "Tightening screw" activity was relatively low. Figure 6 shows the recognition result of "Tightening screw" activity. It shows that none of the 5 "Tightening screw" activity was missed (recall is 100%). On the other hand, the first attempt was recognized as two activities of "tightening a screw" because the probability dropped down bellow the threshold in the middle of the activity. It sometimes happens because the subjects stopped performing the activity for some reason, for example, dropping the screwdriver, which was observed during the experiment. This is reflected on the number of recognized activities, which reduces the precision. One future work is to consider a threshold of the time between the consecutive activities to merge them and reduce this type of false alarm.

Comparing the results of the proposed method with that of the baseline method (Figure 5), we can see the precision performance is close to each other, even though the activities is untrained in our model, while it is trained in the baseline model. On the other hand, the recall rate of the baseline method was better (93%).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 3: Evaluation data for object recognition method.

Number of target classes	4
Target classes	Bag, Bottle, Screw driver, Others(reject class)
Number of training data	Total: 16,000 Bag: 4,000, Bottle: 4,000, Screw driver: 4,000, Others(reject class): 4,000
Number of test data	Total: 3956 Bag: 1007, Bottle: 953, Screw driver: 984, Others(reject class): 1012

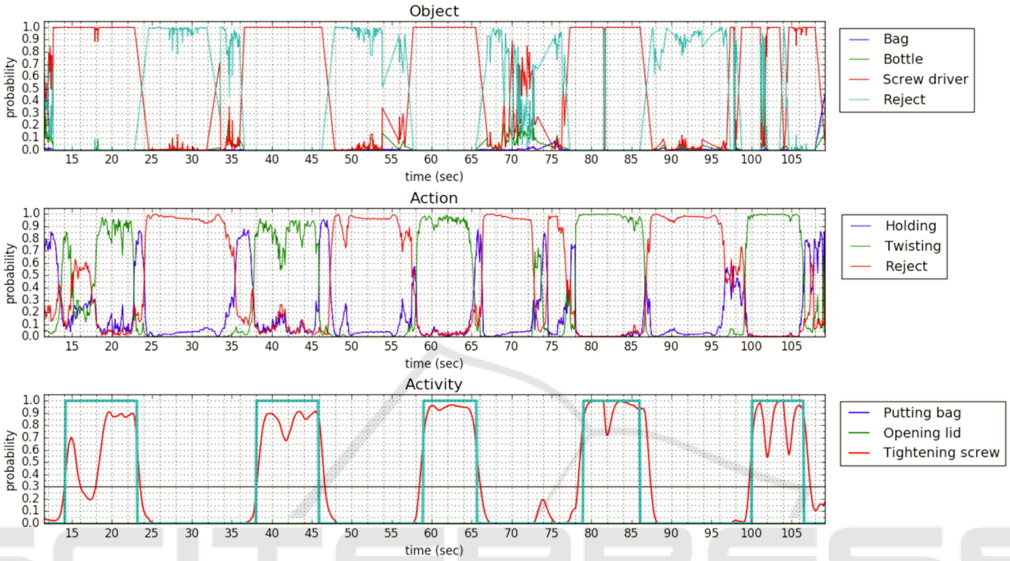


Figure 6: An example of the recognition result. ("tightening a screw" activity).

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 4: Evaluation data for action recognition method.

Number of subject	3
Number of target classes	3
Target classes	Holding, Twisting, Others(reject class)
Number of training data	Total: 10814 Holding: 3583, Twisting: 3429, Others(reject class): 3802
Number of test data	Total: 5853 Holding: 2029, Twisting: 2025, Others(reject class): 1799

This is due to the characteristic of the models. If an activity is defined using the conjunction word "Then", the proposed model can recognize the activity only when the first target is recognized. The baseline model recognizes the activity by combining all the probabilities. It could sometimes recover even when the first target is not recognized if the second target is recognized with high probabilities. The proposed framework sometimes works fine to reduce false alarm, but sometimes leads to lower recall rate like this example.

Table 5: Confusion matrix of the object recognition method.

	Bag	Bottle	Screw driver	Reject	Total	Recall
Bag	443	91	15	458	1007	0.44
Bottle	0	945	1	7	953	0.99
Screw driver	0	10	949	25	984	0.96
Reject	17	16	19	960	1012	0.95
Total	460	1062	984	1450	3956	-
Precision	0.96	0.89	0.96	0.66	-	0.83

Table 6: Confusion matrix of the action recognition method.

	Holding	Twisting	Reject	Total	Recall
Holding	2019	6	4	2029	0.99
Twisting	112	1872	41	2025	0.92
Reject	48	58	1693	1799	0.94
Total	2179	1936	1738	5853	-
Precision	0.93	0.97	0.94	-	0.95

Table 7: Evaluation result of the activity recognition method. Threshold for IOU: 0.2.

Activity	Precision	Recall
Putting bag	0.96 (27 / 28)	0.69 (27 / 39)
Opening lid	0.72 (44 / 61)	0.88 (44 / 50)
Tightening screw	0.73 (43 / 59)	0.88 (37 / 42)
Total	0.77 (114 / 148)	0.82 (108 / 131)

6 CONCLUSIONS

A hierarchical human activity recognition model that recognizes human activities by the combinations of the basic actions and involved objects has been proposed in order to realize an easy-to-deploy activity recognition system. Unlike conventional activity recognition models, the proposed model does not need retraining for recognizing a new activity if the activity is represented by a combination of predefined basic actions and basic objects. Two wearable sensors, namely Myo armband sensor and ETG, have been utilized for the action recognition and object recognition, respectively. The experimental results have shown that the accuracy of both basic modules are reasonably high, and the proposed model could recognize 3 types of activities with precision of 77% and recall rate of 82%. The future works include expansion of target activities as well as enhancing the basic modules.

REFERENCES

Aggarwal, J. (1999). Human Motion Analysis: A Review. *CVIU*, 73(3):428–440.

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys*, 43(3):1–43.

Betancourt, A., Morerio, P., Regazzoni, C. S., and Rauterberg, M. (2015). The evolution of first person vision methods: A survey. *IEEE Trans. Circuits and Systems for Video Technology*, 25(5):744–760.

Cheng, H.-T., Sun, F.-T., Griss, M., Davis, P., Li, J., and You, D. (2013). Nuactiv: Recognizing unseen new activities using semantic attribute-based learning. In *Proc. International Conference on Mobile Systems, Applications, and Services*, pages 361–374. ACM.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Fathi, A., Li, Y., and Rehg, J. M. (2012). Learning to recognize daily actions using gaze. In *ECCV*.

Jain, M., van Gemert, J. C., and Snoek, C. G. (2015). What do 15,000 object categories tell us about classifying and localizing actions? In *CVPR*, pages 46–55.

Lavee, G., Rivlin, E., and Rudzsky, M. (2009). Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. Systems, Man and Cybernetics Part C: Applications and Reviews*, 39(5):489–504.

Li, Y., Ye, Z., and Rehg, J. M. (2015). Delving into egocentric actions. In *CVPR*, pages 287–295.

Liu, J., Kuipers, B., and Savarese, S. (2011). Recognizing human actions by attributes. In *CVPR*, pages 3337–3344. IEEE.

Ma, M., Fan, H., and Kitani, K. M. (2016). Going deeper into first-person activity recognition. In *CVPR*, pages 1894–1903.

Maekawa, T., Yanagisawa, Y., Kishino, Y., Ishiguro, K., Kamei, K., Sakurai, Y., and Okadome, T. (2010). Object-based activity recognition with heterogeneous sensors on wrist. In *Proc. International Conference on Pervasive Computing*, pages 246–264. Springer.

Nguyen, T.-H.-C., Nebel, J.-C., Florez-Revuelta, F., et al. (2016). Recognition of activities of daily living with egocentric vision: A review. *Sensors*, 16(1):72.

Ohashi, H., A. Naser, M., Ahmed, S., Akiyama, T., Sato, T., Nguyen, P., Nakamura, K., and Dengel, A. (2017). Augmenting Wearable Sensor Data with Physical Constraint for DNN-Based Human-Action Recognition. In *Time Series Workshop @ ICML*.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418.

Peng, X. and Schmid, C. (2016). Multi-region two-stream r-cnn for action detection. In *ECCV*, pages 744–759.

Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576.

Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943.

Spriggs, E. H., De La Torre, F., and Hebert, M. (2009). Temporal segmentation and activity classification from first-person sensing. In *CVPR Workshops*, pages 17–24. IEEE.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *CVPR*, pages 1–9.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *ICCV*.

Turaga, P., Chellappa, R., Subrahmanian, V. S., and Udrea, O. (2008). Machine recognition of human activities: A survey. *IEEE Trans. Circuits and Systems for Video Technology*, 18(11):1473–1488.

Wang, H., Kl, A., Schmid, C., and Liu, C.-l. (2011). Action recognition by dense trajectories. In *CVPR*, pages 3169–3176.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *ICCV*, pages 3551–3558.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: towards good practices for deep action recognition. In *ECCV*, pages 20–36.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., and Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. In *ICCV*, pages 1331–1338.