Toyota Smarthome: Real-World Activities of Daily Living

Srijan Das¹,², Rui Dai¹,², Michal Koperski¹,², Luca Minciullo³, Lorenzo Garattoni³, Francois Bremond¹,² and Gianpiero Francesca³

¹Université Côte d’Azur  ²Inria  ³Toyota Motor Europe

Abstract

The performance of deep neural networks is strongly influenced by the quantity and quality of annotated data. Most of the large activity recognition datasets consist of data sourced from the web, which does not reflect challenges that exist in activities of daily living. In this paper, we introduce a large real-world video dataset for activities of daily living: Toyota Smarthome. The dataset consists of 16K RGB+D clips of 31 activity classes, performed by seniors in a smarthome. Unlike previous datasets, videos were fully unscripted. As a result, the dataset poses several challenges: high intra-class variation, high class imbalance, simple and composite activities, and activities with similar motion and variable duration. Activities were annotated with both coarse and fine-grained labels. These characteristics differentiate Toyota Smarthome from other datasets for activity recognition. As recent activity recognition approaches fail to address the challenges posed by Toyota Smarthome, we present a novel activity recognition method with attention mechanism. We propose a pose driven spatio-temporal attention mechanism through 3D ConvNets. We show that our novel method outperforms state-of-the-art methods on benchmark datasets, as well as on the Toyota Smarthome dataset. We release the dataset for research use¹.

1. Introduction

Recent studies show that improvements in recognition methods are often paired with the availability of annotated data. For instance, significant boosts in image recognition accuracy on the AlexNet and VGG architectures [20, 37] were possible thanks to ImageNet [9] dataset. Similarly, the Inflated 3D Convolutional Networks (I3D) for activity recognition [4] largely benefited from the Kinetics [4] dataset.

Most of the available activity recognition datasets such as UCF101 [39], HMDB51 [21], Kinetics [4] are gathered from video web services (e.g. YouTube). Such datasets introduce data bias as they mainly contain activities concerning sports, outdoor activities and playing instruments. In addition, these activities have a significant inter-class variance (e.g. bike riding vs. sword exercising), which usually does not characterize daily living activities. Besides, most video clips only last a few seconds.

ADL datasets proposed in the past years [33, 46, 45, 40] were typically recorded using static cameras from a single viewpoint. The activities were performed in front of cameras by actors (often voluntary students), who were instructed beforehand. As a consequence, activities were performed in a similar, somewhat unnatural way. Finally, most of the datasets do not include complex, composite activities as they focus only on short, atomic motions. Table 1 provides a list of the most popular ADL datasets, outlining their key features along with the limitations mentioned above.

We introduce a new dataset that aims at addressing these limitations: Toyota Smarthome. Toyota Smarthome, hereafter Smarthome, contains approx. 16.1K video clips with 31 activity classes performed by 18 subjects. The challenges of this dataset are characterized by the rich diversity of activity categories performed in a real-world domestic environment. The dataset contains fine-grained activities (e.g. drinking with a cup, bottle or a can) and composite activities (e.g. cooking). The activities were recorded in 3 different scenes from 7 camera viewpoints. Real-world challenges comprise occlusion and high intra-class variation. Another unique feature of Smarthome is that activities are performed by subjects who did not receive any information about how to perform them.

To address the real-world challenges in Smarthome, we propose a novel attention mechanism on top of currently high-performing spatio-temporal convolutional networks [4] (3D ConvNet). Inspired by [11], our method uses both spatial and temporal attention mechanisms. We disassociate the spatial and temporal attention mechanisms (instead
of coupling them). In our architecture, two sub-networks independently regress the attention weights, based on 3D human skeletons inputs. The proposed attention mechanism aims at addressing the diversity of activity categories present in Smarthome. On the one hand, activities with human-object interaction require spatial attention to encode the information on the object involved in the activity. On the other hand, activities with temporal dynamics such as sitting or standing up require temporal attention to focus on the key frames that characterize the motion. The proposed method achieves state-of-the-art results on Smarthome and two public datasets: large-scale NTU-RGB+D [33] and a human-object interaction dataset - Northwestern-UCLA [46].

2. Related work

In this section, we briefly review publicly available daily living activity datasets and state-of-the-art activity recognition algorithms, focusing on attention mechanisms.

2.1. ADL real-world datasets

To deploy activity recognition algorithms on real-world sites, a validation on videos replicating real-world challenges is crucial. To well comprehend the limitations of currently-available datasets, we identify a set of indicators of how well each of these datasets addresses the main real-world challenges. **Context**: The context is the background information of the video. Some activity datasets feature a rich variety of contextual information (context biased). In some cases, the contextual information is so rich that it is sufficient on its own to recognize activities. For instance, in UCF and kinetics, processing the part of the frames around the human is often sufficient to recognize the activities. On the other hand, in datasets recorded in environments with similar backgrounds (context free), the contextual information is lower and thus cannot be used on its own for activity recognition. This is true, for instance, for datasets recorded indoor such as Smarthome and NTU RGB+D [33]. **Spontaneous acting**: This denotes whether the subjects tend to overstate movements following a guided script (low spontaneous acting). Subjects acting freely a loose script tend to perform activities spontaneously in a natural way (high spontaneous acting). **Camera framing**: This describes how the video has been recorded. Internet videos are recorded by a cameraman (high camera framing) and thus capture the subject performing the activity centered within videos and facing the camera. In contrast to this, real-world videos with fixed cameras (low camera framing) capture activities in an unconstrained field of view. **Cross-view challenge**: In real-world applications, a scene may be recorded from multiple angles. As activities can look different from different angles, activity recognition algorithms should be robust to multi-view scenarios. We therefore indicate which of the datasets pose the cross-view challenge. **Duration variation**: The duration of activities may vary greatly both inter-class and intra-class. A high variation of duration is more challenging and more representative of the real-world. We assign high duration variation to datasets in which the length of video samples varies by more than 1 minute within a class; low duration variation
otherwise. **Composite activities:** Some complex activities can be split into sub-activities (e.g., cooking is composed of cutting, stirring, using stove, etc.). This indicator simply states whether the dataset contains composite activities and their sub-activities. **Fine-grained activities:** Recognizing both coarse and fine-grained activities is often needed for real-world applications. For example, *drinking* is a coarse activity with fine-grained details of the object involved in it, say *can*, *cup*, or *bottle*.

Table 1 shows the comparison of the publicly available real-world activity datasets based on the above indicators.

ADL are usually carried out indoor, resulting in low context information. NTU-RGB+D [33] is one of the largest dataset for ADL, comprising more than 55K samples with multi-view settings. However, NTU-RGB+D was recorded in laboratory rooms and the activities are performed by actors with strict guidance. This results in guided activities and actors facing the cameras. MPII Cooking 2 [31] is an ADL dataset recorded for cooking recipes in an equipped kitchen. The dataset has 8 camera views, with composite activities. This dataset focuses on one cooking place, thus limiting the spatial context and the diversity of activity classes. Charades [35] and Something-Something [14] were recorded by hundreds of people in their own home with very fine-grained activity labels. However, self-recorded activities are very short (10 seconds/activity), often not natural, and always performed facing the camera. Hence, current ADL datasets address only partially the challenges of real-world scenarios. This motivates us to propose Smarthome: a dataset recorded in a semi-controlled environment and real-world settings. Here we summarize the key characteristics of Smarthome: (1) The dataset was recorded in a real apartment using 7 Kinect sensors [49] monitoring 3 scenes: dining room, living room and kitchen. Subjects were recorded for an entire day, during which they performed typical daily activities without any script. (3) Activity duration ranges from a couple of seconds to a few minutes. (4) As the camera positions were fixed, the camera-to-subject distance varies considerably between videos. (5) Sub-activity labels are available for composite activities such as *cooking, make coffee*, etc. Our annotations include fine-grained labels together with the coarse activity performed using different objects (e.g., *drink from cup, drink from can, and drink from bottle*).

### 2.2. ADL recognition methods

A large variety of algorithms have been proposed for ADL datasets. For a long time, activity recognition was dominated by approaches using local features, like dense trajectories [43, 44], combined with fisher vector encoding [27]. These approaches are simple and effective on small datasets. To tackle large datasets, researchers usually concatenate local features with those learned from convolutional networks [5, 36, 10]. A common issue with these popular deep learning approaches for activity recognition, such as Two-stream ConvNets [36], is the difficulty of encoding long-range temporal information. As a possible solution, Donahue et al. in [10] extracted spatial features from CNN network to feed sequential networks (LSTM). It was later shown that even when fed with large and sparse CNN features, sequential networks fail to learn the temporal dynamics [8]. Thus, sequential networks are often fed with 3D pose information [48, 33] to model the body dynamics of the subject performing the activity. However, 3D pose information by itself is not sufficient to encode context information such as objects involved in the activities. Spatio-temporal convolutional operations [41] have been used for activity recognition of large scale internet videos. These spatio-temporal operations are inflated from 2D kernels (I3D), pre-trained on ImageNet [9] and Kinetics [4] to recognize diverse activities with high accuracy [4, 12]. However, such 3D convNets do not exploit the salient part of the video. Recently, attention mechanisms on top of deep networks, such as LSTMs [38, 23] and I3D [47], have produced performance improvements.

Attention mechanisms focus on the salient part of the scene relative to the target activity. Attention mechanisms have gained popularity in the activity recognition commu-

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Context</th>
<th>Duration variation</th>
<th>Cross-view challenge</th>
<th>Composite activities</th>
<th>View Type</th>
<th>Spontaneous acting</th>
<th>Camera framing</th>
<th>Fine-grained activities</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTEV/VIRAT</td>
<td>free</td>
<td>Medium</td>
<td>Yes</td>
<td>No</td>
<td>Monitoring</td>
<td>Medium</td>
<td>Low</td>
<td>No</td>
<td>Surveillance</td>
</tr>
<tr>
<td>SVW [32]</td>
<td>biased</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Shooting</td>
<td>High</td>
<td>High</td>
<td>No</td>
<td>Sport</td>
</tr>
<tr>
<td>HMDDB [21]</td>
<td>biased</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Shooting</td>
<td>Medium</td>
<td>Medium</td>
<td>No</td>
<td>YouTube</td>
</tr>
<tr>
<td>Kinetics [4]</td>
<td>biased</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Shooting</td>
<td>Medium</td>
<td>Medium</td>
<td>No</td>
<td>Youtube</td>
</tr>
<tr>
<td>AVA [15]</td>
<td>biased</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Shooting</td>
<td>Medium</td>
<td>Medium</td>
<td>No</td>
<td>Movies</td>
</tr>
<tr>
<td>EPIC-KITCHENS</td>
<td>free</td>
<td>High</td>
<td>No</td>
<td>Yes</td>
<td>Egocentric</td>
<td>Medium</td>
<td>High</td>
<td>Yes</td>
<td>Kitchen</td>
</tr>
<tr>
<td>Something-Something [14]</td>
<td>free</td>
<td>Low</td>
<td>No</td>
<td>No</td>
<td>Shooting</td>
<td>Low</td>
<td>High</td>
<td>Yes</td>
<td>Object interaction</td>
</tr>
<tr>
<td>MPII Cooking2 [31]</td>
<td>free</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>Monitoring</td>
<td>Medium</td>
<td>Medium</td>
<td>Yes</td>
<td>Cooking</td>
</tr>
<tr>
<td>DAHLLA [42]</td>
<td>free</td>
<td>High</td>
<td>Yes</td>
<td>No</td>
<td>Monitoring</td>
<td>Medium</td>
<td>Medium</td>
<td>No</td>
<td>Kitchen</td>
</tr>
<tr>
<td>NUCLA [46]</td>
<td>free</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>Shooting</td>
<td>Low</td>
<td>High</td>
<td>No</td>
<td>Object interaction</td>
</tr>
<tr>
<td>NTU RGB-D [33]</td>
<td>free</td>
<td>Low</td>
<td>Yes</td>
<td>No</td>
<td>Monitoring</td>
<td>Low</td>
<td>High</td>
<td>No</td>
<td>ADL</td>
</tr>
<tr>
<td>Charades [35]</td>
<td>free</td>
<td>Low</td>
<td>Yes</td>
<td>Yes</td>
<td>Shooting</td>
<td>Low</td>
<td>High</td>
<td>Yes</td>
<td>ADL</td>
</tr>
<tr>
<td>Smarthome</td>
<td>free</td>
<td>High</td>
<td>Yes</td>
<td>Yes</td>
<td>Monitoring</td>
<td>High</td>
<td>Low</td>
<td>Yes</td>
<td>ADL</td>
</tr>
</tbody>
</table>
nity \[38, 1, 3\]. Sharma et al. \[34\] proposed an attention mechanism on RGB data where spatial attention weights are assigned to different parts of the convolutional feature map extracted from CNN. Liu et al. \[38\], and Baradel et al. \[1\] extended the aforementioned attention mechanism for both spatial and temporal attention on either 3D joint coordinates or RGB hand patches. Here, the pose driven spatial attention selectively focuses on the pertinent joints or RGB patches, while the temporal attention focuses on the key frames. All these methods \[1, 38, 3, 2\] use spatio-temporal attention for optimizing features computed by RNNs. As discussed earlier, the effectiveness of 3D ConvNets w.r.t. RNNs inspired us to use 3D ConvNets for our spatio-temporal attention mechanism.

Recently, some approaches using high-level I3D features have been proposed \[15, 13\]. The spatio-temporal convolution is guided by object detections in order to focus on the salient part of the images. In \[47\], the authors proposed a module on top of I3D that computes the attention of each pixel as a weighted sum of the features of all pixels along the space-time volume. However, this module is extremely dependent on the appearance of the activity, i.e., pixel position within the space-time volume. As a result, it fails to recognize activities with similar appearance and low motion. Thus, a more robust and general attention mechanism that soft-weights the salient parts of the feature map is required. With this aim, we propose a novel separable spatio-temporal attention mechanism.

3. Toyota Smarthome dataset

Toyota Smarthome is a video dataset recorded in an apartment equipped with 7 Kinect v1 cameras. It contains 31 daily living activities and 18 subjects. The subjects, senior people in the age range 60-80 years old, were aware of the recording but they were unaware of the purpose of the study. Each subject was recorded for 8 hours in one day starting from the morning until the afternoon. To ensure unbiased activities, no script was provided to the subjects. The obtained videos were analyzed and 31 different activities were annotated. The videos were clipped per activity, resulting in a total of 16,115 video samples. The dataset has a resolution of 640 × 480 and offers 3 modalities: RGB + Depth + 3D skeleton. The 3D skeleton joints were extracted from RGB using LCR-Net \[30\]. For privacy-preserving reasons, the face of the subjects is blurred using tinyface detection method \[18\].

Challenges. The dataset encompasses the challenges of recognizing natural and diverse activities. First, as subjects did not follow a script but rather performed typical daily activities, the number of samples for different activities is imbalanced (fig. 2). Second, the camera-to-subject distance varies considerably between videos and sometimes subjects are occluded. Third, the dataset consists of a rich variety of activities with different levels of complexity. Sub-activity labels are available for composite activities such as cooking, make coffee, etc. Fourth, the coarse activity is assigned with fine-grained labels when performed using different objects (for instance, drink from cup, can, or bottle). Finally, the duration of activities varies significantly: from a couple of seconds (for instance, sit down) to a few minutes (for instance, read book or clean dishes). All these challenges make the recognition of activities in Smarthome a difficult task. Figure 1 gives a visual overview of the dataset.

3.1. Evaluation protocols

We define two protocols for activity classification evaluation on Smarthome: cross-subject and cross-view. For each criterion, we report the mean per-class accuracy.

Cross-subject evaluation In cross-subject (CS) evaluation, we split the 18 subjects into training and testing groups. In order to balance the number of videos for each category of activity in both training and testing, the training group consists of 11 subjects with IDs: 3, 4, 6, 7, 9, 12, 13, 15, 17, 19, 25. The remaining 7 subjects are reserved for testing.

Cross-view evaluation For cross-view evaluation we propose two protocols, CV1 and CV2, containing 19 activities \(^2\). Both protocols use camera 2 for testing and camera

\(^2\)Some activities could not be included as they do not appear in the considered cameras.
5 for validation.

For CV1, we pick all samples of camera 1 for training. Camera 1 and camera 2 are both recorded in the dining room, having activities being performed in the same scene from two different viewpoints. This protocol also allows us to verify the generalization of the recognition system as it provides a smaller, highly imbalanced training set.

For CV2, we take samples from all cameras: camera 1, 3, 4, 6, 7 for the training set. We select only the samples of the 19 activities as mentioned in the CV1 protocol.

4. Proposed method

To address ADL recognition challenges, we introduce a new pose driven attention mechanism on top of the 3D ConvNets [4]. The spatial and temporal saliency of human activities can be extracted from the time series representation of pose dynamics, which are described by the 3D joint coordinates of the human body.

4.1. Spatio-temporal representation of a video

The input of our model are successive crops of human body along the video and their 3D pose information. We focus on the pertinent regions of the spatio-temporal representation from 3D ConvNet, which is a 4-dimensional feature map. Starting from the input of 64 human-cropped frames from a video V, the spatio-temporal representation g is the feature map extracted from an intermediate layer of the 3D ConvNet I3D [4]. The intermediate layer we use is the one preceding the Global Average Pooling (GAP) of I3D. The resulting dimension of g is \( t \times m \times n \times c \), where \( t \) is time, \( m \times n \) is the spatial resolution and \( c \) is the channels.

We define two separate network branches, one for spatial and one for temporal attention (see fig. 3). These branches apply the corresponding attention mechanism to the input feature map \( g \) and output the modulated feature maps \( g_s \) (for spatial attention) and \( g_t \) (for temporal attention). \( g_s \) and \( g_t \) are processed by a GAP layer and then concatenated. Finally, the prediction is computed from the concatenated feature map via a \( 1 \times 1 \times 1 \) convolutional operation followed by a softmax activation function.

4.2. Separable spatio-temporal attention

In this section, we elaborate on our pose driven spatio-temporal attention mechanism shown in fig. 4. Coupling spatial and temporal attention is difficult for spatio-temporal 3D ConvNet features as the spatial attention should focus on the important parts of the image, and the temporal attention should focus on the pertinent segments of the video. As these processes are different, our idea is to dissociate them. We learn two distinct attention sets, one for spatial and one temporal weights. These weights are linearly multiplied with the feature map \( g \), to output the modulated feature maps \( g_s \) and \( g_t \).

We use 3D skeleton poses to compute the spatio-temporal attention weights. The inputs to the attention network are the feature vectors calculated by an RNN on the 3D poses. This RNN is a 3 layered stacked LSTM pre-trained on 3D joint coordinates for activity classification. The input is a full set of \( J \) joints per skeleton where the joint coordinates are in the form \( x = (x_1, ..., x_J) \) for \( x_j \in \mathbb{R}^3 \).

The attention network consists of two separated fully connected layers with \( \tanh \) squashing followed by fully connected layers that compute the spatial and temporal attention scores \( s_1 \) and \( s_2 \), respectively (see fig. 4). The scores \( s_1 \) and \( s_2 \) express the importance of the elements of the convolutional feature map \( g \) along space and time. These scores \( s_r \) (i.e., \( s_1 \) and \( s_2 \) for \( r = 1, 2 \)) can be formulated as:

\[
s_r = W_r \tanh(W_h r + b_h) + b_s
\]

where \( W_s, W_h \) are learnable parameters and \( b_s, b_h \) are the biases. \( h^c_r \) is the concatenated hidden state vector of all the timesteps from the stacked LSTM.
The attention weights for spatial ($\alpha$) and temporal ($\beta = \{\beta_1, \beta_2, ..., \beta_t\}$) domain are computed from the scores $s_1$ and $s_2$ as:

$$\alpha = \sigma(W_\sigma s_1 + b_\sigma); \quad \beta_k = \frac{\exp(s_2,k)}{\sum_{i=1}^n \exp(s_2,i)} \quad (2)$$

where $s_2 = \{s_{2,1}, s_{2,2}, ..., s_{2,t}\}$ is obtained from equation 1. Normalizing the high number of $m \times n$ spatial attention weights with softmax leads to extremely low values, which can hamper their effect. To avoid this, we use sigmoid activation as in [38]. This attention weights play the role of soft selection for $m \times n$ spatial elements of the convolutional feature map $g$.

Finally, the modulated feature maps with spatial and temporal attention ($g_s$ & $g_t$) are computed as

$$g_s = \text{reshape}(\alpha) \ast g; \quad g_t = \text{reshape}(\beta) \ast g \quad (3)$$

where $\text{reshape}(x)$ operation is performed to transform $x$ to match the dimension of the feature map $g$. The attention model is joint-trained with the 3D ConvNet.

### 4.3. Training jointly the attention network and 3D ConvNet

Unlike the existing attention networks for activity classification [38, 1], jointly training the separable spatiotemporal attention network and the 3D ConvNet is relatively straightforward. The training phase involves fine-tuning the 3D ConvNet without the attention branches for activity classification. Then, the attention network is jointly trained with the pre-trained 3D ConvNet. This ensures faster convergence as demonstrated in [3]. The 3D ConvNet along with the attention network is trained end-to-end with a regularized cross-entropy loss $L$ formulated as

$$L = L_C + \lambda_1 \sum_{j=1}^{m \times n} \left\| \alpha_j \right\|_2^2 + \lambda_2 \sum_{j=1}^{t} (1 - \beta_j)^2 \quad (4)$$

where $L_C$ is the cross-entropy loss for $C$ activity labels. $\lambda_1$ and $\lambda_2$ are the regularization parameters. The first regularization term is used to regularize the learned spatial attention weights $\alpha$ with the $l_2$ norm to avoid their explosion. The second regularization term forces the model to pay attention to all the segments in the feature map as it is prone to ignore some segments in the temporal dimension although they contribute in modeling activities. Hence, we impose a penalty $\beta_j \approx 1$.

### 5. Experiments

#### 5.1. Other datasets and settings

Along with Smarthome, we performed experiments on two popular human activity recognition datasets: NTU RGB+D Dataset [33] and Northwestern-UCLA Multiview activity 3D Dataset [46].

**NTU RGB+D Dataset (NTU)** - The NTU dataset was acquired with a Kinect v2 camera and consists of 56880 video samples with 60 activity classes. The activities were performed by 40 subjects and recorded from 80 viewpoints. For each frame, the dataset provides RGB, depth and a 25-joint skeleton of each subject in the frame [33]. We performed experiments on NTU using the two split protocols proposed in [33]: cross-subject (CS) and cross-view (CV).

**Northwestern-UCLA Multiview activity 3D Dataset (NUCLA)** - The NUCLA dataset was acquired simultaneously by three Kinect v1 cameras. The dataset consists of 1194 video samples with 10 activity classes. The activities were performed by 10 subjects, and recorded from the three viewpoints. As NTU, the dataset provides RGB, depth, and the human skeleton of the subjects in each frame. We performed experiments on NUCLA using the cross-view (CV) protocol proposed in [46]: we trained our model on samples from two camera views and tested on the samples from the remaining view. For instance, the notation $V_{1,2}^3$ indicates that we trained on samples from view 1 and 2, and tested on samples from view 3.

#### 5.2. Implementation details

**Training** - For separable spatio-temporal attention model, we initialize the I3D base network from the Kinetics-400 classification models. Data augmentation and training procedure for training the I3D on tracks of human body follow [4]. For training the pose driven attention model, we use three-layer stacked LSTM. Each LSTM layer consists of 512, 512 and 128 LSTM units for Smarthome, NTU and NUCLA respectively. Similar to [33], we clip the videos into sub-sequences of 30 (Smarthome), 20 (NTU) and 5 (NUCLA) frames and then sample sub-sequences to input to the LSTM. We use 50% dropout to avoid overfitting. We set $\lambda_1$ & $\lambda_2$ to 0.00001 for all the datasets. For training the entire network, we use Adam Optimizer [19] with an initial learning rate set to 0.001. We use mini-batches of size 16 on 4 GPUs. We sample 10% of the initial training set and use it for validation only, specifically for hyper-parameters optimization, and early stopping. For training the I3D base network for NUCLA, we used NTU pre-trained I3D and then fine-tuned on NUCLA.

**Testing** - Each test video is processed 3 times to extract the human centered crop and two corner crops around the human bounding box. This is to cover the fine detail of the activity, as in [12]. The final prediction is obtained by averaging the softmax scores.
5.3. Comparative study

Tables 2 & 3 show that our model achieves state-of-the-art results on both NTU and NUCLA. We argue that PEM [25], whose results are close to those obtained by our attention mechanism, uses saliency maps of pose estimation. However, these saliency maps can be noisy in case of occlusions, which occur often in Smarthome as well as in most real-world scenarios. On the contrary, our attention mechanism computes attention weights from poses, and the classification ultimately relies on the appearance cue. Our attention mechanism significantly improves the results on these datasets, especially on NTU, by focusing on people interaction and human-object interaction. An important requirement is the availability of a large number of training samples, which is an issue in NUCLA. For this reason, the improvement achieved by our attention mechanism on NUCLA is less significant.

Smarthome consists of very diverse videos of activities performed with or without interactions with objects. Existing state-of-the-art methods fail to address all the challenges posed by Smarthome (see Table 4). The dense trajectories (DT) [43] obtain competitive results for actions with relatively high motion. However, dense trajectories are local motion based features and thus fails to model actions with fine-grained details and to incorporate view-invariance in recognizing activities. LSTM, fed with informative 3D joints, model the coarse activities based on body dynamics of the subject performing the activity, but fails to discriminate fine-grained activities due to the lack of object encoding.

Recent inflated convolutions [4] have shown significant improvement compared to RNNs. As a comparative baseline with our proposed spatio-temporal attention method, we have plugged a non-local module [47] on top of I3D. The non-local behavior along space-time in Smarthome is not view-invariant because its attention mechanism relies on appearance. On the contrary, our proposed attention mechanism is guided by 3D pose information, which is view-invariant. The significant improvement of our separable STA on cross-view protocols shows its view-invariant property compared to existing methods. In fig. 5 we provide some visual example in which our proposed approach outperforms I3D (without attention).

5.4. Other strategies for attention mechanism

Table 5 evaluates other strategies to implement the proposed attention mechanism. Among the strategies we included the implementation of single attention mechanisms (spatial or temporal) and all the different ways to combine them. The strategies included in the study are: I3D base network with (1) no attention (No Att); (2) only $m \times n$ dimensional spatial attention (SA); (3) only $t$ dimensional temporal attention (TA); (4) temporal attention applied after SA (SA+TA); (5) spatial attention applied after TA (TA+SA); and with (6) $m \times n \times t$ spatio-temporal attention at one go from pose driven model (joint STA). For the implementation of SA+TA and TA+SA, we adopt the joint training of SA+TA and TA+SA, we adopt the joint training mechanism proposed in [38]. Our proposed separable STA outperforms all other strategies by a significant margin. It is interesting to note that, unlike in RNNs [38, 1, 2], coupling spatial and temporal attention in 3D ConvNets decreases the classification accuracy. The reason for this can be seen from the classification accuracy achieved by SA and TA separately on the different datasets. In Smarthome and NUCLA, spatial attention is much more effective than temporal attention because several activities of both datasets involve interactions with objects. On the other hand, NTU contains activities with substantial motion (such as kicking, punching) and human-object interaction. Therefore, both spatial and temporal attention contribute to improve the classification accuracy. However, the possibility for the second attention to significantly modify the 3D feature maps is limited once the first attention has modified it. For this reason, we believe that dissociating both attention mechanisms is more effective than coupling them in series.

5.5. Ablation study

Figure 6 compares I3D base network with or without separable STA. The comparison is based on the per-class accuracy on Smarthome and NTU-CS (cross-subject protocol). Our separable STA improves I3D’s accuracy by an average of 4.7% on Smarthome and 6.7% on NTU. For Smarthome, the spatial attention alone contributes to a large improvement due to the ability to recognize fine-grained activities involving interactions with objects, such as Pourfromkettle (+21.4%) for CS and Use laptop (+13.4%), Eat.snack (42.8%) for CV. The temporal attention improves the classification of activities with low and high motion. Examples of this are static activities such as Watch TV (+8.8%) for CS and Readbook (+9.6%) for CV; and dynamic activities such as sidetown (+22.2%). For NTU-CS, the largest accuracy gains are observed for brushing...
Table 2. Results on NTU RGB+D dataset with cross-subject (CS) and cross-view (CV) settings (accuracies in %); Att indicates attention mechanism; □ indicates that the modality has been used only in training.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Pose</th>
<th>RGB</th>
<th>Att</th>
<th>CS</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>STA-LSTM [38]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>73.2</td>
<td>81.2</td>
</tr>
<tr>
<td>TS-LSTM [22]</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>74.6</td>
<td>81.3</td>
</tr>
<tr>
<td>VA-LSTM [48]</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>79.4</td>
<td>87.6</td>
</tr>
<tr>
<td>STA-Hands [1]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>82.5</td>
<td>88.6</td>
</tr>
<tr>
<td>altered STA-Hands [2]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>84.8</td>
<td>90.6</td>
</tr>
<tr>
<td>Glimpse Cloud [3]</td>
<td>○</td>
<td>✓</td>
<td>✓</td>
<td>86.6</td>
<td>93.2</td>
</tr>
<tr>
<td>PEM [25]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>91.7</td>
<td>95.2</td>
</tr>
<tr>
<td>Separable STA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>92.2</td>
<td>94.6</td>
</tr>
</tbody>
</table>

Table 3. Results on North-western-UCLA Multiview activity 3D dataset with cross-view $V_{CV}$2 settings along with indicating input data modalities (accuracies in %); $P_{Pose}$ indicates its usage only in the training phase.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Data</th>
<th>$V_{CV}$2</th>
</tr>
</thead>
<tbody>
<tr>
<td>STA-LSTM [38]</td>
<td>HPM+TM [29]</td>
<td>71.9</td>
</tr>
<tr>
<td>TS-LSTM [22]</td>
<td>HBRNN [17]</td>
<td>78.5</td>
</tr>
<tr>
<td>Ensemble TS-LSTM [22]</td>
<td>Pose</td>
<td>89.2</td>
</tr>
<tr>
<td>nCTE [16]</td>
<td>RGB</td>
<td>75.8</td>
</tr>
<tr>
<td>NKTM [28]</td>
<td>RGB</td>
<td>85.6</td>
</tr>
<tr>
<td>Glimpse Cloud [3]</td>
<td>RGB+$P_{Pose}$</td>
<td>90.1</td>
</tr>
<tr>
<td>Separable STA</td>
<td>RGB+$P_{Pose}$</td>
<td>92.4</td>
</tr>
</tbody>
</table>

5.6. Runtime

Training the separable STA model end-to-end takes 5h over 4 GTX 1080 Ti GPUs on Smarthome in CS settings. Pre-training the I3D base network with RGB human crops and stacked LSTM with 3D poses takes 21h and 2h respectively. At test time, a single forward pass for a video takes 338ms on 4 GPUs.

6. Conclusion

In this paper, we introduced Toyota Smarthome, a dataset that poses several real-world challenges for ADL recognition. To address such challenges, we proposed a novel separable spatio-temporal attention model. This model outperforms state-of-the-art methods on Smarthome and other public datasets. Our comparative study showed that all tested methods achieve lower accuracy on Smarthome compared to the other datasets. We believe that this performance difference is due to the real-world challenges offered by Smarthome. For this reason, we release Toyota Smarthome to the research community. To learn more about Toyota Smarthome dataset please visit the project website3. As future work, we plan to integrate the additional challenge of recognizing activities in untrimmed video streams. This will correspond to a new version of Toyota Smarthome dataset.

Acknowledgement

The authors are grateful to Sophia Antipolis - Mediterranean "NEF" computation cluster for providing resources and support.

References


3https://project.inria.fr/toyotasmarthome
The British Machine Vision Conference (BMVC), September 2018.


[29] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In 2016 IEEE Conference on Co-


