Recognizing Activities of Daily Living with a Wrist-mounted Camera

Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, Tatsuya Harada
Graduate School of Information Science and Technology, The University of Tokyo
{ohnishi, kanehira, kanezaki, harada}@mi.t.u-tokyo.ac.jp

Abstract

We present a novel dataset and a novel algorithm for recognizing activities of daily living (ADL) from a first-person wearable camera. Handled objects are crucially important for egocentric ADL recognition. For specific examination of objects related to users’ actions separately from other objects in an environment, many previous works have addressed the detection of handled objects in images captured from head-mounted and chest-mounted cameras. Nevertheless, detecting handled objects is not always easy because they tend to appear small in images. They can be occluded by a user’s body. As described herein, we mount a camera on a user’s wrist. A wrist-mounted camera can capture handled objects at a large scale, and thus it enables us to skip the object detection process. To compare a wrist-mounted camera and a head-mounted camera, we also developed a novel and publicly available dataset¹ that includes videos and annotations of daily activities captured simultaneously by both cameras. Additionally, we propose a discriminative video representation that retains spatial and temporal information after encoding the frame descriptors extracted by convolutional neural networks (CNN).

1. Introduction

Recently, activity recognition from first-person camera views has been attracting increasing interest, motivated by advances in wearable device technology. Recognition of activities of daily living (ADL) from first-person views is an important task related to activity recognition. ADL are basic activities in a typical human life such as “making coffee” or “cutting paper.” If the system recognizes ADL properly, then it is applicable to nursing services, rehabilitation, and lifestyle habit improvements.

To recognize ADL, it is important to examine objects undergoing hand manipulation specifically. For example, a cleaning activity might be recognized only by recognizing that a user is using a vacuum cleaner. One can also recog-

¹http://www.mi.t.u-tokyo.ac.jp/static/projects/miladl/

Figure 1. Activities of daily living (ADL) captured by a wrist-mounted camera (left) and a head-mounted camera (right).
To train an ADL recognition system without pixel-level annotations or bounding boxes of objects, we consider mounting a wearable camera on the wrist of the user’s dominant arm because the objects are handled mainly by the user’s dominant hand. We designate this camera as a wrist-mounted camera in this paper. For ADL recognition, a wrist-mounted camera has numerous advantages over a head-mounted camera or a chest-mounted camera: (a) Wrist-mounted cameras can capture a large image of the handled objects. (b) Because handled objects are close to a user’s dominant hand, the object positions are restricted in the images from the wrist-mounted camera. (c) Because of the above reasons, we can skip object detection and do not need a dataset of ADL with bounding boxes. We only need a dataset of ADL with annotations about the activity time segments in the videos.

We also propose a recognition system for videos captured by a wrist-mounted camera that has strong spatial bias and weak temporal bias. As shown in Figure 2, an image captured by a wrist-mounted camera has strong spatial bias, meaning that hand-manipulated objects tend to be located at the central area. In addition, the order of manipulated objects is mostly fixed for each action, which we call “weak temporal bias.” The state-of-the-art video representation, which extracts local features containing spatial information from pre-trained convolutional neural networks (CNNs), strongly loses spatial information and completely loses temporal information after encoding. Therefore, we also propose a novel video representation that retains spatial and temporal information after encoding to consider the above mentioned biases.

Our three contributions are the following:

1. We propose the use of a wrist-mounted camera for ADL recognition instead of a head-mounted camera or a chest-mounted camera.
2. We propose a discriminative video representation that retains spatial and temporal information. This is a method for the dataset captured from a wrist-mounted camera that has a large bias of spatial information and a small bias of temporal information.
3. We developed a novel and publicly available dataset that includes videos and annotations of ADL captured from a head-mounted camera and a wrist-mounted camera simultaneously.

2. Related work

2.1. Egocentric vision for ADL recognition

Various approaches for ADL recognition based on handled objects have been proposed [24, 29, 35]. Because wearable devices with cameras such as GoPro and Google Glasses have been developed recently, ADL recognition with viewpoint cameras has received a considerable amount of attention. Some works on egocentric ADL recognition have achieved results in a single environment, such as a kitchen or an office [9, 7, 18, 19].

For more practical settings, Pirsiavash and Ramanan [26] estimated the type of a handled object by detecting the object and arm from a wearable camera’s viewpoint. They demonstrated that action recognition performs well in diverse environments. However, it is necessary to provide positional information of all objects in all frames of the video at the time of learning. In addition, detecting an entire handled object itself is still difficult. Although their dataset has various annotations, such as type of activity and duration of its completion, as well as type of an object and its location, it took over 1 month to label various annotations by 10 part-time annotators. Consequently, expanding the dataset is not practical. In a more practical setting, an ADL recognition system that uses wearable devices in diverse environments should be trained with labels obtained by simpler annotation methods.

2.2. Video representation for action recognition

Video representation has been well studied in the action recognition domain. Some deep-learning approaches for action recognition have been proposed [23, 31]. However, these approaches require the use of large-scale video datasets (e.g. Sports 1M [14]), which are difficult to address and which require enormous amounts of time for the whole learning process.

Motion features: The general pipeline to obtain a video representation for action recognition models the distribution of local features from training videos. Local features representing motion information (e.g., HOG [4], HOF [20], and MBH [5]) are usually used. The combination of local features and improved dense trajectory (iDT) [34], which compensates for camera motion, is the de facto standard. It has shown great performance for action recognition [33].

CNN descriptors: CNN has achieved superior results to the standard pipeline for object recognition [16]. Jain et al. [12] brought CNN to action recognition. They obtain the state of a fully connected layer from each frame in videos and calculate the video representation by averaging all CNN features. Their method therefore exhibits performance that is surprisingly comparable to the combination of iDT, MBH, and Fisher vector (FV) [25]. To obtain more discriminative features containing spatial information, Xu et al. [36] proposed the extraction of latent concept descriptors (LCDs) from the pool3 layer and the application of VLAD [13] instead of averaging. However, spatial information is ignored when applying VLAD. Since our task is an intermediate task of action recognition and object recognition, we developed the CNN-based video representation above to design the video representation for ADL recognition from wrist-mounted cameras, which have strong spatial information bias.
3. Wrist-mounted cameras

Some works in the area of interface research have shed light on wrist-mounted cameras [32, 15]. In ADL recognition, Maekawa et al. [21] conducted multimodal ADL recognition using a wrist-mounted device that has a camera, microphone, acceleration sensor, illuminance meter, and digital compass. However, the color histogram alone is used as an image feature. This system is too simple to identify handled objects. Wrist-mounted cameras have never been evaluated carefully in ADL recognition. Therefore, we discuss the superiority of wrist-mounted cameras in this section.

Wrist-mounted cameras capture handled objects very closely, as shown in Figure 1. In addition, as shown in Figure 2, the user’s hand invariably appears on the right side of the wrist-mounted camera image, although we cannot see anything in the head-mounted camera image. This implies that the user’s hand always appears in the right side and handled objects appear near the center of the wrist-mounted camera image.

4. Video representation

We propose a new video representation based on LCDs [36] to take advantage of the strong spatial bias and weak temporal bias of the video captured by a wrist-mounted camera. Although LCD retains the spatial information in each frame at the descriptor level, spatial information and temporal information are dropped when the descriptors are encoded and aggregated into a video representation. However, the video captured by a wrist-mounted camera has a strong spatial bias, as described in Section 3. Although not as strong as spatial bias, temporal bias also exists because the order of handling objects is fixed roughly in each action class. Therefore, we use the benefits of these spatial and temporal biases specifically for a wrist-mounted camera and ADL.

Our method encodes LCDs at each location in all frames into single VLAD [13] vectors and optimizes the weights for the VLAD vectors to aggregate them into a video representation. The weight for a VLAD vector extracted from each location is designated as a spatial weight. Furthermore, we propose a method that divides a video into short sequences and optimizes the temporal weights for aggregating descriptors. Here, we describe the original LCD in Section 4.1, the proposed method to optimize spatial weights in Section 4.2, and the proposed method to optimize spatial and temporal weights in Section 4.3.

4.1. CNN latent concept descriptors

Latent concept descriptors [36] constitute a state-of-the-art video representation using CNN, which is obtained as follows. (i) Given a video \( E \) including \( T \) frames \( E = \{I_1, I_2, \ldots, I_T\} \), each frame is input to VGG net [27] pre-trained on the ImageNet2012 dataset [6] to obtain the pool\(_3\) layer’s output. The dimension of pool\(_3\) features is \( a \times a \times M \), where \( a \) is the size of the filtered images of the last pooling layer and \( M \) is the number of convolutional filters in the last convolutional layer (in the case of VGG net, \( a = 7 \) and \( M = 512 \)). (ii) The responses of \( M \) filters are concatenated for the respective locations of the pool\(_3\) layer.

Then, a set of \( a^2 \) descriptors \( f^t_{i,j} \in \mathbb{R}^M \) is obtained from the \( t \)-th frame as follows.

\[
\mathcal{F}^t = \{f^t_{(1,1)}, f^t_{(1,2)}, \ldots, f^t_{(a,a)}\}.
\]

(iii) All descriptors in \( \{\mathcal{F}^1, \ldots, \mathcal{F}^T\} \) are encoded with VLAD into a video representation \( v \). Letting \( \{c_1, \ldots, c_K\} \) denote a set of \( K \) coarse centers obtained by K-means, we obtain \( u_k \ (k = 1, \ldots, K) \) as follows:

\[
u_k = \sum_{(t,i,j) \in \{(t,i,j)\mid \text{NN}(f^t_{i,j}) = c_k\}} (f^t_{i,j} - c_k).
\]

Therein, \( \text{NN}(f^t_{i,j}) \) represents the nearest center of \( f^t_{i,j} \). Then, \( v \) is obtained as an \( MK \)-dimensional VLAD encoding vector by concatenating \( u_k \) over all \( K \) centers. (iv) Finally, \( v \) is normalized by power and L2 normalization with intra-normalization [3].
4.2. Discriminative spatial aggregated latent concept descriptors

The original LCD [36] drops the spatial information in the process of VLAD encoding because the descriptors are equally weighted when they are encoded with VLAD into a video representation $v$. We introduce spatial weights for the VLAD encoding vectors distinguished by their locations when we aggregate them into a final video representation. Because of the spatial weights, we can address the spatial bias such that the center area is more important than its surroundings, for example, because hand-manipulated objects tend to be located at the center view of a wrist-mounted camera. Specifically, we obtain a video representation $v(i, j)$ for each cell $(i, j)$ over all the $T$ frames by encoding the descriptors in a set $F \{ (i,j) \}$. In the same manner as in [36], $v(i, j)$ is normalized by power and L2 normalization within each cell. Letting $w(i, j) \in \mathbb{R}^{N_{sp}}$ denote an $N_{sp}$-dimensional weight vector, we obtain a weighted sum of $v(i, j)$ as

$$V = \sum_{i=1}^{a} \sum_{j=1}^{a} w(i, j)v_{sp}(i,j)^T = V_{sp}W_{sp},$$

where $V_{sp} \in \mathbb{R}^{MK \times a^2}$ and $W_{sp} \in \mathbb{R}^{a^2 \times N_{sp}}$ are defined as shown below.

$$V_{sp} = (v(1,1), v(1,2), \ldots, v(a,a)), \quad W_{sp} = (w(1,1), w(1,2), \ldots, w(a,a))^T.$$

As described in this paper, we obtain $W_{sp}$ by arranging $N_{sp}$ eigenvectors $x \in \mathbb{R}^{a^2}$ obtained by partial least squares (PLS) in $N_{sp}$ rows. Note that PLS is a method that can extract common information between sets of observed features. Therefore, $N_{sp}$ represents how many eigenvectors we use were obtained from PLS. Details related to computing the eigenvectors are given in the Supplemental Materials. Finally, we obtain a video representation $v_{DSAR} \in \mathbb{R}^{MK \times N_{sp}}$, which is called discriminative spatial aggregated LCDs (DSAR), by concatenating all elements in $V \in \mathbb{R}^{MK \times N_{sp}}$ in (3). Here, $v_{DSAR}$ is normalized by power and L2 normalization.

The idea to use the eigenvectors obtained by PLS as spatial weights was derived from the discriminative spatial pyramid representation (D-SPR) [10]. Consequently, the proposed method described in this section can be regarded as the combination of LCD and D-SPR. Our method optimizes the weights for $a \times a$ cells in the output of pool3, whereas D-SPR optimizes the weights for areas in a spatial pyramid of each frame.

4.3. Discriminative spatiotemporal aggregated latent concept descriptors

The wrist-mounted camera dataset not only has a strong spatial information bias, but also has a weak temporal information bias. The original LCD described in Section 4.1 and the proposed DSAR described in Section 4.2 lose temporal information in the process of VLAD encoding. Inspired by the idea of spatiotemporal pyramids [17], we introduce temporal weights for the VLAD encoding vectors distinguished by their time stamps when we aggregate them into a final video representation. Because of the temporal weights, we can assign the importance of each frame into a whole video representation. Specifically, we split a video into $2^l$ sequences consisting of equal numbers of frames $T/2^l$. The $s$-th sequence is a set of frames $\{ I_{s1T/2^l+1}, \ldots, I_{sT/2^l} \}$ ($s = 1, \ldots, 2^l$). Then, we obtain a video representation $v_{sp}^l(i, j)$ for each cell $(i, j)$ over all the $T/2^l$ frames in the $s$-th sequence by encoding the descriptors in a set $F_{(i,j)}^s = \{ f_{(i,j)}^{s1T/2^l+1}, \ldots, f_{(i,j)}^{sT/2^l} \}$. Here, we consider multiple levels of the splitting ($l = 0, \ldots, L$) such that we obtain a set of $v_{sp}^l(i, j)$ as follows:

$$V = \{ v_{sp}^l(i, j), \ldots, v_{sp}^l(i, j), \ldots, v_{sp}^{2l}(i, j) \}$$

$$i = 1, \ldots, a, j = 1, \ldots, a.$$ Again, $v_{sp}^l(i, j)$ is normalized by power and L2 normalization with intra-normalization. Letting $w_{sp}^l \in \mathbb{R}^{N_{temp}}$ denote an $N_{temp}$-dimensional weight vector, we obtain a weighted
sum of $v^i_{s}(i, j)$ as follows:

$$V = \sum_{i=1}^{a} \sum_{j=1}^{a} \sum_{i=0}^{2^l} v^i_{s}(i, j) \otimes w(i,j) \otimes w^j_s.$$  \hfill (7)

Here, we define $V(i,j) \in \mathbb{R}^{MK \times d}$, $V^i_s \in \mathbb{R}^{MK \times a^2}$, and $W_{tmp} \in \mathbb{R}^{d \times N_{tmp}}$, where $d = \sum_{l=0}^{L} 2^l = 2^{L+1} - 1$ as shown below.

$$V(i,j) = (v^1_{s}(i, j), v^i_{s}(i, j), \ldots, v^i_{L}(i, j), \ldots, v^i_{L}(i, j)), \hfill (8)$$

$$V^i_s = (v^i_{s}(1, 1), v^i_{s}(1, 2), \ldots, v^i_{s}(a, a)), \hfill (9)$$

$$W_{tmp} = (w^0, w^1, \ldots, w^L, \ldots, w^L_{2L})^T. \hfill (10)$$

As described in this paper, we optimize spatial weights $W_{sp}$ and temporal weights $W_{tmp}$ iteratively and alternately. Specifically, we repeat the following two steps.

**Step 1: optimizing $W_{sp}$**

In this step, we fix $W_{tmp}$ and optimize $W_{sp}$. We obtain $MK N_{tmp}$-dimensional vectors $g(i,j)$ by concatenating all the elements in $V(i,j) W_{tmp}$. Letting $V^I \in \mathbb{R}^{MK N_{tmp} \times a^2}$ denote $(g(1,1), g(1,2), \ldots, g(a,a))$, we can rewrite (7) as

$$V = V^I W_{sp}. \hfill (11)$$

This formulation is identical to (3). Therefore, we optimize $W_{sp}$ in the manner described in Section 4.2.

**Step 2: optimizing $W_{tmp}$**

In this step, we fix $W_{sp}$ and optimize $W_{tmp}$. We obtain $MK N_{sp}$-dimensional vectors $h^i_s$ by concatenating all the elements in $V^I W_{sp}$. Letting $V^H \in \mathbb{R}^{MK N_{sp} \times d}$ denote $(h^0_s, h^1_s, \ldots, h^L_s, \ldots, h^L_{2L})$, we can then rewrite (7) as

$$V = V^H W_{tmp}. \hfill (12)$$

This formulation is identical to (3). Therefore, we optimize $W_{tmp}$ in the manner presented in Section 4.2.

We iterate Step 1 and Step 2 several times. Finally, we obtain a video representation $v_{DSTAR} \in \mathbb{R}^{MK N_{sp} N_{tmp}}$, which is called discriminative spatiotemporal aggregated LCDs (DSTAR), by concatenating all the elements in $V \in \mathbb{R}^{MK N_{sp} N_{tmp}}$ in (7) with power and L2 normalization. An illustration of $v_{DSTAR}$ is shown in Figure 3.

### 5. Dataset: MILADL

We created a new ADL dataset that uses both a wrist-mounted camera and a head-mounted camera because there are as yet no published ADL datasets that use wrist-mounted cameras. In this section, we present the details of our dataset.

Note that it is also important to compare a wrist-mounted camera with a chest-mounted camera instead of with a head-mounted one since a chest-mounted camera is closer to the user’s hands. We encourage to compare wrist-mounting to other mountings [22] for ADL recognition as future work.

<table>
<thead>
<tr>
<th>Action name</th>
<th>Mean of length (s)</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>vacuuming</td>
<td>38.1</td>
<td>17</td>
</tr>
<tr>
<td>empty trash</td>
<td>11.5</td>
<td>22</td>
</tr>
<tr>
<td>wipe desk</td>
<td>35.7</td>
<td>23</td>
</tr>
<tr>
<td>turn on air-conditioner</td>
<td>6.9</td>
<td>27</td>
</tr>
<tr>
<td>open and close door</td>
<td>6.4</td>
<td>34</td>
</tr>
<tr>
<td>make coffee</td>
<td>88.2</td>
<td>24</td>
</tr>
<tr>
<td>make tea</td>
<td>70.0</td>
<td>22</td>
</tr>
<tr>
<td>wash dishes</td>
<td>31.3</td>
<td>29</td>
</tr>
<tr>
<td>dry dishes</td>
<td>15.7</td>
<td>27</td>
</tr>
<tr>
<td>use microwave</td>
<td>33.7</td>
<td>26</td>
</tr>
<tr>
<td>use refrigerator</td>
<td>6.8</td>
<td>42</td>
</tr>
<tr>
<td>wash hands</td>
<td>11.9</td>
<td>32</td>
</tr>
<tr>
<td>dry hands</td>
<td>7.9</td>
<td>29</td>
</tr>
<tr>
<td>drink water from a bottle</td>
<td>13.8</td>
<td>26</td>
</tr>
<tr>
<td>drink water from a cup</td>
<td>7.8</td>
<td>31</td>
</tr>
<tr>
<td>read book</td>
<td>28.5</td>
<td>28</td>
</tr>
<tr>
<td>write on paper</td>
<td>16.6</td>
<td>29</td>
</tr>
<tr>
<td>open and close drawer</td>
<td>6.0</td>
<td>30</td>
</tr>
<tr>
<td>cut paper</td>
<td>14.7</td>
<td>28</td>
</tr>
<tr>
<td>staple paper</td>
<td>7.8</td>
<td>28</td>
</tr>
<tr>
<td>fold origami</td>
<td>68.7</td>
<td>23</td>
</tr>
<tr>
<td>use smartphone</td>
<td>23.1</td>
<td>29</td>
</tr>
<tr>
<td>watch TV</td>
<td>21.6</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 1. Duration of each class and the distribution of the 23 classes in our dataset.

### 5.1. Activity class

We chose activity classes by referring to previous studies of ADL [26, 30, 8]. First, we removed some classes that many users were reluctant to record on video such as “brushing teeth” and “laundry.” Next, to introduce more variety into our dataset, we added some actions referring to other ADL recognition studies [30] and an evaluation of Alzheimer rehabilitation [8]. As Table 1 shows, we strove to recognize 23 ADL classes in this study. Detailed information is given in the Supplemental Materials.

### 5.2. Collection and annotation

To assemble the dataset, we used a GoPro HERO3+ 2 as the head-mounted camera and an HX-A100 3 as the wrist-mounted camera. Each user wore these two cameras, as shown in Figure 4. Each user therefore recorded two videos simultaneously. As in a previous ADL egocentric dataset [26], we did not instruct the users in detail how to act to obtain realistic data. After taking videos, all users manually annotated the duration and the action class in their own videos. The definition of an action includes some initial and final actions related to the action. For example, the action “cutting paper” is defined as follows: The initial action of “cutting paper” is to take scissors from the table and the final action is to put it on the table. We recruited 20 people to perform these tasks. All users were right handed. Our wrist-

---

2http://jp.shop.gopro.com/cameras
3http://panasonic.jp/wearable/a100
mounted camera and head-mounted camera dataset respectively produced 6.5 h (about 690,000 frames) of images.

5.3. Characteristics

Various objects are handled in daily life. Therefore, for ADL recognition, it is important to be able to recognize them in diverse environments. For this study, we asked users to take videos in their own homes. As shown in the examples in Figure 5, the environments caught on camera differ depending on the user. More examples are shown in the Supplemental Materials.

6. Experiments

6.1. Experiment protocols

We used 16-layer VGG net [27] pre-trained on the ImageNet 2012 dataset [6] for the CNN architecture in the same manner as the LCD [36]. Motion features are also important in action recognition. Therefore, we evaluated our dataset not only with CNN descriptors but also with iDT [34]. Following [34], we reduced the dimensions of the descriptors (HOG, HOF, and MBH) by a factor of 2 with PCA and encode them with FV, where the component number of the Gaussian mixture model was 256. We applied power and L2 normalization to aggregated vectors. As a classifier, we used a one-vs.-all SVM with linear kernel, setting $C = 100$. We used leave-one-user-out cross-validation for the evaluation so that the same person does not appear across both training and test data for ADL recognition. The iteration number of our methods is fixed at five because it usually converges in a few iterations.

6.2. Evaluating DSTAR and our dataset

We evaluated our approach on our wrist-mounted camera dataset (WCD) and head-mounted camera dataset (HCD). For fair comparison, we reduced the LCD dimensions from 512-D to a various range of dimensions such as 64-D, 128-D, and 256-D with PCA, and encoded them with various numbers of centers $K$ in VLAD such as $K = 64, 128, 256, 512, 1024$ as in [36] to find the best ones. We also explored the best choice of dimensions, $K$, $N_{sp}$, and $N_{tmp}$, for our method. We describe the best parameters and how they are determined in the Supplemental Materials because of the limited space here.

<table>
<thead>
<tr>
<th>Video Features</th>
<th>WCD</th>
<th>HCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCD+VLAD [36]</td>
<td>78.6</td>
<td>62.4</td>
</tr>
<tr>
<td>LCDsp+VLAD [36]</td>
<td>73.4</td>
<td>51.3</td>
</tr>
<tr>
<td>DSAR (ours)</td>
<td>82.0</td>
<td>61.6</td>
</tr>
<tr>
<td>DSTAR (ours)</td>
<td>83.7</td>
<td>62.0</td>
</tr>
<tr>
<td>STAR*</td>
<td>77.0</td>
<td>53.5</td>
</tr>
</tbody>
</table>

Table 2. Mean classification accuracy of the proposed methods on the wrist-mounted camera dataset (WCD) and the head-mounted camera dataset (HCD). STAR* is the method without weight optimization, which is equivalent to a spatiotemporal pyramid [17].
6.3. Applicability in existing datasets

Table 3 shows how wide our methods can be applied. We first evaluated LCD and our methods on UCIADL [26]. As shown in the table, our methods did not improve the performance since this dataset has low spatial bias. Moreover, we evaluated them on UCF101 [28], which is one of the representative datasets of typical action recognition. Although not as strong as our wrist-mounted dataset, this dataset has substantial bias. Therefore, DST AR showed better performance on this dataset than LCD did. Details are shown in the Supplemental Materials.

6.4. Fusing motion features and cameras

From Tables 2 and 4, we can confirm that iDT features are less discriminative than CNN-based features on WCD.

Comparing the iDT on both cameras, we found that iDT on HCD showed better performance than on WCD unlike the CNN descriptors. We can ascertain this reason from Figure 6. As the figure shows, iDT failed to remove the backgrounds from the video captured by a wrist-mounted camera compared with a head-mounted camera because of the large motion of the camera. Therefore, iDT on HCD is superior to that on WCD.

Jain et al. [11] showed that combining object features extracted by CNN with motion features such as iDT boosts action classification accuracy. Following their conclusion, we also demonstrate how our method was affected by the combination of motion features. We fused our methods with iDT on each dataset by simply averaging the score obtained using our methods and the mean score obtained by all iDT scores. As Table 4 shows, the performance of motion features was boosted by our methods more than by LCD. Although iDT features were more discriminative on HCD than on WCD, the combined features showed better performance on WCD than on HCD. Unlike action recognition, object features are more effective than motion features in ADL recognition because the critical key is the handled object. Therefore, we can find that wrist-mounted cameras are more suitable for ADL recognition than head-mounted cameras.

In case the user wears both a head-mounted camera and a wrist-mounted camera, we can choose superior information from wrist-mounted cameras and head-mounted cameras. Better object information is obtainable from wrist-mounted cameras, but better motion information is obtainable from head-mounted cameras. Therefore, we combined DST AR on WCD with iDT on HCD to achieve the best accuracy of 89.7% on our dataset.

7. Discussion

7.1. Visualized weights of DST AR

Figures 7 and 8 show the absolute values of spatial and temporal weights $W_{sp}$ and $W_{imp}$ calculated using DST AR on WCD. This figure presents the optimal discriminative weights for the respective cells.

**Spatial weight:** In Figure 7, it is apparent that cells near the center are important for classification, whereas cells on the right side are less important. The user’s palm always appears. No object appears on the right side of a wrist-
mounted camera image. Therefore, the right side area in the image has less information for recognition. The features obtained from the upper left cell and the bottom left cell are also less discriminative because backgrounds unrelated to the user’s action are often captured in these cells. However, handled objects often appear in the middle area of a wrist-mounted camera image. Discriminative features can be obtained from the cells of these areas.

**Temporal weight:** Although not as strong as the spatial bias of the image captured by a wrist-mounted camera, each ADL class has weak temporal bias. As Figure 8 shows, although full-length features (level 0) are the most important, temporally cut features (levels 1 and 2) have different weights. Using temporally cut pyramids improved the recognition performance, as presented in Table 2. Additionally, slight differences are apparent at the same level. At level 2, the beginning and the end of the action are slightly more important than the middle of the action.

7.2. Analysis of ADL classification results

We analyze the classification results here. Figure 9 presents the results of four different methods: LCD on HCD, LCD on WCD, DSTAR on WCD, and, finally, DSTAR on WCD and iDT on HCD.

**Comparing HCD with WCD:** We first compared both cameras using LCD. Results showed that 18 classes showed superior performance on WCD over HCD; these classes improved by 28.1% on average. Especially, “write on paper,” “cut paper,” and “staple paper” were improved significantly. These classes are actions wherein users use small objects such as a pen, scissors, and a stapler. A head-mounted camera captures these objects at a small scale. However, a wrist-mounted camera can capture large-scale images even of small objects. Four classes, however, showed inferior performance on WCD compared to HCD, and “dry dishes” is the class in which the classification accuracy declined considerably: 18.5%. On WCD, “dry dishes” was more often confused with “wash dishes” than on HCD, although all actions of “wash dishes” were recognized correctly.

**Comparing LCD with DSTAR:** Next, we compared DSTAR with LCD on WCD. Using DSTAR instead of LCD improved the accuracy on 14 classes. Its average improvement rate was 10.1%. One significantly improved action class was “vacuuming.” When we used a vacuum cleaner with a wrist-mounted camera, the floor and other unrelated backgrounds appeared on the left side of the wrist-mounted camera image. Actually, DSTAR was considered to improve the performance by reducing the importance of the features extracted from these areas. On five classes, the accuracy decreased, but the average rate of decrease was only 5.5%.

**Adding iDT on HCD with DSTAR on WCD:** Finally, we noticed how adding iDT affected HCD. Actually, 14 classes improved their performance by adding iDT on HCD; these classes improved by 10.4% on average. Especially, “wipe desk” was highly improved. Because the object information of “wipe desk” on HCD was often confused with other actions done near the table, using a wrist-mounted camera improved the performance. However, the “wipe desk” motion was distinctive. Consequently, adding iDT on HCD boosted the performance again. Only on five classes did the accuracy decline, the average of which was only 4.7%. In case the user wears both cameras, we can obtain better recognition.

8. Conclusion

This study examined the recognition of ADL with a wrist-mounted camera. We developed a publishable dataset of videos taken with a head-mounted camera and a wrist-mounted camera. Additionally, we proposed a novel video representation that aggregated CNN descriptors spatially and temporally, and optimized their weights both iteratively and alternately. Finally, using the proposed dataset, we quantitatively demonstrated the benefits of a wrist-mounted camera over a head-mounted camera and those of our proposed method over previous methods. We believe that our work will help spread the use of cameras attached to wrist-mounted devices.
References


