Anticipating Daily Intention using On-Wrist Motion Triggered Sensing

Tz-Ying Wu,* Ting-An Chien,* Cheng-Sheng Chan, Chan-Wei Hu, Min Sun
Dept. of Electrical Engineering,
National Tsing Hua University, Taiwan.
{gina9726, tingan0206, james121506, huchanwe1204, aliensunmin}@gmail.com

Abstract

Anticipating human intention by observing one’s actions has many applications. For instance, picking up a cellphone, then a charger (actions) implies that one wants to charge the cellphone (intention) (Fig. 1). By anticipating the intention, an intelligent system can guide the user to the closest power outlet. We propose an on-wrist motion triggered sensing system for anticipating daily intentions, where the on-wrist sensors help us to persistently observe one’s actions. The core of the system is a novel Recurrent Neural Network (RNN) and Policy Network (PN), where the RNN encodes visual and motion observation to anticipate intention, and the PN parsimoniously triggers the process of visual observation to reduce computation requirement. We jointly trained the whole network using policy gradient and cross-entropy loss. To evaluate, we collect the first daily “intention” dataset consisting of 2379 videos with 34 intentions and 164 unique action sequences (paths in Fig. 1). Our method achieves 92.68%, 90.85%, 97.56% accuracy on three users while processing only 29% of the visual observation on average.

1. Introduction

Thanks to the advance in Artificial Intelligence, many intelligent systems (e.g., Amazon Echo, Google Home) have become available on the markets. Despite their great ability to interact with humans through a speech interface, they are currently not good at proactively interacting with humans. Thus, we argue that the key for proactive interaction is to anticipate user’s intention by observing their actions. Given the anticipated intention, the intelligent system may provide service to facilitate the intention. More specifically, the ability to anticipate a large number of daily intentions will be the key to enable a proactive intelligent system.

Many researchers have tackled tasks related to intention anticipation. [11, 28, 18] focus on early activity prediction -- predicting actions before they have completed. However, the time-to-action-completion of this task is typically very short. Hence, there are only a few scenarios that intelligent systems may take advantage of the predicted activity. Kitani et al. [15] propose to forecast human’s trajectory. Forecasting trajectory is very useful, but it does not directly tell you the “intention” behind a trajectory. [4, 12, 13] anticipate the future events on the road such as making a left turn or involving in an accident. Although these events can be considered as intentions, only few intentions (at most five) are studied. Moreover, none of the work above leverages heterogeneous sensing modalities to reduce computing requirement.

In this work, we anticipate a variety of daily intentions (e.g., “go outside”, “charge cellphone”, in Fig. 1) by sensing motion and visual observation of actions. Our method is unique in several ways. Firstly, we focus on On-Wrist sensing: (1) an on-wrist camera (inspired by [24, 2]) is used to observe object interactions reliably, and (2) an on-wrist accelerometer is used to sense 3D hand motion efficiently. Since both on-wrist sensors are unconventional, we collect auxiliary object appearance and motion data to pre-train two

*indicates equal contribution
Our main contributions can be summarized as follows. (1) We adapt on-wrist sensors to reliably capture daily human actions. (2) We show that our policy network can effectively select the important images while only slightly sacrificing the anticipation accuracy. (3) We collected and will release one of the first daily intention datasets with a diverse set of action sequence and heterogeneous on-wrist sensory observations.

2. Related Work

We first describe works related to anticipation. Then, we mention other behavior analysis tasks. Finally, we describe a few works using wearable sensors for recognition.

2.1. Anticipation

The gist of anticipation is to predict in the future. We describe related works into groups as follows.

Early activity recognition. [11, 28, 18] focus on predicting activities before they are completed. For instance, recognizing a smile as early as the corners of the mouth curve up. Ryoo [28] introduces a probability model for early activity prediction. Hoai et al. [11] proposed a max-margin model to handle partial observation. Lan et al. [18] propose the hierarchical moves representation for predicting future activities.

Event anticipation. [17, 13, 12, 3, 33] anticipate events even before they appear. Jain et al. [13, 12] propose to fuse multiple visual sensors to anticipate the actions of a driver such as turning left or right. Fu et al. [3] further propose a dynamic soft-attention-based RNN model to anticipate accidents on the road captured in dashcam videos. Recently, Vondrick et al. [33] propose to learn temporal knowledge from unlabeled videos for anticipating actions and objects. However, the early activity recognition and anticipation approaches focus on activity categories and do not study risk assessment of objects and regions in videos. Bokhari and Kitani [11] propose to forecast long-term activities from a first-person perspective.

Intention anticipation. Intention has been explored more in the robotic community [35, 17, 16, 22]. Wang et al. [35] propose a latent variable model for inferring human intentions. Koppula and Saxena [17] address the problem by observing RGB-D data. A real robotic system has executed the proposed method to assist humans in daily tasks. [16, 22] also propose to anticipate human activities for improving human-robot collaboration. Hashimoto et al. [8] recently propose to sense intention in cooking tasks via the knowledge of access to objects. Recently, Rhinehart and Kitani [27] propose an on-line approach for first-person videos to anticipate intentions including where to go and what to acquire.

Others. Kitani et al. [15] propose to forecast human trajectory by surrounding physical environment (e.g., road, pavement). The paper shows that the forecasted trajectory can be used to improve object tracking accuracy. Yuen and Torralba [39] propose to predict motion from still images. Julian et al. [44] propose a novel visual appearance prediction method based on mid-level visual elements with temporal modeling methods.

Despite many related works, to the best of our knowledge, this is the first work in computer vision focusing on leveraging a heterogeneous sensing system to anticipate daily intentions with low computation requirement.

2.2. High-level Behavior Analysis

Other than activity recognition, there are a few high-level behavior analysis tasks. Joo et al. [14] propose to predict the persuasive motivation of the photographer who captured an image. Vondrick et al. [33] propose to infer the motivation of actions in an image by leveraging text. Recently, many methods (e.g., [38, 25, 26, 40, 32, 37]) have been proposed to generate sentence or paragraph to describe the behavior of humans in a video.
2.3. Recognition from Wearable Sensors

Most wearable sensors used in computer vision are first-person (i.e., ego-centric) cameras. [23, 31, 6, 19] propose to recognize activities. [21, 7] propose to summarize daily activities. Recently, two works [24, 1] focus on recognition using on-wrist camera and show that it outperforms ego-centric cameras. Inspired by them, we adapt a similar on-wrist sensor approach.

3. Our Approach

We first define the problem of intention anticipation. Next, we introduce our RNN model encoding sequential observations and fusing multiple sensors’ information from both hands. Then, we talk about our novel motion-triggered process based on a policy network. Finally, we describe how we pre-train the representation from auxiliary data.

3.1. Problem Formulation

**Observations.** At frame $t$, the camera observes an image $I_t$, and the motion sensor observes the 3D acceleration of hands $A_t \in R^3$.

**Representations.** The image $I$ and 3D acceleration $A$ are raw sensory values which are challenging to be used directly for intention anticipation, especially when lacking training data. Hence, we propose to learn visual object (referred to as object) $f_{o,t}$ and hand motion (referred to as motion) $f_{m,t}$ representations from other tasks with a larger number of training data. Note that, for all the variables, we use superscript to specify left or right hand (when needed). For instance, $f_{o,t}^L$ indicates left-hand object representation.

**Goal.** At frame $t$, our model predicts the future intention $y_t \in \mathcal{Y}$ based on the observations, where $\mathcal{Y}$ is the set of intention indices. Assuming the intention occurs at frame $T$, we not only want the prediction to be correct but also to predict as early as possible (i.e., $T - t$ to be large).

### 3.2. Our Recurrent and Fusion Model

Intention anticipation is a very challenging task. Intuitively, the order of observed objects and hand motions should be a very strong cue. However, most orders are not strict. Hence, learning composite orders from limited training data is critical.

**Recurrent Neural Network (RNN) for encoding.** We propose to use an RNN with two-layers of Long-Short-Term-Memory (LSTM) cell to handle the variation (Fig. 2-Top) as follows,

$$
g_t = \text{Emb}(W_{emb}, \text{con}(f_{m,t}, f_{o,t})) ,
$$

$$
h_t = \text{RNN}(g_t, h_{t-1}) ,
$$

$$
p_t = \text{Softmax}(W_y, h_t) ,
$$

$$
y_t = \arg \max_{y \in \mathcal{Y}} p_t(y) ,
$$

where $p_t \in R^{|\mathcal{Y}|}$ is the softmax probability of every intention in $\mathcal{Y}$, $W_y$ is the model parameter to be learned, $h_t$ is the learned hidden representation, and $g_t$ is a fixed dimension output of $\text{Emb}(\cdot)$. $W_{emb}$ is the parameter of embedding function $\text{Emb}(\cdot)$, $\text{con}(\cdot)$ is the concatenation operation, and $\text{Emb}(\cdot)$ is a linear mapping function (i.e.,
process of computing Eq. 7, when a CNN inference. When a and hidden state of RNN fation f a t to anticipate the correct intention as early as possible. Where our goal is to predict at π (the policy network, the policy
3.3. RL-based Policy Network
We propose a policy network π which is ideal for anticipating intentions.

3.4. Learning Representations from Auxiliary Data
Due to the limited daily intention data, we propose to use two auxiliary datasets (object interaction and hand motion) to pre-train two encoders: an object Convolutional Neural Network (CNN) and a hand motion 1D-CNN. In this way, we can learn a suitable representation of object and motion.

Object CNN. It is well-known that ImageNet [5] pre-trained CNN performs well on classifying a variety of objects. However, Chan et al. [2] show that images captured by on-wrist camera are significantly different from images in ImageNet. Hence, we propose to collect an auxiliary image dataset with 50 object categories captured by our on-wrist camera, and fine-tuned on ImageNet [5] pre-trained
3@150\times1

Figure 3. Illustration of our 1D-CNN pre-trained to classify six motions. Conv, MP, FC stand for Convolution, Max Pooling, and Fully Connected, respectively. 3@150 \times 1 denotes that there are three 150 \times 1 matrices. Since the second dimension is always one, it is a 1D-CNN. Our model has three stacks of Conv+MP layers and a FC layer at the end.

ResNet-based CNN [9]. After the model is pre-trained, we use the model to extract object representation \( f_o \) from the last layer before softmax.

**Hand motion 1D-CNN.** Our accelerometer captures acceleration in three axes \( s \in \mathbb{R}^3 \) with a sampling rate of 75Hz. We calibrate our sensor so that the acceleration in 3 axes are zero when we placed it on a flat and horizontal surface. We design a 1D-CNN to classify every 150 samples (2 seconds) into six motions: lift, pick up, put down, pull, stationary, and walking. The architecture of our model is shown in Fig. 3. Originally, we plan to mimic the model proposed by [4], which is a 3-layer 2D-CNN model with 1 input channel. Considering that there are no stationary properties among three acceleration values for each sample, we adjust the input channel number to 3 and define the 1D-CNN. For training the model, we have collected an auxiliary hand motion data with ground truth motions (Sec. 4). After the model is trained, we use the model to extract motion representation \( f_m \) at the FC4 layer (see Fig. 3).

3.5. Implementation Details

**Intention anticipation model.** We design our intention anticipation model to make a prediction in every half second. All of our models are trained using a constant learning rate 0.001 and 256 hidden states.

**Policy Network.** Our policy network is a neural network with two hidden layers. For joint training, we set learning rate 0.001, \( \lambda = 0.1 \) for joint loss. The reward of \( R^+ \) and \( R^- \) are 100 and -100, respectively.

**Object CNN.** Following the setting of [2], our object CNN aims at processing 6 fps on NVIDIA TX1. This frame rate is enough for daily actions. Since most of the actions will last a few seconds, it’s unnecessary to process at 15 or 30 fps. We take the average over 6 object representations as the input of our model. Different from [2], our on-wrist camera has a fish-eye lens to ensure a wide field-of-view capturing most objects. For fine-tuning the CNN model on our dataset, we set maximum iterations 20000, step-size 10000, momentum 0.9, every 10000 iteration weight decay 0.1, and learning rate 0.001. We also augment our dataset by horizontal flipping frames.

**Hand motion 1D-CNN.** Motion representation is extracted for a 2-second time segment. Hence, at every second, we process a 2-second time segment overlapped with previous processed time segment for 1 second. For training from scratch, we set the base learning rate to 0.01 with step-size 4000, momentum 0.9 and weight decay 0.0005. We adjust the base learning rate to 0.001 when fine-tuning.

4. Setting and Datasets

We introduce our setting of on-wrist sensors and describe details of our datasets.

4.1. Setting of On-wrist Sensors

Following similar settings in [24, 2], our on-wrist camera\(^2\) and accelerometer\(^3\) are mounted as shown in Fig. 4. Both camera and accelerometer are secured using velcro. We use the fisheye lens to ensure a wide field-of-view. We list some simple rules to be followed by users. First, the camera is under the arm, toward the palm. Second, the camera should roughly align the center of the wrist. This ensures that camera can easily record the state of the hand.

4.2. Datasets

We collect three datasets\(^4\) for the following purposes. (1) Daily Intention Dataset: for training our RNN model to anticipate intention before the intention occurs. (2) Object Interaction Dataset: for pre-training a better object interaction encoder to recognize common daily object categories. (3) Hand Motion Dataset: for pre-training a better motion encoder to recognize common motions.

\(^2\)fisheye lens mounted on noIR camera module with CSI interface.

\(^3\)MPU-6050.

\(^4\)Our dataset and code can be downloaded from [http://aliensunmin.github.io/project/intent-anticipate](http://aliensunmin.github.io/project/intent-anticipate)
4.2.1 Daily Intention Dataset

Inspired by Sigurdsson et al. [29], we select 34 daily intentions such as charge cellphone, go exercise, etc. Note that each intention is associated with at least one action sequence, and each action consists of a motion and an object (e.g., pick up+wallet). We propose two steps to collect various action sequences fulfilling 34 daily intentions.

**Exploring stage.** At this stage, we want to observe various ways to fulfill an intention (Fig. 1). Hence, we ask a user (referred to as user A) to perform each intention as different as possible. At this step, we observed 164 unique action sequences.

**Generalization stage.** At this stage, we ask user A and other users (referred to as user B and user C) to follow 164 action sequences and record multiple samples for each action sequence. This setting simulates when an intelligent system needs to serve other users. We show by experiment that our method performs similarly well on three users.

In Table 1, we summarize our intention dataset. Note that the number of action sequences recorded by user A is much more than others. Since we will train and validate on user A, selecting the proper hyper-parameters (e.g., design reward function). Next, we’ll apply the same setting to the training process of all users, and evaluate the result. This can exam the generalization of our methods. Design of reward function is described in the Sec.3 of the supplementary material.

<table>
<thead>
<tr>
<th>User</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td># of action sequences</td>
<td>1540</td>
<td>358</td>
<td>481</td>
</tr>
<tr>
<td>avg. per sequence</td>
<td>9.4</td>
<td>2.2</td>
<td>2.9</td>
</tr>
</tbody>
</table>

Table 1. Statistics of our intention dataset.

4.2.2 Object Interaction Dataset.

We select 54 object categories and collect a set of 940 videos corresponding to 909 unique object instances. Each video records how an object instance is interacted by a user’s hand. We sample 362 frames on average in each video. At the end, we collected an auxiliary dataset consisting of 340, 218 frames in total to pre-train our object encoder. Example frames of the dataset are shown in Fig. 6.

4.2.3 Hand Motion Dataset

Inspired by [4], we select six motions. We ask eight users to collect 609 motion sequences from the right hand and one user to collect 36 motion sequences from the left hand. For the right-hand data collected by eight users, we aim at testing cross users generalizability. For the left-hand data, we aim at testing cross hand generalizability.

5. Experiments

We first conduct pilot experiments to pre-train object and hand motion encoders. This helps us to select the appropriate encoders. Next, we conduct experiments for intention anticipation with policy network and evaluate our method in various settings. Finally, we show typical examples to highlight the properties of our method.
5.1. Preliminary Experiments

**Object pre-training.** We evaluate multiple Convolutional Neural Network (CNN) architectures on classifying 50 object categories in our object intention auxiliary dataset. These architectures include VGG-16 [30] and Residual Net (ResNet) [10] with 50, 101, 152-layers. We separate the whole dataset into two parts: 80% of object instances for training and 20% for testing. The testing accuracy is reported on Table 2. Our results show that deeper networks have slightly higher accuracy. Another critical consideration is the speed on the embedded device. Hence, we report the processed frames per second (fps) on NVIDIA TX1 in the last column of Table 2. Considering both accuracy and speed, we decide to use ResNet-50 since we designed our system to process at 6 fps similar to [2].

For hand motion, We describe two experiments to (1) select the best model generalizing across users, and (2) select the pre-processing step generalizing to the left hand.

**Generalizing across users.** Given our dataset collected by eight different users, we conduct a 4-fold cross validation experiment and report the average accuracy. We compare a recent deep-learning-based method [1] (1ch-3layer model) with our 3ch models trained from scratch in Table 2. The results show that our 3ch-3layer model generalizes the best across different users. At the end, we pre-train our 3-layer model on data collected by [20] to leverage more data. Then, we fine-tune the model on our auxiliary data.

**Generalizing across hands.** We propose the following pre-process to generalize our best model (3ch-3layer trained on right hand data) to handle left hand. We flip the left hand samples by negating all values in one channel (referred to as flip). This effectively flips left-hand samples to look similar to right-hand samples. In the last two rows of Table 3, we show the accuracy of left-hand data. Our method with flip pre-processing achieves better performance. In the intention anticipation experiment, we use “3ch-3layer” and apply flip pre-process on left hand.

### Table 2. Classification accuracy and processing speed of different models. We highlight the best performance using bold font.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Acc.</th>
<th>Testing Acc.</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG-16</td>
<td>98.58%</td>
<td>77.51%</td>
<td>4 fps</td>
</tr>
<tr>
<td>ResNet-50</td>
<td><strong>99.92%</strong></td>
<td>80.77%</td>
<td><strong>6 fps</strong></td>
</tr>
<tr>
<td>ResNet-101</td>
<td>97.45%</td>
<td>82.79%</td>
<td>4 fps</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>96.83%</td>
<td><strong>83.09%</strong></td>
<td>3 fps</td>
</tr>
</tbody>
</table>

**Table 3. Motion classification accuracy of different models. We highlight best performance using bold font.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Acc.</th>
<th>Testing Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ch-3layer</td>
<td>100.00%</td>
<td>81.41%</td>
</tr>
<tr>
<td>3ch-1layer</td>
<td>100.00%</td>
<td>77.21%</td>
</tr>
<tr>
<td>3ch-2layer</td>
<td>100.00%</td>
<td>78.37%</td>
</tr>
<tr>
<td>3ch-3layer</td>
<td>100.00%</td>
<td><strong>83.92%</strong></td>
</tr>
<tr>
<td>left</td>
<td>52.78%</td>
<td></td>
</tr>
<tr>
<td>left-flip</td>
<td><strong>83.33%</strong></td>
<td></td>
</tr>
</tbody>
</table>

5.2. Motion Triggered Intention Anticipation

For intention anticipation, we evaluate different settings on all three users. In the following, we first introduce our setting variants and the evaluation metric. Then, we compare their performance in different levels of anticipation (e.g., observing only the beginning X percent of the action sequence).

**Setting variants.**

1. **Object-only (OO):** RNN considering only object representation $f_o$.
2. **Motion-only (MO):** RNN considering only motion representation $f_m$.
3. **Concatenation (Con.):** RNN considering both object $f_o$ and motion $f_m$ representations.
4. **Motion-Triggered (MTr):** RNN with policy network, where the input of RNN is determined by the policy network. In this setting, we also report the ratio of triggered moments (referred to as Ratio). The lower the ratio the lower the computation requirement.

**Metric.** We report the intention prediction accuracy when observing only the beginning 25%, 50%, 75%, or 100% of the action sequence in a video.

Comparisons of different variants on all users (referred to as user A, B, and C) are shown in Table 4. We summarize our findings below. Object-only (OO) outperforms Motion-only (MO). This proves that object representation is much more influential than motion representation for intention anticipation. We also found that concatenating motion and object (Con.) does not consistently outperform Object-only (OO). Despite the inferior performance of MO, the tendency of MO under different percentage of observation is pretty steady. This implies that there are still some useful information in the motion representation. Indeed, MTr can take advantage of motion observation to reduce the cost of processing visual observation to nearly 29% while maintaining a high anticipation accuracy (92.6%, 90.85%, 97.56%).

In Fig. 8, we control the ratio of triggered moments and change the anticipation accuracy by adjusting the threshold of motion triggers. The results show that increasing the ratio of triggered moments leads to higher accuracy on intention anticipation. Most interesting, the accuracy only decrease slightly when the ratio is larger than 20%. Note that the default threshold is 0.5, which means the policy will decide to trigger when the probability of trigger is larger than non-trigger. Some quantitative results are described in Sec.4 of the supplementary material.
Table 4. Intention anticipation comparison. **OO** stands for object-only observation. **MO** stands for motion-only observation. **Con.** stands for concatenating \(f_o\) and \(f_m\). **Mtr.** stands for motion-triggered. **Ratio** stands for triggered ratio. In the second row, **25%** denotes only the beginning 25% of the action sequence is observed. All methods are evaluated on A, B, and C users. Note that **Mtr.** is significantly better than **MO** and only slightly worse than **Con.** while processing only about 29% of the frames.

<table>
<thead>
<tr>
<th></th>
<th>User A</th>
<th></th>
<th>User B</th>
<th></th>
<th>User C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25%</td>
<td>50%</td>
<td>75%</td>
<td>100%</td>
<td>25%</td>
</tr>
<tr>
<td>Con.</td>
<td>88.41%</td>
<td>90.24%</td>
<td>92.68%</td>
<td>94.51%</td>
<td>95.12%</td>
</tr>
<tr>
<td>OO</td>
<td>89.63%</td>
<td>92.68%</td>
<td>94.51%</td>
<td>95.73%</td>
<td>90.85%</td>
</tr>
<tr>
<td>MO</td>
<td>65.85%</td>
<td>70.73%</td>
<td>75.61%</td>
<td>75.61%</td>
<td>62.20%</td>
</tr>
<tr>
<td>Mtr.</td>
<td>86.58%</td>
<td>90.24%</td>
<td>92.68%</td>
<td>92.68%</td>
<td>84.75%</td>
</tr>
<tr>
<td>Ratio</td>
<td>34.00%</td>
<td>28.42%</td>
<td>30.72%</td>
<td>32.34%</td>
<td>33.23%</td>
</tr>
</tbody>
</table>

5.3. Typical Examples

We show typical examples in Fig. 7. In the first example, our Policy Network (PN) efficiently peeks at various objects (e.g., keys, cellphone, backpack, etc.). In other examples, PN no longer triggers after some early peeks. Specifically, in the second example, once the cellphone is observed and the wire is plugged in, PN is confident enough to anticipate cellphone charging without any further triggered operation.

6. Conclusion

We propose an on-wrist motion triggered sensing system for anticipating daily intentions. The core of the system is a novel RNN and policy networks jointly trained using policy gradient and cross-entropy loss to anticipate intention as early as possible. On our newly collected daily intention dataset with three users, our method achieves impressive anticipation accuracy while processing only 29% of the visual observation. In the future, we would like to develop an on-line learning based method for intention anticipation in the wild.

7. Acknowledgement

We thank MOST 104-2221-E-007-089-MY2 and MediaTek for their support.
References


