# Chapter 1
# A Step Towards Automated Functional Assessment of Activities of Daily Living

Bappaditya Debnath and Mary O'brien and Swagat Kumar and Ardhendu Behera

**Abstract** Current activity recognition approaches have achieved a great success due to the advancement in deep learning and the availability of huge public benchmark datasets. These datasets focus on highly distinctive actions involving discriminative body movements, body-object and/or human-human interactions. However, in real-world scenarios, e.g., functional assessment of a rehabilitation task, which requires the capability of differentiating the execution of same activities performed by individuals with different impairments, their recognition accuracy is far from being satisfactory. To address this, we develop Functional-ADL, a challenging novel dataset to take action recognition to a new level. Compared to the existing datasets, Functional-ADL is distinguished in multi-label and impaired-specific executions of different Activities of Daily Living (ADL) to contribute towards vision-based automated assessment and rehabilitation of physically impaired persons. We also propose a novel pose-based two-stream multi-label activity recognition model consisting of a spatial and a temporal stream. The proposed approach significantly outperforms the state-of-the-art by a considerable margin. This new Functional-ADL dataset presents significant challenges for human activity recognition, and we hope this could advance research towards activity understanding and monitoring.

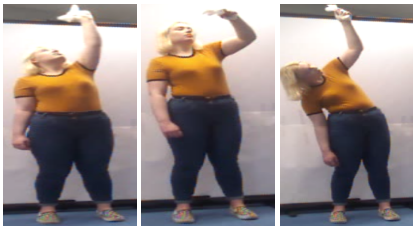
**Key words:** Physical Rehabilitation, Functional Activity Recognition, Computer Vision, Deep Learning, Body-pose Sequence, Fisher Vectors

## 1.1 Introduction

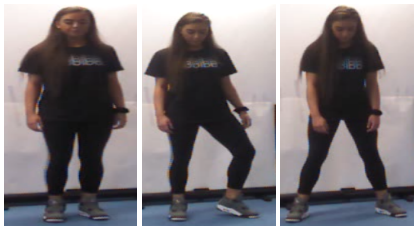
Activity recognition is an important and challenging problem in computer vision with many applications linking assistive and rehabilitative robotics for health and

Bappaditya Debnath and Mary O'brien and Swagat Kumar and Ardhendu Behera
Edge Hill University, St Helen's Road, Ormskirk, L394QP, UK, e-mail: {debnathb,obrienm,kumars,beheraa}@edgehill.ac.uk

social care services. This research aims to contribute towards this where there has been an increased interest in using vision-based human motion understanding for rehabilitation and assessment of physically impaired patients [27]. Physically impaired people (e.g., affected by stroke, spinal cord injury, etc.) often experience problems with physical movement and balance. As a result, they face difficulties in performing day to day tasks, known as Activities of Daily Living (ADL) [9]. To recover, improve or avoid further loss of physical functionality, such patients undergo physical rehabilitation programs [9] involving repetitive therapeutic exercises or ADL. These activities are usually guided by health professionals (clinicians, occupational therapists and physiotherapists) at home or in a clinic [9]. The assessment part of this rehabilitation process is often carried out via direct observation, which requires the observer to note down the detailed movements of the patients performing a given ADL. The process is time-consuming, laborious and often requires a significant attention from the observer. This process could be automated by using vision-based autonomous systems that can recognise and evaluate the difference between normal and impaired physical activities. This has the potential to lower cognitive load on the observers, time and overall cost.



(a) Reaching Above: Normal, Elbow Rigidity, and Shoulder Weakness



(b) Walking: Normal, Knee Rigidity, and Wider Gait

Fig. 1.1: We introduce a novel multi-label functional ADL dataset consisting of 10 activities and four different physical impairment-specific executions of each ADL for fine-grained activity recognition in rehabilitation videos. a) "Reaching Above" activity is executed by a normal person, an individual with 'Elbow Rigidity', and a person with "Shoulder Weakness" impairment (left to right). b) Similarly, "Walking" activity is performed 'Normally', with 'Knee Rigidity' and 'Wider Gait'.

Functional assessment through ADL is widely carried out for assessing a patient's condition and progress. To measure it, there exists various methods [12] that focus on complex manual assessment. The current study is only the first step towards automating functional assessment of various ADL. The main aim of this study is to recognise different ADL (e.g., eating, drinking, etc.) as well as their impairment-specific variations executed by individuals with physical impairments (e.g., ataxia, elbow rigidity, etc.). This will help in monitoring the extent of their progress and achievement while undergoing impairment-specific rehabilitation. In recent times, computer vision researchers have focused on Deep Learning (DL) [28, 1, 16] to

improve human activity recognition accuracy. This is possible due to the availability of large-scale datasets. However, for vision-based ADL assessment and rehabilitation, authors have mainly used their own small in-house datasets [27, 4, 25, 37]. Lack of suitable publicly available datasets is one of the reasons why there has been less participation from the vision community in developing models for solving such problem.

To address this, we present a novel multi-label functional ADL dataset that is targeted towards automated functional assessment of physically impaired persons through ADL. Physically impaired persons would perform an ADL differently from healthy individuals resulting in a different spatio-temporal pattern, which is dependent on the type of impairment they are suffering from. For example, a person having tremors would shake his/her hand while drinking water and the spatio-temporal trajectory would be different from a drinking action without tremors. The existing human activity datasets (Table 1.1) are not appropriate to develop and validate solutions targeting this issue. Thus, this study presents a dataset that consists of a normal and various physical impairment-specific executions of the same ADL. The proposed dataset contains 5685 samples of 10 common ADL performed by 10 subjects, captured in video, depth and human body-pose sequence format. For each ADL, the dataset presents one normal and four different physical impairment-specific executions. Thus, each sample has two labels, one for the 'Activity' (e.g., drinking, walking, etc.) and the other one for the 'Impairment' (e.g., normal, ataxic, etc.) and hence, the name multi-label functional ADL dataset. Furthermore, we present a novel pose-based two-stream multi-label activity recognition model based on TCN-ResNet [16] that comprehensively outperforms the TCN-ResNet on our dataset and the well-known NTU-RGBD ADL recognition dataset [28]. The two-stream architecture inspired by [6], which consists of a spatial and a temporal stream. The spatial stream contains a Spatial Encoding Unit (SEU), which provides an enriched representation that *learns to capture* the structural relationships between various body joints in a given video frame. Similarly, the temporal stream includes a Temporal Encoding Unit (TEU) that *learns to encode* the temporal relationship of each body joint over the duration of a given sequence. The performance of the network is further enhanced by the introduction of a Fisher Vector (FV) based activity-aware learn-able pooling mechanism introduced at the end of each stream to replace the Global Average Pooling (GAP) in the TCN-ResNet [16]. Our novel contributions are: 1) A novel functional ADL recognition dataset that presents a normal and four different physical impairment-specific versions of each ADL. 2) A pose-based (skeleton) two-stream functional ADL recognition model that integrates a spatial-temporal body-pose encoding mechanism with FV-based pooling in a novel manner.

## 1.2 Related Works

**Datasets**: Major advances have been made in human activity recognition influenced by the availability of large-scale datasets. Well-known datasets in this domain are

shown in Table 1.1. However, the existing datasets are largely targeted towards normal human activity recognition and are not suitable for functional assessment of physically impaired patients through ADL. These datasets consist different normal ADL but do not capture various physical impairment-specific versions of the same ADL (Table 1.1, Column 4). Most of the datasets in Table 1.1 are single-label datasets whereas we present a multi-label dataset where there are two labels ('Activity' and 'Impairment') for each sample. The Chardes [30] and the UA-Concurrent [38] datasets present multiple labels for a single activity sample but these are multi-label normal activities. The NTU-RGBD 120 dataset [21] presents 12 medical conditions including neck pain, fall, etc. However, it is a single-label dataset which does not demonstrated the difference between impairment-specific executions of the same ADL. To the best of our knowledge, this is the first dataset that illustrates the difference between various physical impairment-specific versions of the same ADL.

| Datasets | #Videos | #Activities | #Impairments | #Subjects | Data Modalities |
|---|---|---|---|---|---|
| MSRDailyActivity3D [35] | 320 | 16 | 0 | 10 | R,D,P |
| UTKinect [39] | 200 | 10 | 0 | 10 | R,D,P |
| MSR-Action3D [20] | 567 | 20 | 0 | 10 | R,D,P |
| CAD-60 [32] | 60 | 12 | 0 | 4 | R,D,P |
| CAD-120 [17] | 120 | 20 | 0 | 4 | R,D,P |
| Northwestern-UCLA [36] | 1475 | 10 | 0 | 10 | R,D,P |
| NTU-RGBD [28] | 60K | 60 | 0 | 40 | R,D,P |
| Chardes [30] | 10K | 157 | 0 | 267 | R |
| NTU-RGBd 120 [21] | 120K | 120 | 0 | 106 | R,D,P |
| Toyota Smart Home [5] | 16K | 51 | 0 | 18 | R,D,P |
| UA-Concurrent [38] | 201 | 35 | 0 | NA | R,D,P |
| Ours | 5865 | 10 | **8** | 10 | R,D,P |

Table 1.1: Comparison of the proposed dataset with other popular activity recognition datasets. The proposed multi-label functional ADL recognition dataset represents a normal and four different physical impairment-specific executions for each ADL. R: RGB, D: Depth, P: Pose

**Pose-based activity recognition**: The availability of cheap depth sensors (e.g., Microsoft Kinect) has significantly influenced pose-based activity recognition. Processing 3D pose (body skeleton) information is computationally much less expensive than RGB video processing and thus, researchers have increasingly relied on pose-based methods for human activity recognition. Most works in this area have explored recurrent networks such as RNN, LSTM and GRU which are specially designed for processing sequential information such as trajectory of human body joints in a given activity. [22] advanced the human tree-structure to model spatio-temporal features learned from a modified gating mechanism of LSTM. [31] introduced LSTM-based spatial and temporal networks with attention mechanism. Temporal Convolutional Network (TCN) which are stack of 1D convolutional layer have been explored as an alternate to recurrent mechanism. [18] proposed TCN with an Encoder-Decoder

and a dilated convolution model for activity recognition. An LSTM cell processes each time-step sequentially whereas no such constraint exists within TCN. This makes TCN inherently faster than LSTMs. Kim et al. [16] present a pose-based TCN-ResNet model which combines residual connections with TCN and shown to be computationally inexpensive without compromising the recognition accuracy on NTU-RGBD dataset [28]. The proposed model is inspired by this lightweight architecture which is necessary for home-based or in-clinic assessment of patients where high-performance computational facilities (e.g., servers, GPUs, etc.) are not available. However, pose-based models do not benefit from contextual cues such as hand-object interactions, background information other than body pose. Thus, authors have focused on enriching the pose-information with physics-based measurements such velocities and acceleration [7, 42], different normalisation techniques [42], relative body joints positions [15] etc. Instead of handcrafting such features, [6] uses SEU and TEU to automatically learn enhanced representations that can capture structural information and various inter-joint dependencies of the human body joints. Inspired by this, we adapt the TCN-ResNet model [16] to use a spatial-temporal architecture involving SEU and TEU layer to advance the light-weight human activity recognition approaches.

**Learnable-Pooling:** To further enhance the performance of our model, a FV-based pooling mechanism is used that replaces the GAP layer typically present towards the end of many standard convolutional architectures including TCN-ResNet [16]. Similar to GAP, the literature presents other statistical pooling methods like average or max-pooling [13, 14], rank-pooling [8], context-gating [24] and high-dimensional Feature encoding [40]. Pooling using statistical methods do not consider spatial-temporal and other semantic information in feature maps produced by CNN, TCN or LSTM. Thus, learn-able pooling methods have been explored by researchers to pool the most relevant features based on learned representations. In [10], authors present a second order attentional pooling, in which the output map from a CNN is multiplied with a weighted version of itself. A well-known technique called VLAD for image feature representation is integrated by Girdhar et al. [11] for learn-able pooling-based activity recognition. In [24], authors introduce learn-able FV (NetFV) to semantically cluster and pool audio and video features by integrating it to a deep model. In this study, NetFV is adapted for semantically clustering information present in TCN-ResNet maps which further enhances the model performance.

## 1.3 Dataset

Human motion manifests in a wide variety of forms and so does its abnormalities. It is not feasible to capture whole range of ADL and their corresponding impairments. The idea is to prepare a dataset that would meet the following constraints: 1) The dataset should contain enough samples uniformly spread across subjects, activities and impairments that would suffice the needs of DL-based models. 2) The dataset should contain enough activities that would collectively cover a wide range of body

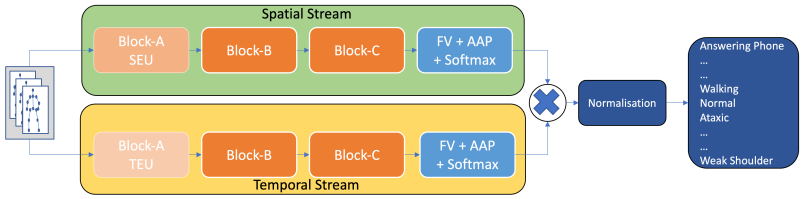



Fig. 1.2: The proposed model consists of a spatial and a temporal stream where each stream uses a TCN-ResNet [16]. Block-A of the spatial stream is replaced with the SEU [6] while the same block in temporal stream is exchanged with the TEU [6]. The GAP layer of the TCN-ResNet [16] is replaced by a FV-based pooling mechanism [24]. The Soft-max output of both the streams are multiplied (indicated by $\times$) and normalised for the final output. The model is trained through a multi-hot encoded label where in each label vector there are two '1's indicating 'Activity' and 'Impairment' label.

movements and capture a few common abnormalities. To assess a patient's condition and to determine their functional independence, clinicians often require them to perform ADL [12]. The initial idea was to capture patients performing these activities and annotate each action with an 'Activity' and an 'Impairment'. To create a dataset, one needs multiple samples for each annotation, ideally uniformly spread across number of subjects. It is very difficult to ask patients to perform multiple repetitions of each of the activities owing to their physical constraints. It is easy to see that a patient with a 'bent knee' would face difficulty in performing sit to stand multiple times and would not be able to provide a sample of a normal sit to stand sequence. Thus, to address this, the workaround was to film the activities with healthy subjects acting like patients. To make sure that activities performed by healthy subjects accurately reflect the performance of real patients, help was sought from an occupational therapist. Under the guidance of the occupational therapist the ADL were chosen in a manner that would collectively cover a wide range of body movements and test various parts of the musculo-skeletal system. Each ADL filmed for this dataset is captured in one healthy and four different physical impairment-specific executions of the same. The activities "Sitting", "Standing" and "Walking" cover lower torso and leg movements. "Drinking", "Brushing Hair" and "Wearing Glasses" test the functionality of upper limbs. "Brushing Floor", "Answering Phone" and "Clapping" are preformed while standing and thus they require close co-ordination between upper and lower halves of the body. The impairments 'Weakness to One Side', 'Knee Rigidity' and 'Wider Gait' are represented in all the lower-limb activities. The impairments 'Shoulder Weakness', 'Tremors' and 'Elbow Rigidity' are present for all the upper-limb activities. All the activities exhibit the 'Normal' and the 'Ataxic' versions. The dataset presents seven impairments in total while each of the 10 activity is represented through four different impairments in addition to a

regular healthy execution. Altogether there are 5685 samples, each annotated with an activity and an impairment, performed by 10 (5 female and 5 male) subjects.

## 1.4 Proposed Approach

There are many aspects of designing a pose-based model and the proposed model aims to address the following aspects: 1) Effectively capture the saptio-temporal information contained within human body-pose sequence. 2) Semantically cluster meaningful information represented by the body-pose network. The proposed model is based on the TCN-ResNet architecture [16], which is basically a combination of 1D convolutions with residual connections. We use two TCN-ResNet models, one for a spatial stream and another one is for a temporal stream. In the spatial stream, Block-A of the TCN-ResNet is replaced with SEU and similarly, TEU substitutes Block-A in the temporal stream (Fig. 1.2). As in [6], for each frame the SEU captures the structural relationship between various body joints and enhanced/augmented representation of the human body pose sequence to rest of the network. On the other hand, the goal of the TEU is to encode the frame-wise positions of body joints and present a temporally rich representation for each joint, individually. Furthermore, we introduce a novel FV-based learn-able pooling mechanism in each stream replacing the GAP layer in the original TCN-ResNet. This is mainly due to the fact that learn-able pooling approaches have shown to be more effective in pooling more relevant features instead of statistical pooling (e.g., average or max-pool). This has been further discussed in related works. This learn-able pooling method integrates FV-based clustering mechanism which semantically clusters the spatial and temporal structures contained within the respective streams. Thus, it has significantly improved the recognition accuracy as shown in the experimental evaluation section.

**TCN-ResNet:** The TCN-ResNet model is basically a stacking of 1D convolutional layers followed by the standard GAP + FC layers. As shown in Fig 1.2, the network is composed of three 1D convolutional blocks (Block-A, Block-B and Block-C) and each of the three block is composed of three layers of 1D convolutions. Each convolutional operation is followed by BN and a ReLU activation function. The convolutional operation at the start of Block-B and Block-C is of stride 2, which means the input is halved along the first dimension (normally time dimension) as it passes from Block-A to Block-B and then from Block-B to Block-C. There are two paths between any two layers: 1) First is through 1D convolutional operation followed by BN and a ReLU activation function; 2) Second, through a residual or skip-connection (omitted in Fig. 1.2 for simplicity). Let $T$ be number of frames in a sequence, $J$ the number of body joints, $D$ dimension of each joint (3 for 3D pose) and $F$ the total number of filters in a layer. Then, with input pose map $V \in \mathbf{R}^{T \times JD}$, 1D convolution operations in each block performs the following transformation:

$$BlockA : V_{T,J\times D} \rightarrow M^{a}_{T,F_a} \tag{1.1}$$

$$BlockB : M^{a}_{T,F_a} \rightarrow M^{b}_{T/2,F_b} \tag{1.2}$$

$$BlockC : M^{b}_{T/2,F_b} \rightarrow M^{c}_{T/4,F_c} \tag{1.3}$$

Here, $F_a = 64, F_b = 128, F_c = 256$ indicate the number of filters and $M^a$,$M^b$, $M^c$ imply the output maps of Block-A, Block-B and Block-C (Fig. 1.2), respectively. The output of Block-C is passed through a standard GAP layer followed by a FC layer with Soft-max activation function.

**Spatial Stream:** The spatial stream is a TCN-ResNet [16] model. The difference is, it adapts the Block-A (Fig. 1.2) to the SEU introduced by [6]. Normal 1D convolutions process all the frames in a body-pose sequence together as shown in Eq. 1.1. In contrast, the SEU processes each frame through separate and independent convolutional operations and then concatenate the outputs. The enables the network to learn relationships and dependencies between various body joints for each point in time. In other words, the network learns the body-structure spatially for each time-step for a given sequence. This structural learning is absent in case of TCN-ResNet which involves normal 1D convolutions. Formally the SEU in Block-A performs the following transformation [6]:

$$\hat{M}^{at}_{J,F_a} = \hat{U}_t(F_a, V_{J,D}) \tag{1.4}$$

$$\hat{M}^{at}_{J,F_a} \rightarrow \hat{M}^{a}_{T,J\times F_a} \tag{1.5}$$

where $\hat{U}$ is the convolution operation parameterised by filters $F_a$. $\hat{M}$ indicates a map for the spatial stream, which corresponds to $M$ for the TCN-ResNet (Eqs. 1.1-1.3). Normally, body-pose sequence is represented by a map where for each frame, the body-pose is represented by a vector of size $J \times D$. In TCN-ResNet [16], the convolution operation in Block-A transforms this vector into a vector of length $F_a$ (Eq. 1.1), which represents the body-pose for each frame transformed through 1D convolutions. In contrast, the transformation by SEU produces a body-pose that is represented through a vector of size $J \times F_a$ for each frame (Eq. 1.4). The SEU transforms joints in each frame separately to produces maps $\hat{M}^{at}_{J,F_a}$ for each individual frame (Eq. 1.4). These maps are aggregated to form the final SEU output (Eq. 1.5). Thus, instead of $F_a$ for all the joints in TCN-ResNet (Eq. 1.1), the SEU represents each joint by a vector of size $F_a$ (Eq. 1.5). Through this enhanced representation from $D$ co-ordinates (normally 3 for 3D pose) in input to $F_a$ (=64) at the output of Block-A (Fig. 1.2), in each frame, the SEU encodes the relationships and dependencies between various body joints that is *learnt* [6]. The SEU increases the number of parameters from SEU by a factor of number of joints (Eq. 1.5). Including more blocks (Block-B, Block-C) in the SEU, increases the number of parameters because in TCN-ResNet the filter count is doubled each time a block is traversed (Eqs. 1.1,1.2,1.3). Empirically it was observed that including more block for the SEU, made the model slower while having no positive impact on the performance. The output of Block-C in TCN-ResNet has the temporal dimension reduced by a fourth, as shown in Eq. 1.3. Thus from Eq. 1.3 (TCN-ResNet), the spatial stream

(Fig. 1.2) performs the following transformation:

$$V_{T,J\times D} -> \hat{M}^{c}_{T/4,F_c} \tag{1.6}$$

Note that that because of SEU, the Block-A (Eq 1.5) in the spatial stream produces different output than Block-A (Eq 1.1) in the TCN-ResNet. However, this does not make any difference in the overall output of the spatial stream at the end of Block-C which is determined by the number of filters at the end of Block-C and the number of frames (time-steps) in the body-pose sequence. Effectively, output of Block-C in spatial stream (Eq. 1.6) is same as output of Block-C in TCN-ResNet (Eq. 1.3).

**Temporal Stream:** Similar to the spatial stream, a TCN-ResNet [16] is used for the temporal stream. The first block (Block-A, Fig. 1.2) is used for TEU as done in [6]. Formally, the TEU performs the following transformation:

$$V_{T,J\times D} \rightarrow \bar{V}_{J\times D,T} \tag{1.7}$$

$$\bar{M}^{a}_{J\times D,F_a} = \bar{U}(F_a, \bar{V}_{J\times D,T}) \tag{1.8}$$

$$\bar{M}^{a}_{J\times D,F_a} \rightarrow \bar{M}^{a}_{F_a,J\times D} \tag{1.9}$$

Here, $\bar{U}$ is the convolution operation. $\bar{M}$ indicates maps for temporal stream corresponding to $M$ in TCN-ResNet (Eqs. 1.1-1.3). Normally, the input consists of a map $V_{T,J\times D}$ which is transposed to $\bar{V}_{J\times D,T}$ in case of TEU. This means in TEU each input data point (row) represents the temporal variation of a body joint over $t = 1$ to $t = T$ and hence the name TEU. This is in contrast to a normal convolution operation where each data point consists of body joint $j = 1$ to $j = J$ for a single frame. Similar to the spatial stream the output map $\bar{M}^{a}_{F,J\times D}$ of the TEU is passed to Block-B and Block-C (Fig. 1.2) in the temporal stream which perform the following transformation (Eqs. 1.2,1.3):

$$\bar{M}^{a}_{F_a,J\times D} -> \bar{M}^{c}_{F_a/4,F_c} \tag{1.10}$$

**Streams fusion:** The potential points for fusion of the two streams are at the end of each block. The SEU and the TEU produce maps of different dimensions (Eqs. 1.5,1.10) at the end of Block-A (Fig. 1.2). Moreover, the TCN-ResNet [16] reduces the temporal dimensions through Block-B and Block-C (Eqs. 1.1,1.2,1.3). Thus, the spatial and temporal streams produces maps of different dimensions at the end of each blocks. For example, the Block-C of spatial stream produces map $\hat{M}^{c}_{T/4,F_c}$ (Eq. 1.6), whereas the temporal stream has the map $\bar{M}^{c}_{F_a/4,F_c}$ (Eq. 1.10). The different sizes of the maps do not allow the maps to be fused with either concatenation or addition in a semantic manner. At the end of Block-C, the two streams can be fused by flattening and concatenating however, flattening disturbs the spatial and temporal structural organisation of the maps. FV-based clustering mechanism relies on such meaningful representations for semantic clustering [26]. Empirically, it was observed that flattening the two streams at this stage for fusion lead to poor performance. To preserve the structural organisation of the maps and to cluster them semantically,

each stream uses its own learn-able FV pooling. FVs [26] are computed as the aggregation of cluster weights, means and co-variances computed from Gaussian Mixture Model (GMM). Instead of calculating the FVs from GMM, NetFV (FV with neural network) learns these parameters [24]. Let $M_{R,S}$ be the input to FV-based pooling. In spatial stream, $M_{R,S}$ corresponds to $\hat{M}^c_{T/4,F_c}$ (Eq. 1.6) where $R = T/4$ and $S = F_c$. Similarly, in temporal stream, $M_{R,S}$ corresponds to $\bar{M}^c_{F_a/4,F_c}$ (Eq. 1.10) where $R = F_a/4$ and $S = F_c$. Let $r \in (1, R)$. The idea is to assign each $S$-dimensional data point of $M$ i.e., $M_r$ to a cluster as a soft-assignment [24]:

$$\alpha_k(\mathbf{M_r}) = \frac{e^{W_k^{T_r}\mathbf{M_r}+b_k}}{\sum_{j=1}^{K} e^{W_j^{T_r}\mathbf{M_r}+b_j}} \tag{1.11}$$

Here matrix $W_j$ and bias-vector $b_j$ are learn-able parameters. The soft-assignment $\alpha_k(\mathbf{M_r})$ to the $k^{th}$ cluster indicates how close $M_r$ is to the cluster $k$. Here, $j \in (1, K)$ where $K$ is the total number of clusters. Using the above soft-assignment, the Fisher vector is computed using the NetFV representation by [24]:

$$\begin{aligned} FV_1(j,k) &= \sum_{r=1}^{R} \alpha_k(\mathbf{M_r}) \left( \frac{\mathbf{M_r}(j) - c_k(j)}{\sigma_k(j)} \right) \\ FV_2(j,k) &= \sum_{r=1}^{R} \alpha_k(\mathbf{M_r}) \left( \left( \frac{\mathbf{M_r}(j) - c_k(j)}{\sigma_k(j)} \right)^2 - 1 \right) \end{aligned} \tag{1.12}$$

$FV_1$ and $FV_2$ respectively are the first-order and second-order statistics FV. $c_k$ and $\sigma_k$ are the learned cluster centre and the diagonal co-variance of the $k_{th}$ cluster, where $k \in (1, K)$. Here, $c_k$ and $\sigma_k$ are randomly initialised and are learned independently from the parameters of the soft-assignment $\alpha_k$ as in Eq. 1.11. As in [24], the FVs are then L2 normalised and concatenated and to get the final $FV = [FV_1, FV_2]$. In [24], the learned FV from the video stream is concatenated with the FV from the audio stream to form a FC layer which is further processed with context gating and mixture of experts identifier. Our implementation is different from the approach of [24], where a weighted pooling mechanism is used to output the final class maps:

$$Pooling(FV) = \text{softmax}(W_p FV + b_p) \tag{1.13}$$

Here, matrix $W_p \in \mathbf{R}^{|FV| \times C}$ and bias vector $b_p$ are learn-able parameters, and $C$ is number of human activity classes. The class-maps from both the stream are multiplied and normalised to get the final output. By avoiding concatenation of FVs we preserve the intelligently pooled features from the semantic FV-based clusters to form the class-maps. The multiplication and normalisation (Fig. 1.2) operation ensures that the network automatically learns the contribution weightage of each stream without the need for a further FC layer. Thus, in contrast to NetFV [24] we are able to produce class-maps without any further processing.

## 1.5 Training and Evaluation

Apart from the ground truth and evaluation method, the model is trained in a similar manner to standard single-label classification models. The ground truth is presented as multi-hot encoded labels to train the multi-label model. Two separate one-hot encoded labels, prepared as 'activity' labels and 'impairment' labels are concatenated to form the final ground truth labels. Let there be $A$ activity classes and $I$ impairment classes. For $a_{th}$ activity class where $a \in \{1...A\}$ and $i_{th}$ impairment class where $i \in \{1...I\}$, the one-hot encoded labels for activity and impairment respectively are:

$$AL_{m\in A} = \begin{cases} 1, & \text{if } m = a, \\ 0, & \text{if } m \neq a. \end{cases} \quad IL_{n\in I} = \begin{cases} 1, & \text{if } n = i, \\ 0, & \text{if } n \neq i. \end{cases} \tag{1.14}$$

$$GT = AL_{m\in A} \oplus IL_{n\in I} \tag{1.15}$$

To create the final ground truth label $GT$, the two labels are simply concatenated (Eq. 1.15). Thus, each of the ground truth label vectors $GT$ has two '1' values indicating activity and impairment. In $GT$, the 'activity' label comes from the first $A$ elements whereas the 'impairment' label is determined from final $I$ elements. Thus, to evaluate the model, the prediction probability vector (i.e., the model output) is split into two parts where the first part contains the first $A$ elements indicating the 'activity' class probabilities and the rest $I$ elements indicate the 'impairment' probabilities. Then, the accuracy for the 'activities' and 'impairments' are calculated separately. Finally, prediction by the model is considered to be true if both the 'activity' and 'impairment' predictions are correct.

## 1.6 Experiments and Results

We evaluate the proposed pose-based model in both single-label and multi-label mode. The well-known and challenging NTU-RGBD [28], which contains around 60K samples distributed over 60 action classes has been used for evaluation in single-label mode. We use the authors [28] protocol of cross-subject (CS) evaluation which is harder than the cross-view evaluation. Table 1.2 compares the proposed model to existing state-of-the art approaches and shows that the model achieves competitive performance under the constraints of data modality (pose-based, RGB video-based), end-to-end trainable and random initialisation (i.e., not pre-trained). The proposed model has the advantage to being end-to-end trainable as compared to [29, 33, 23]. Also, in contrast to [2, 19, 23, 33] we do not pre-train the proposed model which reflects the true capacity of model to learn without prior information. Given these constraints the best performance is achieved by ST-GCN [41] and the proposed model achieves almost similar performance while being a very light-weight model. ST-GCN [41] requires 8 Nvidia Titan X GPUs for training while we use only one Titan X GPU.

| Model | Mode | E2E | RI | CS(%) |
|---|---|---|---|---|
| TCN-ResNet [16] | P | ✓ | ✓ | 74.3 |
| Synth-CNN [23] | P | **x** | **x** | 80.0 |
| ST-GCN [41] | P | ✓ | ✓ | **81.5** |
| DPRL+GCNN [33] | P | **x** | **x** | 83.5 |
| 3Scale-ResNet152 [19] | P | ✓ | **x** | 85.0 |
| Glimpse Clouds [2] | R | ✓ | **x** | 86.6 |
| Learned-Encoding [6] | RP | ✓ | **x** | 87.7 |
| DGNN [29] | P | **x** | ✓ | 89.9 |
| Ours | P | ✓ | ✓ | 80.2 |

Table 1.2: The proposed model achieves competitive accuracy when compared with other pose-based state-of-the art models given the constraints of data mode (P: Pose, R: RGB-video), being end-to-end trainable (E2E) and random initialisation (RI). Given these constraints ST-GCN achieves the best performance and we achieve performance close to ST-GCN.

| **Model** | **Mode** | **A** | **I** | **Final** |
|---|---|---|---|---|
| I3D [3] | R | 87.2 | 65.9 | 55.9 |
| C3D [34] | R | 90.1 | 73.2 | 63.3 |
| TCN-ResNet [16] | P | 91.2 | 69.0 | 63.4 |
| Ours | P | **97.1** | **80.7** | **78.7** |

Table 1.3: Evaluation of the proposed dataset using different models that predict 'Activity' (A) and 'Impairment' (I) and a model's prediction is considered as correct if both the 'Activity and 'Impairment' predictions are true. Mode: Pose (P), RGB (R)

| **Model** | **Split 1** | | | **Split 2** | | | **Weighted-Average** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **I** | **Final** | **A** | **I** | **Final** | **A** | **I** | **Final** |
| TCN-ResNet [16] | 90.4 | 72.0 | 65.3 | 90.0 | 69.0 | 61.9 | 90.2 | 70.5 | 63.6 |
| Two-Stream (TEU) | 92.9 | 73.1 | 69.3 | 96.2 | 75.7 | 73.3 | 94.6 | 74.4 | 71.3 |
| Two-Stream (TEU + SEU) | 93.1 | 74.4 | 69.8 | 95.0 | 79.1 | 75.6 | 94.0 | 76.8 | 72.7 |
| Two-Stream (SEU + TEU) + FV | **96.9** | **77.7** | **75.3** | **97.3** | **83.7** | **82.2** | **97.1** | **80.7** | **78.7** |

Table 1.4: Ablation study demonstrating the effectiveness of the two-stream architecture and FVs

Next, we evaluate our model using the proposed multi-label functional ADL recognition dataset using the following protocol. A cross-validation approach is used where the dataset is split into two subject-wise folds for good generalisation. The first fold uses subjects 1, 3, 5, 7, 9 for training while subjects 2, 4, 6, 8, 10 are used for validation and vice-versa for the second fold. Thus, out of 5685 samples, the dataset is split into two groups of approximately equal groups of 2869 and 2816 samples which indicates a very good generalisation protocol. We do not use any data-augmentation or any transfer-learning approach to understand the true capacity

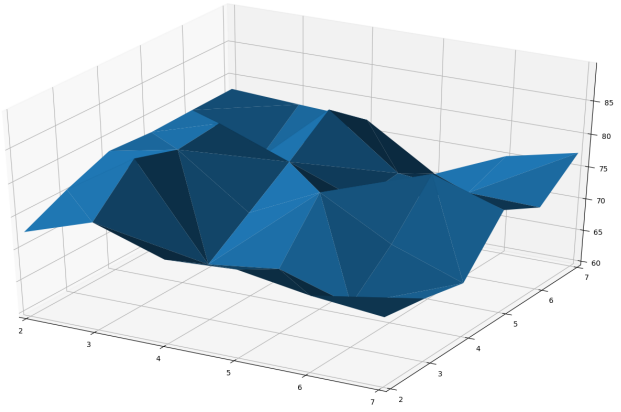



Fig. 1.3: Grid search for appropriate cluster-sizes show several parameter choices provide close to peak performance. This indicates that the TCN maps can be semantically clustered in multiple ways. Search range: $2^n$, where $n = 2, 3, 4, 5, 6, 7$

of the model to learn. The results in Table 1.3 are weighted average average of the two-fold cross-validation mentioned above. For each sample, the models in Table 1.3 predict 'activity' (A) and 'impairment' (I) classes and a model's 'Final' prediction is considered to be true if both the 'activity' and 'impairment' predictions are true.

### 1.6.1 Ablation study

In Table 1.4, we demonstrate the evolution of the model from the base TCN-ResNet to the final model through step-by-step inclusion of spatial-temporal architecture and the FV-based pooling. First, we experiment with the original TCN-ResNet (Row 1), and then we experiment with the two-stream architecture consisting of two parallel TCN-ResNets (Row 2). Here, TEU is introduced to Block-A (Fig. 1.2) of one of the streams making the stream temporal in nature, which is otherwise identical to the other stream. This greatly improves the accuracy which is further enhanced by the introduction of SEU to the spatial stream (Row 3). The final model accuracy is the greatly enhanced by the introduction of FV-based pooling to both the streams (Row 4). The number of clusters in FV is a tune-able hyper-parameter and for a grid-search was performed within the search space $2^n$, where $n = 2, 3, 4, 5, 6, 7$. The search results illustrated in Fig 1.3 which show that there are several peaks indicating higher performance with multiple cluster-size settings. The best performance (78.8%) is obtained with a cluster-size (CS) of $2^3$ for both the streams. Similar, results (78.0%) are obtained with CS is set at 8 (Spatial) and 16 (Temporal). CS of 16 (Spatial) and 64 (Temporal) gave 78.6% while CS of 64 (Spatial) and 32 (Temporal) gave 77.2% accuracy. The results suggest that the TCN maps can be semantically clustered in

a more than one way. However, with increased CS the model parameters increases and thus the aim should be to keep the CS to a minimum.

## 1.7 Conclusion

The paper is a step towards a robot or an automated systems-based assessment of physically impaired persons through ADL. To this end, we propose a dataset that consists of 10 different ADL that can test functional capacity of different body parts. Further, we present a normal and four different physical impairment-specific executions of each ADL. To our knowledge, the dataset is first to explore robot or an automated system's perception of functional assessment through ADL. The paper also presents a novel multi-label functional ADL recognition model that integrates a spatial-temporal body-pose encoding method with FV-based pooling in a novel manner. The dataset and the model will be made publicly available along with the publication of this paper.

**Acknowledgement:** We are immensely thankful to Dr Helen Carey for guiding us through the data collection process and Nvidia for providing the GPU.

## References

[1] Fabien Baradel, Christian Wolf, and Julien Mille. Human activity recognition with pose-driven attention to rgb. In *BMVC*, 2018.

[2] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *CVPR*, 2018.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[4] Alana Da Gama, Pascal Fallavollita, Veronica Teichrieb, and Nassir Navab. Motor rehabilitation using kinect: a systematic review. *Games for health journal*, 4(2):123–135, 2015.

[5] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the ICCV*, pages 833–842, 2019.

[6] Bappaditya Debnath, Mary O'Brient, Swagat Kumar, and Ardhendu Behera. Attention-driven body pose encoding for human activity recognition. *ICPR*, 2021.

[7] Girum G Demisse, Konstantinos Papadopoulos, Djamila Aouada, and Bjorn Ottersten. Pose encoding for robust skeleton-based action recognition. In *in Proc. of the CVPR*, 2018.

[8] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *PAMI*, 2016.

[9] L Ferrucci, C Koh, S Bandinelli, and JM Guralnik. Disability, functional status, and activities of daily living. In *Encyclopedia of gerontology*, pages 427–436. Elsevier Inc., 2010.

[10] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *in Proc. of the NIPS*, pages 34–45, 2017.

[11] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *in Proc. of the CVPR*, pages 971–980, 2017.

[12] John Green and John Young. A test-retest reliability study of the barthel index, the rivermead mobility index, the nottingham extended activities of daily living scale and the frenchay activities index in stroke patients. *Disability and rehabilitation*, 23(15):670–676, 2001.

[13] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *PAMI*, 2016.

[14] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Unified embedding and metric learning for zero-exemplar event detection. In *in Proc. of the CVPR*, pages 1096–1105, 2017.

[15] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *In Proc. of the CVPR*, pages 3288–3297, 2017.

[16] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *in Proc. of the CVPRW*, pages 1623–1631. IEEE, 2017.

[17] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[18] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *in Proc. of the CVPR*, pages 156–165, 2017.

[19] Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin, and Mingyi He. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In *ICMEW*. IEEE, 2017.

[20] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *CVPR Workshops (CVPRW)*, pages 9–14. IEEE, 2010.

[21] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on PAMI*, 42(10):2684–2701, 2019.

[22] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal LSTM with trust gates for 3D human action recognition. ECCV, 2016.

[23] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017.

[24] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

[25] Hossein Mousavi Hondori and Maryam Khademi. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering*, 2014, 2014.

[26] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*. IEEE, 2007.

[27] Supriya Sathyanarayana, Ravi Kumar Satzoda, Suchitra Sathyanarayana, and Srikanthan Thambipillai. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*, 9(2):225–251, Apr 2018.

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, 2019.

[30] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526. Springer, 2016.

[31] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *in Proc. of the AAAI*, 2017.

[32] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. In *AAAI Workshop*, 2011.

[33] Yansong Tang, Yi Tian, Jiwen Lu, Peiyang Li, and Jie Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In *CVPR*, 2018.

[34] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *in Proc. of the ICCV*, pages 4489–4497, 2015.

[35] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*. IEEE, 2012.

[36] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014.

[37] David Webster and Ozkan Celik. Systematic review of kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation*, 11(1):108, 2014.

[38] Yi Wei, Wenbo Li, Yanbo Fan, Linghan Xu, Ming-Ching Chang, and Siwei Lyu. 3d single-person concurrent activity detection using stacked relation network. *AAAI*, 2020.

[39] Lu Xia, Chia Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *CVPR Workshops*. IEEE, 2012.

[40] Zhongwen Xu, Yi Yang, and Alex G Hauptmann. A discriminative cnn video representation for event detection. In *in Proc. of the CVPR*, 2015.

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *in Proc. of the AAAI*, 2018.

[42] Mihai Zanfir, Marius Leordeanu, and Cristian Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *in Proc. of the ICCV*, 2013.