StrokeRehab: A Benchmark Dataset for Sub-second Action Identification

Aakash Kaku[*1], **Kangning Liu**[*1], **Avinash Parnandi**[*2], **Haresh Rengaraj Rajamohan**[1],
Kannan Venkataramanan[1], **Anita Venkatesan**[2], **Audre Wirtanen**[2], **Natasha Pandit**[2],
Heidi Schambra[†2],**Carlos Fernandez-Granda**[†1,3]
[1]NYU Center for Data Science [2]NYU School of Medicine
[3] Courant Institute of Mathematical Sciences

Abstract

Automatic action identification from video and kinematic data is an important machine learning problem with applications ranging from robotics to smart health. Most existing works focus on identifying coarse actions such as running, climbing, or cutting vegetables, which have relatively long durations and a complex series of motions. This is an important limitation for applications that require identification of more elemental motions at high temporal resolution. For example, in the rehabilitation of arm impairment after stroke, quantifying the training dose (number of repetitions) requires differentiating motions with sub-second durations. Our goal is to bridge this gap. To this end, we introduce a large-scale, multimodal dataset, StrokeRehab, as a new action-recognition benchmark that includes elemental short-duration actions labeled at a high temporal resolution. StrokeRehab consists of high-quality inertial measurement unit sensor and video data of 51 stroke-impaired patients and 20 healthy subjects performing activities of daily living like feeding, brushing teeth, etc. Because it contains data from both healthy and impaired individuals, StrokeRehab can be used to study the influence of distribution shift in action-recognition tasks. When evaluated on StrokeRehab, current state-of-the-art models for action segmentation produce noisy predictions, which reduces their accuracy in identifying the corresponding sequence of actions. To address this, we propose a novel approach for high-resolution action identification, inspired by speech-recognition techniques, which is based on a sequence-to-sequence model that directly predicts the sequence of actions. This approach outperforms current state-of-the-art methods on StrokeRehab, as well as on the standard benchmark datasets 50Salads, Breakfast, and Jigsaws.

1 Introduction

In domains ranging from robotics to smart health, automatically identifying action from video and kinematic data is an important machine learning problem. In some of these applications it is critical to identify motions at high temporal resolution. This is the case in data-driven stroke rehabilitation, which requires classifying and counting sub-second single-goal motions. In order to advance methodology for high-resolution action identification, it is crucial to establish appropriate benchmark datasets. Existing benchmarks, such as 50Salads [52], Breakfast [32], Jigsaws [18], or Kinetics [28] contain very few short-duration actions (see Figure 3). To address this, we introduce a large-scale, multimodal dataset, StrokeRehab, as a new action-recognition benchmark that includes elemental short-duration actions labeled at a high temporal resolution. The dataset consists of high-quality wearable sensor and video data of 51 stroke patients and 20 healthy subjects. These

[*]Equal Contribution

[†]Joint Last Author

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

https://doi.org/10.52202/068431-0122

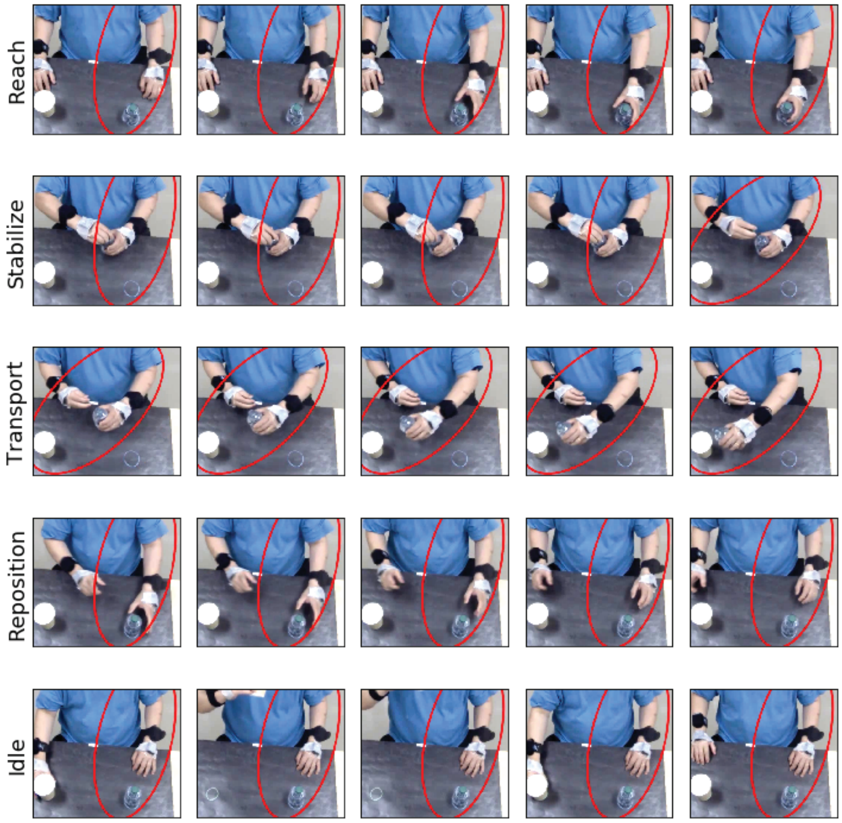


Figure 1: A stroke patient performing a functional activity (drinking) from the StrokeRehab activities battery. Using the functional motion taxonomy, the activity can be decomposed into its constituent functional primitives as follows: *reach*, upper extremity (UE) motion to bring it into contact with a target object (e.g. water bottle); *stabilize* minimal UE motion to keep a target object still (e.g. holding the bottle to allow the other UE to open the cap); *transport*, UE motion to move a target object (e.g. moving the bottle to pour some water); *reposition*, UE motion proximate to a target object (e.g. to move the to the initial neutral spot); *idle*, minimal UE motion to stand at the ready near a target object.

subjects performed nine activities of daily living like drinking, eating, applying deodorant, etc. in a rehabilitation gym (see Figure 1 for an example). The elemental actions performed by the subjects in each session were meticulously labeled by trained annotators overseen by an expert, who examined one third of their labels. These labeled actions are called functional primitives, consisting of five main classes: reach (upper extremity (UE) motion to make contact with a target object), reposition (UE motion to move into proximity of a target object), transport (UE motion to convey a target object in space), stabilization (minimal UE motion to hold a target object still), and idle (minimal UE motion to stand at the ready near a target object) [48].

Evaluation of current state-of-the-art models for action segmentation on StrokeRehab reveals that these approaches are not as effective when applied to short-duration elemental actions. The reason is that the boundaries of these actions are not clearly defined, even for human-expert annotators. As a result, segmentation-based approaches produce noisy estimates, which limits their accuracy. To address this limitation, we introduce an approach to action identification inspired by speech-recognition models, which achieves state-of-the-art performance on StrokeRehab, and also outperforms existing approaches on the standard benchmark datasets like Breakfast, 50Salads, and Jigsaws for the task of sequence-identification. This showcases how StrokeRehab can contribute to methodological advances in machine-learning methodology for action identification.

StrokeRehab contains data from healthy and stroke-impaired individuals. It therefore provides a real-world example of distributional shift. We show that models trained on impaired patients are able to generalize to healthy subjects, but the opposite is not true. In addition, we show that the

https://doi.org/10.52202/068431-0122

performance of models trained on moderately-impaired patients does not generalize effectively to severely-impaired patients. StrokeRehab therefore provides a challenging benchmark dataset to evaluate methods addressing distributional shift.

StrokeRehab also has the potential to advance data-driven stroke rehabilitation. Basic research in animals indicates that the repeated practice of functional motions early after stroke markedly boosts recovery [25, 4, 42]. The same is believed to be true for humans undergoing rehabilitation for stroke-induced disability. However, there has been no systematic quantification of how many training repetitions are needed early after stroke for optimal recovery [31]. Data-driven quantification requires identifying and counting motions at high temporal (sub-second) resolution. StrokeRehab provides a high-quality labeled benchmark dataset for this task.

To summarize, StrokeRehab is a multimodal, labeled real-world dataset, which provides a benchmark for (1) sub-second action identification, (2) generalization in the presence of realistic distributional shift, and (3) data-driven quantification of rehabilitation dose.

2 Description of the dataset

2.1 Clinical motivation

Stroke is the leading cause of disability in the United States. It affects nearly 800,000 individuals per year, with the numbers of stroke cases increasing as our population ages [19, 46, 6]. Stroke affects the arm in 77% of patients, causing long-lasting motor impairment [36, 34]. By six months, most of these patients remain unable to independently perform activities of daily living, such as feeding, bathing, grooming, etc. This loss of independence reduces the quality of life of both the patients and their caretakers [45, 7] and exacts a heavy societal toll, with annual caretaking and healthcare costs predicted to skyrocket to $240 billion by 2030 [23]. Due to the profound impact of stroke on arm function and its downstream consequences, we focus on the arm in our study.

Following stroke, some spontaneous recovery occurs because of brain plasticity, but this plasticity alone does not fully restore function. In animal models of stroke, training high numbers of functional arm motions not only increases this plasticity [29, 4], but also markedly boosts recovery [42, 25]. It is increasingly believed that if intensive rehabilitation training can be delivered early after stroke in humans, recovery could be similarly accelerated [31].

In rehabilitation training, patients use their impaired arm to practice activities of daily living (ADLs). ADLs are composed of five elemental actions called functional primitives: reach, transport, reposition, stabilize, and idle [48]. For example, in a drinking activity, our arm would "idle" as it rests at our side, then "reach" for a glass and "transport" the glass to our mouth, then "transport" the glass back to the table, and finally "reposition" back to our side for an "idle." Examples of these primitives for a drinking activity can be seen in Figure 1.

A major clinical question is how many repetitions of functional motions are needed to boost recovery. In animal research, the number of repetitions that promote recovery has been quantified [25]. For humans, this quantification has not been done. A handful of studies have observed that patients train about 10 times fewer repetitions than what recovering animals receive, suggesting pronounced under-training in human rehabilitation [35, 30]. However, the optimal number of training repetitions to boost recovery remains uncertain in humans. Currently, the best way to quantify rehabilitation is hand tallying. If performed in real time by the rehabilitation therapist, hand tallying distracts from treatment delivery. If the session is videotaped and annotated offline, tallying is laborious and slow: **it takes one hour of manual effort to label one minute of recorded training**. Hand tallying thus incurs time, effort, and personnel costs that render it unscalable.

Therefore, to facilitate the quantification of training motions in stroke rehabilitation, we developed an approach that combines unobtrusive motion capture with automated identification. We collected the StrokeRehab dataset that consists of labeled sensor and video data from stroke patients and healthy subjects. We then used the StrokeRehab dataset to train models to automatically identify and count functional primitives. The dataset is hosted at SimTK website - `https://simtk.org/projects/primseq`. The license agreement and datasheet for the dataset can be found in Appendix I and Appendix J, respectively.

2.2 Cohort selection

We collected sensor and video data from 51 stroke patients and 20 healthy subjects in an inpatient rehabilitation gym.

https://doi.org/10.52202/068431-0122

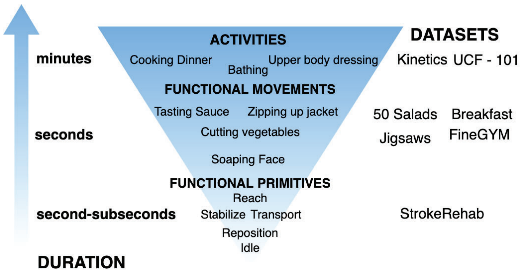


Figure 2: Action recognition datasets ordered according to a hierarchy of the labeled actions they contain. Our dataset StrokeRehab consists of short-duration elemental actions, called functional primitives.

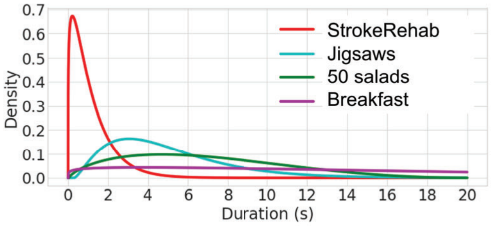


Figure 3: Distribution of action duration for various benchmark datasets and StrokeRehab. This illustrates the extreme fine-grained nature of actions (functional primitives) in the StrokeRehab dataset in comparison to existing ones.

Stroke patients: Individuals were included if they were $\geq$ 18 years old, premorbidly right-handed, and had unilateral arm weakness from an ischemic or hemorrhagic stroke that occurred at least 6 months prior.

Patients were excluded if they had traumatic brain injury; any musculoskeletal or non-stroke neurological condition that interferes with motor function; contracture at the shoulder, elbow, or wrist; moderate arm dysmetria or truncal ataxia; visuospatial neglect; apraxia; global inattention; or legal blindness. A trained assessor quantified arm impairment with the upper extremity Fugl-Meyer Assessment, where a maximum score of 66 signifies no impairment [16]. Stroke impaired patients were divided into two sub-cohorts: moderately to mildly impaired (scores 23-65) and severely impaired patients (scores 8-23). For healthy controls, individuals were included if they were $\geq$ 18 years old, had a right-handed dominance and no motor impairment (score = 66).

Table 1 describes the demographic and clinical characteristics of the stroke impaired patients and healthy controls.

Table 1: Demographic and clinical characteristics of the stroke impaired patients and healthy controls in the cohort. Mean and ranges in parentheses are shown. The cohort is divided into a training set and a test set of mildly and moderately-impaired patients. There is no overlap of patients between the training and test set.

	Training set (Mild + Moderate)	Test set (Mild + Moderate)	Severe set	Healthy control
n	35	8	8	20
Age (in years)	56.56 (21.2-82.7)	60.8 (42.6-84.2)	59.73 (41-74.3)	62.47 (42-82.9)
Gender (Female : Male)	19 F : 16 M	4 F : 4 M	5 F : 3 M	9 F : 11 M
Time since stroke (in years)	6.5 (0.3-38.4)	3.1 (0.4-5.7)	3.46 (1.14-6.43)	NA
Paretic side (Left : Right)	20 L : 15 R	4 L : 4 R	4 L : 4 R	NA
Fugl-Meyer Assessment score	48.1 (26-65)	49.4 (27-63)	16 (8-23)	66

2.3 Data acquisition and labelling

Upper body motion was recorded while subjects performed activities of daily living commonly used during stroke rehabilitation. The activities included: washing the face, applying deodorant, combing the hair, donning and doffing glasses, preparing and eating a slice of bread, pouring and drinking a cup of water, brushing teeth, and moving an object horizontal and vertical target arrays. Subjects performed five repetitions of each activity. See Appendix H.1 for detailed descriptions of the activities.

Description of kinematic data: Upper body motion was recorded using nine Inertial Measurement Units (IMUs, Noraxon, USA) attached to the upper body, specifically the cervical vertebra C7, the thoracic vertebra T12, the pelvis, and both arms, forearms, and hands (see Figure 8 in Appendix H). The sensors are lightweight (34 g) and small (matchbook-size). They are adhered to the back of the hand with thin tape that does not interfere with finger movement or grasp. Similarly, the straps holding the sensors to the forearms and arms do not cross any joints. Neither the location nor the methods used to affix the sensors are expected to interfere with natural motion. These IMUs captured

76-dimensional kinematic features of 3D linear accelerations, 3D quaternions, and joint angles from the upper body (see Appendix H.2 for details). As an additional feature, for stroke-impaired patients, we included the paretic (stroke-impaired) side of the patient (left or right) encoded in a one-hot vector, increasing the dimension of the feature vector to 77. For healthy subjects, the movements of both hands are labeled. Therefore, the 77th feature equals *right* if we are making predictions for the right hand and *left* if we are making predictions for the left hand.

Each IMU captures 3D linear accelerations and angular velocities at 100 Hz. Angular velocities are converted to sensor-centric unit quaternions, representing the rotation of each sensor on its own axes, with coordinate transformation matrices. In addition, proprietary software (Myomotion, Noraxon) generates 22 anatomical angle values using a rigid-body skeletal model scaled to patient height. See Appendix H.2 for a detailed description of these angles. Each entry (except for the feature encoding paretic side) was mean-centered and normalized separately for each task repetition in order to remove spurious offsets introduced during sensor calibration.

Description of video data: Video data were synchronously captured using two high definition cameras (1088 x 704, 60 frames per second or 100 frames per second; Ninox, Noraxon) placed orthogonally < 2 m from the subject. We extract frame-wise features from the raw videos using the X3D model [14], a 3D convolutional network designed for video classification. The model is pretrained on the Kinetic dataset [28], which consists of coarse actions like running, climbing, sitting, etc. Since the StrokeRehab dataset consists of elemental, sub-second actions, we fine-tuned the X3D model on the training set of StrokeRehab. In order to fine-tune, we used video sequences as input and trained the model to identify the primitive happening in the center frame of the sequence.

Data labeling: The elemental actions performed by the patients in each session were meticulously labeled by trained annotators overseen by an expert, who examined one-third of their labels. Interrater reliability between the coders and expert was high, with Cohen's kappa ≥ 0.96 between the coders and the expert.

2.4 Training and test sets

Healthy subjects. We randomly assigned 16 subjects to a training set and 4 subjects to the *Healthy subject test set*.

Stroke patients. We used a sub-cohort of 43 mild and moderately impaired patients to create the stroke patient training and test sets. Patients were separated into eight subgroups, balancing for impairment level and paretic side (left or right). One patient in each group was randomly removed and assigned to the *Stroke patient test set*.

Severely impaired stroke patients. A second sub-cohort of 8 severely impaired patients with impairment scores in the range 8-23 [16] was used to create a challenging test set to evaluate model generalizability to patients with higher impairment levels. We call this the *Severe impairment test set*.

2.5 Data/Label quality

An expert in the functional motion taxonomy (AP) [48] individually trained the annotators, who underwent one month of intensive training on a series of increasingly complex activities. Once annotators labeled the training videos with high accuracy (less than 2% errors), they were given new videos to label independently. The annotators identified and labeled functional primitives in the video, which simultaneously labeled primitives in the IMU data. To ensure consistent labeling, the expert inspected one-third of all labeled videos. Inter-rater reliability between the annotators and expert was high across primitives (Cohen's kappa: reaches, 0.96; repositions, 0.97; transports, 0.97; stabilizations, 0.98; idles, 0.96). When computed in terms of Action Error Rate (AER in Section 3.2), the inter-rater reliability for the entire dataset is 0.021. One minute of recording took on average 79.8 minutes to annotate. There were two experts and seven annotators.

2.6 Comparison to existing benchmark datasets

The dataset consists of 3,372 trials of rehabilitation activities performed by 51 stroke-impaired and 20 healthy subjects. Cumulatively, they performed 120,891 functional primitives, which is more than existing benchmark datasets such as FineGym (32,697 annotated sub-actions), Breakfast (11,656 annotated actions), Jigsaws (1,701 annotated actions), and 50Salads (999 annotated actions). A non-trivial amount of manual effort (approximately 2,700 human hours) was required to label the functional primitives, which span 43.48 hours of recorded training.

https://doi.org/10.52202/068431-0122

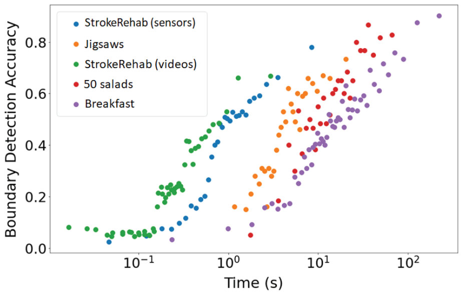


Figure 4: Boundary accuracy achieved by the segmentation models vs duration of the actions for several datasets. Boundary-detection accuracy is directly proportional to action duration.

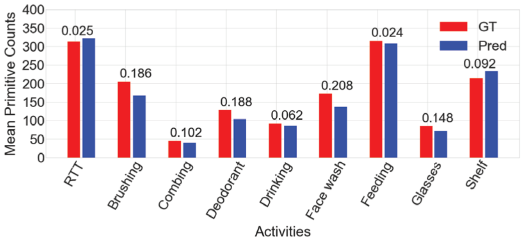


Figure 5: Comparison of ground-truth and predicted mean counts for the different high-level activities in the StrokeRehab dataset. The relative error is very small for structured activities like moving objects on/off a shelf (Shelf), and larger for unstructured activities like brushing.

Figure 2 shows a hierarchy of human actions. Existing datasets focus on high level actions associated to particular activities or objects (e.g. cutting vegetables, zipping up jacket), which typically have long durations (seconds or longer) and execute several goals. In contrast, as shown in Figure 3, the actions in the StrokeRehab dataset are much shorter (sub-second). They correspond to functional primitives that execute one goal, and are therefore qualitatively different to the higher-level actions in existing benchmark datasets. In particular, they are not associated to specific objects or contexts; identifying them correctly requires learning to distinguish elemental human movements.

3 Action sequence identification

The main task associated to StrokeRehab is action sequence identification, i.e. identifying the correct sequence of actions carried out by an individual. This is often the ultimate goal in applications of action recognition [41, 43, 1], and is particularly critical in data-driven rehabilitation [26].

3.1 Problem definition

Let $\mathbf{x} = (x_1, ..., x_T)$ be an input sequence with length T, which may correspond for example to high-dimensional sensor or video features. The goal of action sequence identification is to estimate the corresponding sequence of actions encoded as $\mathbf{y} = (y_1, ..., y_{T'})$, where y_i $(1 \leq i \leq T')$ is one of c different actions, so $y_i \in \{1, \ldots, c\}$. The length T' of this sequence is much shorter than the input sequence because each action takes place over several time steps. For example, a 100-frame video sequence from the StrokeRehab dataset may correspond to the 3-action sequence: *reach*, *transport*, *stabilize*.

3.2 Evaluation metric: Action error rate (AER)

In this section we introduce a metric to evaluate methods for action sequence identification based on the Levenshtein distance, which is a distance tailored to sequence estimation tasks. The Levenshtein distance $\mathrm{L}(G, P)$ between a ground-truth sequence G and a predicted sequence P is the minimum number of insertions, deletions, and substitutions required to convert P to G. For example, if G = [*reach*, *idle*, *stabilize*] and P = [*reach*, *transport*], then $\mathrm{L}(G, P) = 2$ (*transport* is substituted for *idle* and *stabilize* is inserted). An important consideration is how to normalize this distance in order to obtain a metric to evaluate sequence identification. Existing works in action recognition [13, 57, 24, 38] use the edit score (ES):

$$\mathrm{ES}(G, P) := \left(1 - \frac{\mathrm{L}(G, P)}{\max(\mathrm{len}(G), \mathrm{len}(P))}\right) \times 100, \tag{1}$$

where $\mathrm{len}(G)$ and $\mathrm{len}(P)$ are the lengths of G and P respectively. Due to the normalization factor consisting of the maximum between the estimated and ground-truth sequences, this metric is lenient with long estimated sequences containing noisy, spurious estimates. This can be addressed by normalizing with respect to the length of the ground-truth sequence. We call the corresponding metric action error rate (AER), since it is analogous to word error rate, a standard metric in speech

Table 2: Results for action sequence identification on the stroke-patient test set of StrokeRehab (top table) and on existing benchmark datasets (bottom table). We report the mean of the metrics of interest with 95% confidence intervals computed via bootstrapping (see Appendix G.2). In both tables, * indicates models selected based on the best validation frame-wise accuracy. Overall, we observe that seq2seq models tend to outperform segmentation-based approaches. Additional metrics like true positive rate and false discovery rate can be seen in Appendix D

StrokeRehab (stroke-patient test set)

	Model	Video Data		Sensor Data	
		Edit Score	Action Error Rate	Edit Score	Action Error Rate
Segmentation-based model	MS-TCN* [13]	60.7 (59.1 - 62.2)	0.408 (0.388 - 0.428)	66.9 (65.0 - 68.9)	0.372 (0.335 - 0.406)
	MS-TCN [13]	62.2 (60.8 - 63.6)	0.392 (0.371 - 0.413)	68.7 (67.3 - 70.5)	0.330 (0.307 - 0.354)
	+ Smoothing window	62.7 (61.3 - 64.1)	0.390 (0.370 - 0.410)	**68.8 (67.3 - 70.3)**	0.317 (0.297 - 0.338)
	ASRF* [24]	56.9 (55.2 - 58.6)	0.449 (0.427 - 0.472)	68.2 (66.7 - 69.9)	0.328 (0.309 - 0.348)
	ASRF [24]	58.7 (57.3 - 60.2)	0.436 (0.417 - 0.456)	67.9 (66.3 - 69.5)	0.349 (0.326 - 0.372)
seq2seq	Seg2seq	**67.6 (66.4 - 68.8)**	**0.322 (0.307 - 0.339)**	63.0 (61.3 - 64.7)	0.337 (0.311 - 0.363)
	Raw2seq	66.6 (65.4 - 67.9)	0.329 (0.312 - 0.345)	**68.8 (67.4 - 70.3)**	**0.305 (0.284 - 0.324)**

Existing benchmark datasets

	Model	50 salads		Breakfast		Jigsaws	
		Edit Score	Action Error Rate	Edit Score	Action Error Rate	Edit Score	Action Error Rate
Segmentation-based model	MS-TCN* [13]	68.8	0.47	62.0	1.16	55.24	0.96
	MS-TCN [13]	70.8	0.43	61.7	0.97	61.44	0.82
	+ Smoothing window	76.4	0.32	69.1	0.51	76.54	0.31
	ASRF* [24]	74.0	0.34	71.2	0.44	71.29	0.37
	ASRF [24]	75.2	0.33	70.9	0.45	74.63	0.31
Seq2seq	Seg2seq	**76.9**	**0.30**	**73.7**	**0.37**	**83.87**	**0.17**
	Raw2seq	69.4	0.54	64.1	0.55	70.13	0.35

recognition:

$$\mathrm{AER}(G,P) := \frac{\mathrm{L}(G,P)}{\mathrm{len}(G)} \tag{2}$$

AER penalizes longer and shorter predictions equally. For example, if G = [*reach*, *idle*, *stabilize*], P_1 = [*reach*,*idle*], and P_2 = [*reach*, *idle*, *stabilize*, *transport*], then $\mathrm{ES}(G, P_1)$ = 0.67 and $\mathrm{ES}(G, P_2)$ = 0.75, but $\mathrm{AER}(G, P_1)$ = $\mathrm{AER}(G, P_2)$ = 0.33.

3.3 Limitations of segmentation-based models

As described in Appendix A, where we provide a detailed description of the state of the art, most existing methods [13, 57, 24, 38] address the task of action sequence identification by performing segmentation of the input data and then removing label repetitions at consecutive steps (see Figure 6 for a concrete example). Experimentally we observe (see Section 3.4) that this may result in systematically over-segmented noisy outputs, as reported also in [13, 57, 24]. Recent works [57, 24] address this by training a separate action-boundary detection network. The boundary predictions are then used to refine the frame-wise predictions. As pointed out by [49], boundaries of high-level actions are more detectable because the adjacent motions are distinctive; for example, the motions associated with cutting tomatoes versus tossing a salad are very different. In contrast, boundaries of fine-grained actions are harder to identify because their transitions are more elemental; for example, the boundary between the end of a *reach* and the beginning of a *transport* in the StrokeRehab dataset is determined by when the finger pads have fully contacted a target object.

Figure 4 shows the accuracy of boundary-detection models for actions with different durations for several datasets. For all datasets, the accuracy of boundary detection is directly proportional to duration. This suggests that segmentation-based approaches may be fundamentally limited for identification of elemental action sequences at high time resolution. This is very problematic for the proposed dataset StrokeRehab where the action durations are very short (see Figure 3), but it also limits performance on existing benchmarks like 50salads and Breakfast (see Section 3.4). Motivated by this limitation, we propose a sequence-to-sequence method that predicts the sequence of actions directly in the following section.

3.4 Proposed methodology: Sequence-to-sequence models

Motivated by the limitations of segmentation-based approaches described in Section 3.3, we propose to directly predict the sequence of actions from the input data using sequence-to-sequence (seq2seq)

https://doi.org/10.52202/068431-0122

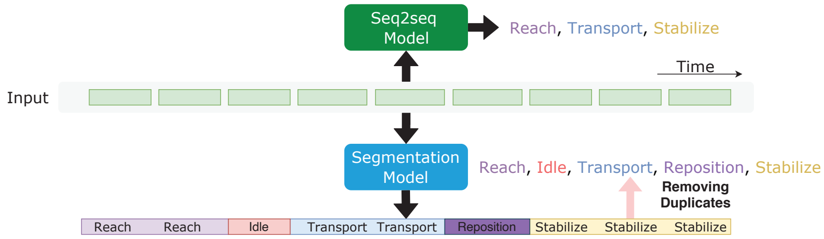


Figure 6: Comparison of sequence-to-sequence (seq2seq) and segmentation models. For an input frame sequence, the segmentation model outputs frame-wise action predictions with the same total length as the input frames. The frame-wise prediction can then be converted to a sequence estimate by removing the duplicates. In contrast, the seq2seq model learns a mapping of a variable-length input sequence to a variable-length output sequence directly.

Table 3: In order to study the effect of distribution shift on StrokeRehab we evaluate a Raw2seq model trained only on healthy subjects (HS), only on stroke patients (SP) and on both (HS+SP) using different test datasets (see Section 2.4). The model trained only on healthy subjects, fails to generalize to the other two cohorts. In contrast, models trained on stroke patients generalize well to healthy subjects. All models have difficulties generalizing to severely impaired patients (but the performance of the HS-trained model is particularly poor) (see Appendix C). We report the mean of the metrics of interest with 95% confidence intervals computed via bootstrapping (see Appendix G.2).

Tested on	Healthy subjects	Stroke patients	Severely impaired
Trained on	Action Error Rate	Action Error Rate	Action Error Rate
Healthy Subjects (HS)	0.281 (0.263 - 0.299)	0.405 (0.383 - 0.425)	0.819 (0.746 - 0.898)
Stroke patients (SP)	0.286 (0.268 - 0.303)	0.305 (0.284 - 0.324)	0.612 (0.562 - 0.676)
HS + SP	0.287 (0.269 - 0.304)	0.297 (0.279 - 0.318)	0.604 (0.558 - 0.656)

models inspired by speech recognition (see Figure 6). These models are designed to map input sequences to output sequences of different length, and are therefore well suited to the action-sequence estimation problem. The key idea is to *encode* the input data as a hidden vector of fixed dimension. The hidden vector is then *decoded* sequentially to produce the estimated sequence. During training we apply the seq2seq models on overlapping windows of short duration. During inference, we divide the input data into non-overlapping windows and concatenate the estimates removing duplicates at the boundaries.

Computational constraints limit the window size of inputs to sequence-to-sequence models to 500-1000 time steps. This is sufficient for applications in natural language processing [3] and speech recognition [9]. However, in order to identify high level actions, it may be necessary to model long-term dependencies. To overcome this limitation, we propose a version of our seq2seq model, which uses frame-wise predictions from a segmentation-based model (specifically an MS-TCN model) as inputs. These frame-wise predictions can be interpreted as features capturing long-term dependencies. To differentiate the two versions of our proposed approach we call *Raw2seq* the method that uses raw sequences of sensor data or video features as input, and *Seg2seq* the method that uses frame-wise predictions from a segmentation-based model as input. Appendix G.3 provides a more detailed explanation of our proposed seq2seq models. Additional implementation details are provided in Appendix G.4.

4 Experiments and Results

4.1 Action sequence identification

We use StrokeRehab to compare the proposed seq2seq methods, described in Section 3.4, to two segmentation-based baselines: MS-TCN [13] and ASRF [24], which was chosen because it outperforms other action segmentation models [57, 38, 17, 10] on existing benchmark datasets (see Appendix G.1). Table 2 reports the results for the *Stroke-patient test set* of StrokeRehab (see Appendix B for additional results on healthy subjects), and also includes results of a comparison based on existing benchmark datasets. To ensure a fair comparison, we optimized the segmentation models using our metric of interest for sequence identification (AER). Interestingly, optimizing this metric, as opposed to framewise accuracy, substantially boosts the performance of the MS-TCN baseline (see Table 2). Our validation and evaluation procedures are explained in detail in Appendix G.2.

The results in Table 2 indicate that sequence-to-sequence models tend to outperform segmentation-based models. Raw2seq achieves better performance on the sensor data of the StrokeRehab data, where the actions are very localized and do not require modeling long-time dependencies, Seg2seq is superior on the remaining datasets. The baselines achieve edit score values that are close to those of seq2seq on the sensor data of StrokeRehab (but not on the video data), and on 50Salads, but the AER of seq2seq is better. The reason is that, as explained in Section 3.2, the edit score is more lenient with the false positives that tend to be produced by segmentation-based models. Interestingly, the refinement strategies of smoothing and boundary detection (ASRF models) do not improve the baselines on StrokeRehab, highlighting that it contains actions that are qualitatively different from those of the other datasets. Additionally, models selected based on best validation AER generally outperform models selected based on best frame-wise accuracy. This underscores the importance of optimizing metrics that are specifically tailored to sequence identification.

4.2 Comparison between data modalities

StrokeRehab contains both video and sensor data, which makes it possible to compare both modalities. Figure 7 shows confusion matrices of the predictions produced by the best models based on video and sensor data respectively. In order to compute AER, there are four components that need to be computed: correctly predicted primitives, substituted primitives, inserted primitives and deleted primitives. The confusion matrices only shows the correctly predicted and substituted primitives. Complete confusion matrices with two more columns namely deleted primitives, and inserted primitives can be seen in Appendix F. The nature of errors of two models were complementary to some extent. For example, the model trained on video data confused reaches with idles, whereas the model trained on sensor data confused reaches with stabilizes. For the video data, confusion occurs mainly between primitives which are usually performed one after the other (eg. idles and reaches). For example, the functional movement that follows idle is a reach 69.3% of the time. This is because the patients tend to be idle right before reaching out to touch an object. In contrast, for the sensor data, there is more confusion between primitives that look similar, such as reaches and transports or idles and stabilizes. Since the nature of errors is disparate, this makes a good case for training multi-modal models that can leverage both modalities to perform the task of sequence estimation.

4.3 Studying distribution shift

The different training and test sets in StrokeRehab (see Section 2.4) can be used to evaluate the effect of distribution shift on models trained for action sequence identification. To this end, we trained Raw2seq models on only healthy subjects, only stroke patients, and on both. We then tested these models on the three test sets containing healthy subjects, mildly and moderately impaired stroke patients, and severely impaired patients. Table 3 shows that the model trained only on healthy subjects failed to generalize to the other two cohorts. In contrast, the model trained on stroke patients (mildly and moderately impaired), generalized well to healthy subjects. Therefore, movements observed in stroke patients are relevant to healthy patients, but the opposite is not true. In addition, all models struggled to generalize to the severely impaired patients. This indicates that generalization to different impairment levels is a challenging problem, and that it is paramount to build training sets containing stroke-patient data in order to train models for data-driven rehabilitation. More analysis is provided in Appendix E.

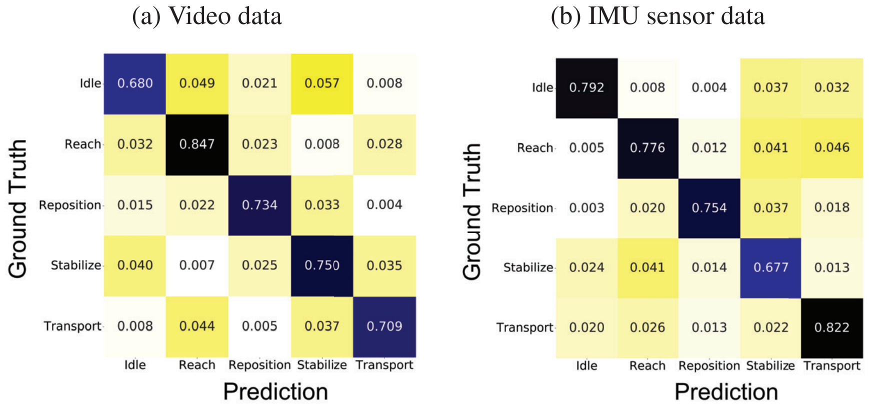
4.4 Automatic quantification of rehabilitation dose

In stroke rehabilitation, action identification can be used for quantifying dose by counting functional primitives. Figure 5 shows that the raw2seq version of the seq2seq model produces accurate counts for all activities in the StrokeRehab dataset. Performance is particularly good for structured activities such as moving objects on/off a shelf, in comparison to less structured activities such as brushing, which tend to be more heterogeneous across patients

5 Discussion and Conclusion

In this work, we introduce a large-scale, multimodal dataset, StrokeRehab, as a new benchmark for identification of action sequences that includes labled short-duration actions. In addition, we introduce a novel sequence-to-sequence approach, which outperforms existing methods on StrokeRehab as well

https://doi.org/10.52202/068431-0122


(a) Video data

	Idle	Reach	Reposition	Stabilize	Transport
Idle	0.680	0.049	0.021	0.057	0.008
Reach	0.032	0.847	0.023	0.008	0.028
Reposition	0.015	0.022	0.734	0.033	0.004
Stabilize	0.040	0.007	0.025	0.750	0.035
Transport	0.008	0.044	0.005	0.037	0.709

(b) IMU sensor data

	Idle	Reach	Reposition	Stabilize	Transport
Idle	0.792	0.008	0.004	0.037	0.032
Reach	0.005	0.776	0.012	0.041	0.046
Reposition	0.003	0.020	0.754	0.037	0.018
Stabilize	0.024	0.041	0.014	0.677	0.013
Transport	0.020	0.026	0.013	0.022	0.822

Figure 7: Confusion matrices for the best performing models on the StrokeRehab video and sensor dataset for the stroke-patient test set. The diagonal entries are the fractions of primitives estimated correctly. The off-diagonal entries are the fractions of primitives that were substituted.

as on existing benchmark datasets. A known limitation of sequence-to-sequence approaches is that they have difficulties capturing long-term dependencies. Here, we address this by using the output of a segmentation-based network as an input for the sequence-to-sequence model. An interesting direction for future research is to design sequence-to-sequence models capable of directly learning these dependencies. Our results also show that models based on video and wearable-sensor data have different strengths and weaknesses (see Section 4.4), which suggests that multimodal approaches may have significant potential.

Acknowledgments

We would like to thank the volunteers who contributed to label the dataset: Ronak Trivedi, Adisa Velovic, Sanya Rastogi, Candace Cameron, Sirajul Islam, Bria Bartsch, Courtney Nilson, Vivian Zhang, Nicole Rezak, Christopher Yoon, Sindhu Avuthu, and Tiffany Rivera. We thank Dawn Nilsen, and OT EdD for expert advice on the testing battery. This work was supported by an AHA postdoctoral fellowship 19AMTG35210398 (AP), NIH grants R01 LM013316 (AK, KL, HR, CFG, HMS) and K02 NS104207 (HMS), NSF NRT-HDR Award 1922658 (AK, KL, HR, CFG).

References

[1] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[2] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat. Physical human activity recognition using wearable sensors. *Sensors*, 15(12):31314–31338, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] J. A. Bell, M. L. Wolke, R. C. Ortez, T. A. Jones, and A. L. Kerr. Training intensity affects motor rehabilitation efficacy following unilateral ischemic insult of the sensorimotor cortex in c57bl/6 mice. *Neurorehabilitation and neural repair*, 29(6):590–598, 2015.

[5] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv preprint arXiv:1506.03099*, 2015.

[6] J. P. Broderick. William m. feinberg lecture: stroke therapy in the year 2025: burden, breakthroughs, and barriers to progress. *Stroke*, 35(1):205–211, 2004.

[7] F. J. Carod-Artal and J. A. Egido. Quality of life after stroke: the importance of a good recovery. *Cerebrovascular diseases*, 27(Suppl. 1):204–214, 2009.

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[9] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*, 2015.

[10] M.-H. Chen, B. Li, Y. Bao, G. AlRegib, and Z. Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020.

[11] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. *arXiv preprint arXiv:1506.07503*, 2015.

[12] S. Chung, J. Lim, K. J. Noh, G. Kim, and H. Jeong. Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors*, 19(7):1716, 2019.

[13] Y. A. Farha and J. Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.

[14] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[15] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6202–6211, 2019.

[16] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scandinavian journal of rehabilitation medicine*, 7(1):13–31, 1975.

[17] S.-H. Gao, Q. Han, Z.-Y. Li, P. Peng, L. Wang, and M.-M. Cheng. Global2local: Efficient structure search for video action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16805–16814, 2021.

[18] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014.

[19] A. S. Go, D. Mozaffarian, V. L. Roger, E. J. Benjamin, J. D. Berry, M. J. Blaha, S. Dai, E. S. Ford, C. S. Fox, S. Franco, et al. Executive summary: heart disease and stroke statistics—2014 update: a report from the american heart association. *Circulation*, 129(3):399–410, 2014.

[20] A. Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

[22] J. Guerra, J. Uddin, D. Nilsen, J. Mclnerney, A. Fadoo, I. B. Omofuma, S. Hughes, S. Agrawal, P. Allen, and H. M. Schambra. Capture, learning, and classification of upper extremity movement primitives in healthy controls and stroke patients. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pages 547–554. IEEE, 2017.

[23] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, et al. Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association. *Circulation*, 123(8):933–944, 2011.

[24] Y. Ishikawa, S. Kasai, Y. Aoki, and H. Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2322–2331, 2021.

https://doi.org/10.52202/068431-0122

[25] M. S. Jeffers, S. Karthikeyan, M. Gomez-Smith, S. Gasinzigwa, J. Achenbach, A. Feiten, and D. Corbett. Does stroke rehabilitation really matter? part b: an algorithm for prescribing an effective intensity of rehabilitation. *Neurorehabilitation and neural repair*, 32(1):73–83, 2018.

[26] A. Kaku, A. Parnandi, A. Venkatesan, N. Pandit, H. Schambra, and C. Fernandez-Granda. Towards data-driven stroke rehabilitation via wearable sensors and deep learning. In *Machine Learning for Healthcare Conference*, pages 143–171. PMLR, 2020.

[27] A. Kaku, K. Liu, A. Parnandi, H. R. Rajamohan, K. Venkataramanan, A. Venkatesan, A. Wirtanen, N. Pandit, H. Schambra, and C. Fernandez-Granda. Sequence-to-sequence modeling for action identification at high temporal resolution. *arXiv preprint arXiv:2111.02521*, 2021.

[28] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[29] S. Y. Kim, J. E. Hsu, L. C. Husbands, J. A. Kleim, and T. A. Jones. Coordinated plasticity of synapses and astrocytes underlies practice-driven functional vicariation in peri-infarct motor cortex. *Journal of Neuroscience*, 38(1):93–107, 2018.

[30] T. J. Kimberley, S. Samargia, L. G. Moore, J. K. Shakya, and C. E. Lang. Comparison of amounts and types of practice during rehabilitation for traumatic brain injury and stroke. 2010.

[31] J. W. Krakauer, S. T. Carmichael, D. Corbett, and G. F. Wittenberg. Getting neurorehabilitation right: what can be learned from animal models? *Neurorehabilitation and neural repair*, 26(8): 923–931, 2012.

[32] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014.

[33] H. Kuehne, J. Gall, and T. Serre. An end-to-end generative framework for video segmentation and recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pages 1–8. IEEE, 2016.

[34] G. Kwakkel, J. M. Veerbeek, E. E. van Wegen, and S. L. Wolf. Constraint-induced movement therapy after stroke. *The Lancet Neurology*, 14(2):224–234, 2015.

[35] C. E. Lang, J. R. MacDonald, D. S. Reisman, L. Boyd, T. J. Kimberley, S. M. Schindler-Ivens, T. G. Hornby, S. A. Ross, and P. L. Scheets. Observation of amounts of movement practice provided during stroke rehabilitation. *Archives of physical medicine and rehabilitation*, 90(10): 1692–1698, 2009.

[36] E. S. Lawrence, C. Coshall, R. Dundas, J. Stewart, A. G. Rudd, R. Howard, and C. D. Wolfe. Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population. *Stroke*, 32(6):1279–1284, 2001.

[37] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[38] P. Lei and S. Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6742–6751, 2018.

[39] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7083–7093, 2019.

[40] M.-T. Luong, I. Sutskever, Q. V. Le, O. Vinyals, and W. Zaremba. Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*, 2014.

[41] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6588–6597, 2019.

[42] Y. Murata, N. Higo, T. Oishi, A. Yamashita, K. Matsuda, M. Hayashi, and S. Yamane. Effects of motor training on the recovery of manual dexterity after primary motor cortex lesion in macaque monkeys. *Journal of neurophysiology*, 99(2):773–786, 2008.

[43] Y. B. Ng and B. Fernando. Human action sequence classification. *arXiv preprint arXiv:1910.02602*, 2019.

[44] Y. B. Ng and B. Fernando. Forecasting future action sequences with attention: a new approach to weakly supervised action forecasting. *IEEE Transactions on Image Processing*, 29:8880–8891, 2020.

[45] D. S. Nichols-Larsen, P. Clark, A. Zeringue, A. Greenspan, and S. Blanton. Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke*, 36(7):1480–1484, 2005.

[46] B. Ovbiagele, L. B. Goldstein, R. T. Higashida, V. J. Howard, S. C. Johnston, O. A. Khavjou, D. T. Lackland, J. H. Lichtman, S. Mohl, R. L. Sacco, et al. Forecasting the future of stroke in the united states: a policy statement from the american heart association and american stroke association. *Stroke*, 44(8):2361–2375, 2013.

[47] A. Parnandi, A. Kaku, A. Venkatesan, N. Pandit, A. Wirtanen, H. Rajamohan, K. Venkataramanan, D. Nilsen, C. Fernandez-Granda, and H. Schambra. Primseq: a deep learning-based pipeline to quantitate rehabilitation training. *arXiv preprint arXiv:2112.11330*, 2021.

[48] H. M. Schambra, A. R. Parnandi, N. G. Pandit, J. Uddin, A. Wirtanen, and D. M. Nilsen. A taxonomy of functional upper extremity motion. *Frontiers in neurology*, 10:857, 2019.

[49] D. Shao, Y. Zhao, B. Dai, and D. Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020.

[50] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1961–1970, 2016.

[51] Y. Souri, M. Fayyaz, L. Minciullo, G. Francesca, and J. Gall. Fast weakly supervised action segmentation using mutual consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] S. Stein and S. J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[53] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[55] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[56] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[57] Z. Wang, Z. Gao, L. Wang, Z. Li, and G. Wu. Boundary-aware cascade networks for temporal action segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 34–51. Springer, 2020.

[58] T. Wen and R. Keyes. Time series anomaly detection using convolutional neural networks and transfer learning. *arXiv preprint arXiv:1905.13628*, 2019.

[59] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018.

https://doi.org/10.52202/068431-0122

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see the datasheet J.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see the datasheet J.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included the code in supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix G
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See bottom table in Table 2
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G.4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We provide citations for all baselines.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code that is necessary to reproduce this work in the appendix. Upon acceptance, we will release code on Github.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix datasheet J
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Appendix datasheet J
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Appendix datasheet J
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See Appendix datasheet J
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/068431-0122