Automated Movement Assessment in Stroke Rehabilitation

Tamim Ahmed[1]*, Kowshik Thopalli[2]*, Thanassis Rikakis[1], Pavan Turaga[2], Aisling Kelliher[3], Jia-Bin Huang[4] and Steven L. Wolf[5]

[1] Department of Biomedical Engineering, Virginia Tech, Blacksburg, VA, United States, [2] Geometric Media Lab, School of Arts, Media and Engineering, Arizona State University, Tempe, AZ, United States, [3] Department of Computer Science, Virginia Tech, Blacksburg, VA, United States, [4] Department of Electrical and Communication Engineering, Virginia Tech, Blacksburg, VA, United States, [5] Department of Rehabilitation Medicine, Emory University, Atlanta, GA, United States

We are developing a system for long term Semi-Automated Rehabilitation At the Home (SARAH) that relies on low-cost and unobtrusive video-based sensing. We present a cyber-human methodology used by the SARAH system for automated assessment of upper extremity stroke rehabilitation at the home. We propose a hierarchical model for automatically segmenting stroke survivor's movements and generating training task performance assessment scores during rehabilitation. The hierarchical model fuses expert therapist knowledge-based approaches with data-driven techniques. The expert knowledge is more observable in the higher layers of the hierarchy (task and segment) and therefore more accessible to algorithms incorporating high level constraints relating to activity structure (i.e., type and order of segments per task). We utilize an HMM and a Decision Tree model to connect these high level priors to data driven analysis. The lower layers (RGB images and raw kinematics) need to be addressed primarily through data driven techniques. We use a transformer based architecture operating on low-level action features (tracking of individual body joints and objects) and a Multi-Stage Temporal Convolutional Network(MS-TCN) operating on raw RGB images. We develop a sequence combining these complimentary algorithms effectively, thus encoding the information from different layers of the movement hierarchy. Through this combination, we produce a robust segmentation and task assessment results on noisy, variable and limited data, which is characteristic of low cost video capture of rehabilitation at the home. Our proposed approach achieves 85% accuracy in per-frame labeling, 99% accuracy in segment classification and 93% accuracy in task completion assessment. Although the methodology proposed in this paper applies to upper extremity rehabilitation using the SARAH system, it can potentially be used, with minor alterations, to assist automation in many other movement rehabilitation contexts (i.e., lower extremity training for neurological accidents).

Keywords: stroke rehabilitation, automation, cyber-human intelligence, HMM, MSTCN++, transformer, segmentation, movement assessment

1. INTRODUCTION

As the US and global populations age, we observe an increasing need for effective and accessible rehabilitation services for survivable debilitating illnesses and injuries, such as stroke and degenerative arthritis (1, 2). Effective rehabilitation requires intensive training and the ability to adapt the training program based on patient progress and therapeutic judgment (3). Intensive and adaptive rehabilitation is challenging to administer in an accessible and affordable way; high intensity therapy necessitates frequent trips to the clinic (usually supported by a caregiver), and significant one-on-one time with rehabilitation experts (4). Adaptation requires a standardized, evidence-based approach, coordinated amongst many specialists (5–7). Active participation by the patient is also critical for improving self-efficacy and program adherence (8), although, without significant dedicated effort from a caregiver, in many cases, active participation and adherence are difficult to achieve (9).

Telemedicine and telehealth are gaining significance as viable approaches for delivering health and wellness at the home and in the community at scale (10). Applying existing telemedicine approaches to physical rehabilitation in the home is not yet possible, owing to the challenges of automating the observation, assessment, and therapy adaptation process used by expert therapists. For upper extremity rehabilitation for stroke survivors, which is the focus of this paper, more than 30 low-level movement features need to be tracked as the patient performs functional tasks in order to precisely and quantitatively characterize movement impairment (5). High precision sensing and tracking systems can work well in spacious and supervised clinical environments, but are currently not yet appropriate for a typical home setting. The use of marker-based tracking systems or complex exoskeletons are simply too expensive, challenging to use, and obtrusive in the home (11–13).

Banks of video camera arrays can seem intrusive in the home and lead patients and/or their families to feel as if they are under surveillance (6). More promisingly, networks of wearable technologies (e.g., IMUs, smart skins, pressure sensors) can provide useful tracking data for overall movement and detailed features, but they can also be hard to put on correctly, irritating to wear for long periods of time, and sometimes require a perceived excessive number of wearables to capture all movement features correctly (14). A final concern with respect to the patient's home environment concerns the physical footprint of any technology introduced into their home. Disturbing the home setting can be understood negatively, which has the knock-on effect of reducing adoption by stroke survivors and/or other family members in the home (15).

That said, accurate, low-cost capture of movement data is only part of the challenge. Automation of assessment is also difficult because the processes used by therapists are largely tacit and not well standardized (16). Clinicians are trained to use validated clinical measures (e.g., Fugl-Meyer, ARAT, and WMFT), utilizing a small range of quantitative scales (0–2, 0–3, and 0–5, respectively) for assessing performance of functional tasks that map to activities of daily life (17, 18). These scales provide standardized rubrics for realizing

this assessment and use a uniform activity space with exact measurements for each subcomponent of the space to assist the standardized performance of the tasks. The high-level rating of task performance provided by experts through these measures is difficult to map directly to specific aspects of movement and related detailed kinematics extracted through computational means. Even expert clinicians cannot simultaneously observe all aspects of upper extremity pathological movement or compare such observations to a standardized value. There is considerable evidence showing that individual therapists direct their attention toward different elements and assess them differently when evaluating performance *in situ* and real-time, or when later rating videos of performance (19–22). The relation of movement quality to function is therefore difficult to ascertain in a standardized quantitative manner (5, 16). This results in approaches for structuring and customizing therapy that are partly based on subjective experience rather than a standardized quantifiable framework (16, 20). In turn, this results in a lack of large-scale data on the structuring and customization of adaptive therapy, and on the effects of customization and adaptation choices on functionality in everyday life (23). Therefore, full scale automation of the real-time functions of the therapist at the home is not yet feasible.

In light of these limitations, we are developing the novel Semi-Automated Rehabilitation At the Home (SARAH) system. The SARAH system comprises two video cameras, a tablet computer, a flexible activity mat, and eight custom-designed 3D printed objects, as shown in **Figure 1** below. The objects are designed to support a broad range of perceived affordances (24), meaning they can be gripped, moved, and manipulated in a wide variety of ways (25). Each object is unique in terms of size and color to assist identification of objects by patients, and to enable easier identification and tracking using computer vision methods. The flexible activity mat is screen-printed with high-contrast guidance lines indicating to the patient the four primary activity spaces (near and distal ipsilateral and contralateral) through increasingly colored lines. In addition, four rows of circles on the mat assist the computer vision system with boundary detection between activity space for consistently analyzing patient activities.

The system can be easily installed on a regular kitchen or living room table. The two cameras initiate and record only during training, and they are activated and controlled by the patient using a custom-designed application on the tablet computer. The system aims to integrate expert knowledge with data driven algorithms to realize coarse real-time automated assessment of movement of stroke survivors during therapy at the home (26). This assessment can drive high-level feedback on results and performance after execution of each training task, to assist patients with self assessment, and to help them plan their next attempt(s) of the training task. Daily summaries of the interactive training will be transmitted to remote therapists to assess overall progress (within and across sessions), adjust the therapy structure, and provide text or audio based feedback and directions to the patient via the tablet. Continuous and effective training monitoring accompanied by feedback on the patient's immediate performance, combined with expert customization of

therapy to their needs and learning styles, increases the likelihood of patients adopting home-based rehabilitation systems (8).

The SARAH system is optimized for daily at home therapy of stroke survivors with moderate and moderate-to-mild impairment. Stroke survivors who score above 30 on the Fugl Meyer test and can initiate even minimal movement into extension of the elbow, wrist and digits when discharged from the clinic post stroke, can show significant improvement in function through repetitive training lasting 2 weeks to 6 months (27–29). Currently, third party insurance in the US provides financial support for up to 6 months of outpatient and/or in home therapy after release from the clinic. Patients with moderate impairment and a Mini Mental score >25 can follow instructions given by the SARAH system, use their unimpaired limb to control the SARAH system and can engage (at least partially) all SARAH objects (25). The SARAH system promotes active learning by the patient who is expected to interpret the coarse feedback provided by the system and to plan the next action so as to improve her performance. Furthermore, the generalizable design of the objects and tasks of SARAH encourages the patient to actively map the task to multiple ADLs. These active learning characteristics of the SARAH system, combined with the variety of included tasks in the system makes the system feasible for in-home rehabilitation lasting 2–8 weeks (25, 26).

This paper focuses on our hybrid knowledge-based data-driven approach to automated assessment of human movement in the home. Our approach leverages expert rubrics for standardized rating of overall task performance to inform automated rating of movement performance based on low cost, limited, noisy, and variable kinematic data. The assessment process and outcomes need to be compatible with therapist assessment approaches so as to assist remote therapists in using summaries of the computational assessment when remotely monitoring progress and structuring therapy at the home. Our approach has two components: (i) making the expert raters process as observable as possible; and (ii) leveraging the expert rating process to inform the structuring and improved performance of computational algorithms. In previous publications, we have presented in detail our research and

development activities for the first component (26, 30, 31). Inspired by clinical measures for rating rehabilitation movement, we developed the SARAH system to utilize a standardized activity space with eight well defined sub-spaces that are drawn as bounding boxes on the video capture of therapy (see **Figure 2**). We designed the SARAH training objects to facilitate generalized mapping of training tasks to ADLs for the patients, while also facilitating tracking through low-cost video cameras (25). We used participatory design processes and custom-designed interactive video rating tools to help expert therapists reveal and reflect on their rating process and internalized (tacit) rating schemas (26).

In order to manage the complexity of real-time movement observation and to make generalizable observations across different therapy tasks, therapists tend to segment tasks into a few segments that can be combined in different sequences to generate targeted therapy tasks. Even though most therapists use intuitive segmentation of movement for observation and assessment, the segment vocabulary is *not* standardized. We worked with expert therapists to standardize the segment vocabulary into a state machine that can produce all 15 tasks of the SARAH system (26). The segments are: Initiation + Progression + Termination (IPT), Manipulate and Transport (MTR), Complex Manipulation and Transport (CMTR), and Release and Return (RR). As an example, a drinking related task can be described by the following codification: subject reaches out and grasps a cone object (IPT) and brings it to their mouth (MTR), then returns the object to the original position (MTR), and releases the object and returns the hand to the rest position (RR).

To make the assessment of segments in real-time manageable, the therapist significantly limits the features observed per segment. This limitation is achieved by using their own experience to develop a probabilistic filtering of irrelevant low-level features for a segment (i.e., digit positioning is likely not that relevant to movement initiation), and probabilistic composite observation of relevant features (i.e., a strategy for quick impressions of shoulder and torso compensation during movement initiation). This process is not well standardized as the filtering and compositing activities are based on individual

FIGURE 1 | (A) SARAH system and objects setup. **(B)** SARAH activity mat.
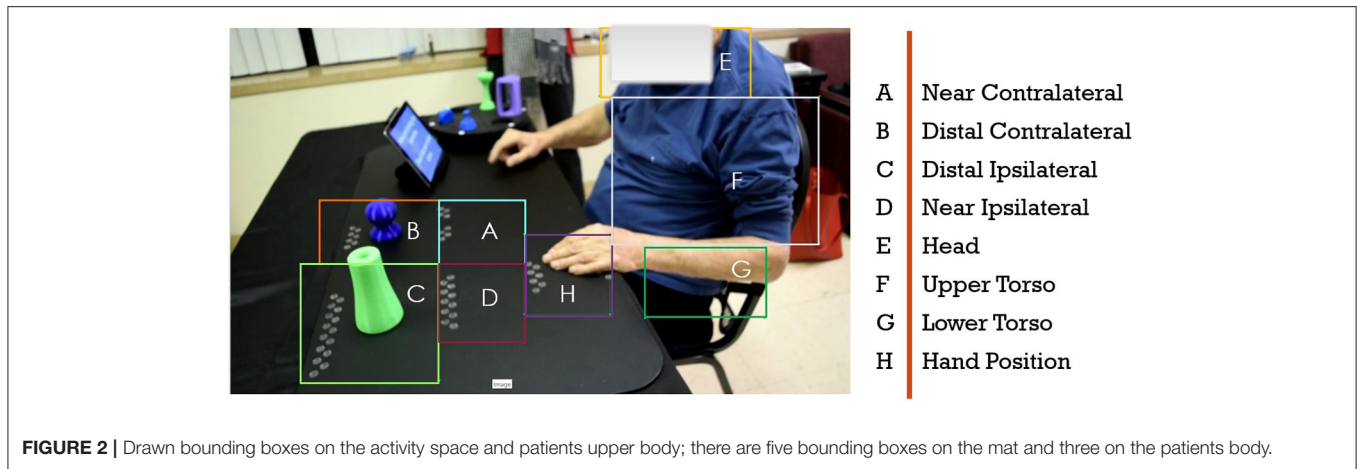
FIGURE 2 | Drawn bounding boxes on the activity space and patients upper body; there are five bounding boxes on the mat and three on the patients body.

experience and training. We further worked with expert therapists to define a consensus-limited set of composite movement features that are important when assessing the performance of each segment in our model (26). For example, the resulting rubric identifies four key features to assess during the Complex Manipulation and Transport stage: (i) appropriate initial finger positioning, (ii) appropriate finger motion after positioning, (iii) appropriate limb motion following finger positioning, and (iv) limb trajectory with appropriate accuracy. The rubric also establishes operational definitions of terms used to evaluate movement quality and inform rating. For example, the word "appropriate" used in the above instructions is defined as "the range, direction, and timing of the movement component for the task compared to that expected for the less impaired upper extremity." Although therapists do not explicitly track and assess raw kinematic features, in previous work we proposed computational approaches connecting therapist's assessment of composite features to computationally tracked raw kinematics (5, 32).

2. MATERIALS AND METHODS

2.1. Analysis Framework

Based on the background discussed above, we propose a cyber-human movement assessment hierarchy for upper extremity rehabilitation. The five layers of the hierarchy (listed top to bottom) are: overall task rating, segment rating, composite feature assessment, raw kinematics, and raw RGB images (see also **Figure 3**). As we move down the hierarchy, the expert knowledge becomes less observable and less standardized. Computational knowledge works in reverse, with more confidence in the lower levels (raw kinematics or derivatives) that gradually diminishes moving up in the hierarchy toward complex decision-making relying on expert experience (e.g., overall task rating). The proposed hierarchy aims to reveal the grammar (or compositional structure) of therapy, from the signal level to the meaningful activity level (training tasks). Knowing the composition rules of a complex human activity, whether that is language, sport, or a board game like
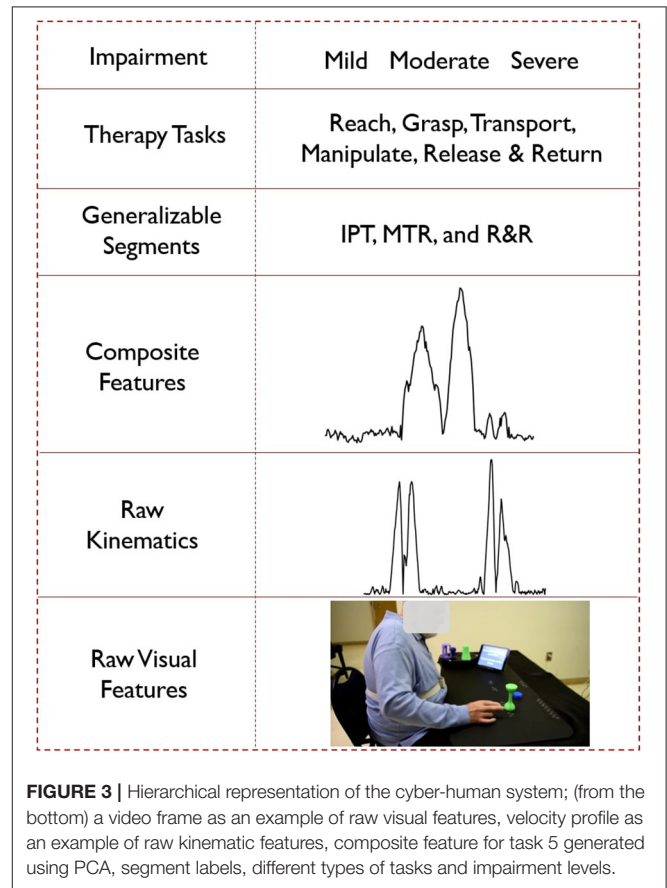
FIGURE 3 | Hierarchical representation of the cyber-human system; (from the bottom) a video frame as an example of raw visual features, velocity profile as an example of raw kinematic features, composite feature for task 5 generated using PCA, segment labels, different types of tasks and impairment levels.

chess, facilitates the meaningful computational analysis of the activity in a manner that is comprehended and leveraged by expert trainers and trainees (33, 34). Although the hierarchy proposed in this paper is established for upper extremity rehabilitation using the SARAH system, it can potentially be used, with minor alterations, to assist automation in many other movement rehabilitation contexts (i.e., lower extremity training for neurological accidents). Our proposed methodology can also

transfer to other complex human activity contexts (i.e., training firefighters or athletes).

In the following sections, we show how we leverage the observable expert knowledge of the higher levels of the hierarchy to improve the performance of computational algorithms using raw kinematics and visual features for automated task segmentation, segment classification, and task performance assessment at the home. As the patient selects the task they want to attempt using the tablet computer, the system knows the desired sequence of segments for satisfactory performance of the task. The expected topology of each segment of each task within our standardized activity space is also specified. We can thus utilize a feature-matching approach to measure the "distance" between the observed sequence of segments by the system and the expected sequence of states for the performed task to determine whether the task was successfully performed.

The state-time characteristics of our hierarchy (i.e., segment state machine) motivates our use of automated segmentation though a hidden Markov model (HMM). We also leverage recent advances in machine-learning models for processing video data. Specifically, we tested the use of transformers (35) operating on low-level action features (tracking of individual body joints and object tracking). We also tested machine learning algorithms that operate on raw RGB images (MS-TCN) (36). We chose the MS-TCN algorithm because its classifiers are trained on recognizing human activity based on kinetics and thus can synergize well with automated assessment of task performance in rehabilitation. All three types of algorithms required slightly different features for achieving best performance in connecting the raw data layers to the segment layer (and achieving automated segmentation). In addition, the features used by the algorithms were different from the composite features used by the human experts. This established the segments layer as the key integration layer for different assessment approaches, and reinforced the important role of a generalizable segment vocabulary proposed by expert therapists. Each approach showed different strengths and weaknesses, therefore we utilized an ensemble approach across the algorithms to finalize the automated segmentation and segment classification decisions. Automated assessment of therapy at the home using the SARAH system will need to be realized through noisy and variable data because of the low cost infrastructure, the varying environments of installation in the home, and the significant variation of movement impairment of stroke survivors (30). Furthermore, at the early stages of implementation of the SARAH system the data sets will be limited. To explore the resilience of our methodology we recreated these conditions in a clinic and gathered a limited data set for testing the algorithms. As we show in detail in the next section, even though the individual data layers of our cyber-human hierarchy are noisy, the composite automated decisions (successful performance of individual tasks and sequence of tasks assigned) produced across layers are robust. In future work, we will explore the expanded use of this approach for detailed automated assessment of elements of movement quality and the relation of movement quality to functionality.

2.2. Data Collection

We used an earlier version of the SARAH system (37) to record videos of nine stroke survivors (seven men and two women) performing the first 12 SARAH tasks. The nine stroke survivors had different levels of impairment ranging from mild/moderate (Fugl-Meyer score >40) to moderate impairment (Fugl-Meyer score >30) and had different types of movement challenges. Thus, our overall dataset represents good range of conditions and movement variability. The participants in the study were asked to attempt each of the first 12 SARAH tasks, repeating each task four times if possible.
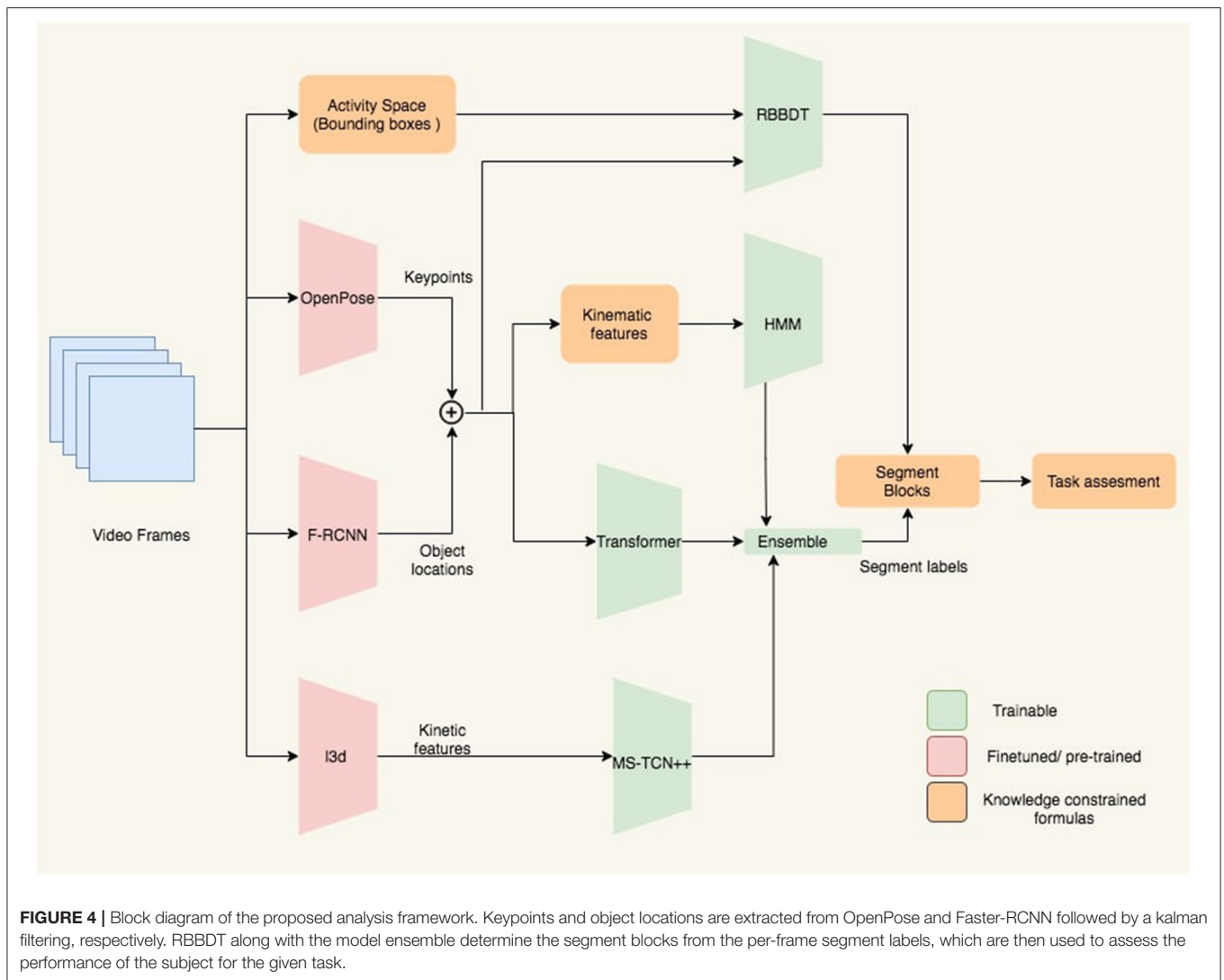
All patients could follow the instructions given by the SARAH system and use their unimpaired limb to control the system. All patients could engage all objects of the SARAH system (25, 26) at least minimally using the impaired limb. Most patients could not perform the final two of the 12 SARAH tasks since the last two are the most difficult. For this reason, in our analysis below, we only use 450 videos of the individual performances of the first 10 tasks by the nine patients. When we begin to implement the SARAH system in the home we expect patients to gradually engage the more complex SARAH tasks as their training at home progresses.

The movement of the patients in this study was recorded using one consumer grade video camera (we used the side view camera of the SARAH system to capture the profile of the body and impaired arm as this is the preferred viewing point of the therapist). The camera placement instructions were relatively high-level to ensure that they could be implemented quickly and without interfering with the patient's therapy session. The research assistant collecting the videos was asked to place the camera on the side of the impaired limb of the patient. The camera had to be far enough so as to not interfere with the performance of the tasks and be able to capture the full upper body of the patient. No specific distance, height or viewing angle were given for camera placement. As expected, this process provided minimal interference and could be realized very quickly but produced high variability in the captured videos in terms of location of the patient, activity space in the image, lighting, camera height, and viewing angle.

All collected videos were rated by expert therapists using a custom developed SARAH rating interface that allows therapists to rate the task, segment and composite movement features of each video recording in an integrated manner (26). Therefore, there are at least two expert ratings per recorded video. These expert ratings can be used as the ground truth for comparing the effectiveness of the computational analysis algorithms.

3. AUTOMATED SEGMENTATION AND SEGMENT CLASSIFICATION

Our segmentation and classification framework is illustrated in **Figure 4**. It has three sets of blocks denoted with different colors in the figure. The first set of blocks are fine-tuned or pre-trained models that we implement as feature extractor. The second set of blocks include different algorithms such as HMM, Transformer, MSTCN++, and RBBDT. The third set of blocks are knowledge constrained formulas that we use as feature extractor for the

FIGURE 4 | Block diagram of the proposed analysis framework. Keypoints and object locations are extracted from OpenPose and Faster-RCNN followed by a kalman filtering, respectively. RBBDT along with the model ensemble determine the segment blocks from the per-frame segment labels, which are then used to assess the performance of the subject for the given task.

RBBDT, HMM, and also use to predict segment blocks and do task assessment. The HMM, Transformer, and MSTCN++ blocks generate per-frame state probabilities from the input video data. We generate the per-frame segment labels by a fusion of the state probabilities. With the incorporation of the design constrained denoising and the candidates from the RBBDT block (the decision tree), we calculate segment blocks from the per-frame ensemble predictions. Finally, we calculate task performance assessment scores using the segment blocks. The following subsections cover the detailed description of different blocks in the analysis framework.
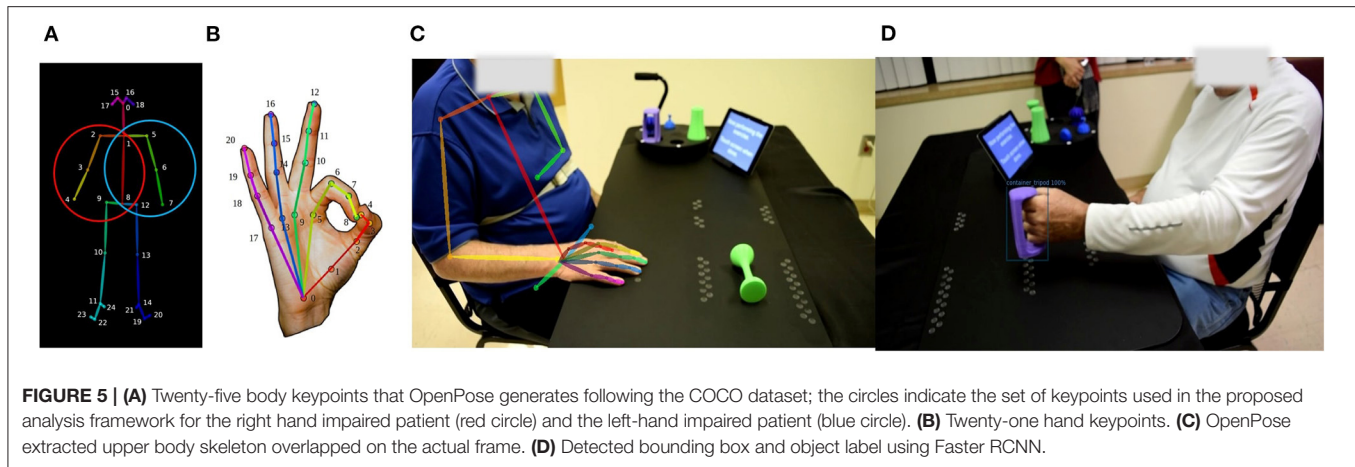
3.1. RGB Image Pre-processor

The transformer, the HMM, MSTCN++, and the rule-based decision tree pipelines have a feature extractor as demonstrated in **Figure 4** that converts the raw RGB frames into suitable input features. The frames capture the activity space that includes the patients upper body, the objects and the activity mat. All 12 tasks in the SARAH system involve movement in relation with

one or two objects. Therefore, it is imperative to detect key movement features associated with both the patient's upper body and the objects.

3.1.1. Patient Skeleton Detection

We use OpenPose (38), an open source pose estimation technique, to extract 2D patient skeletons from videos. OpenPose generates 135 keypoints per-frame that include 25 body keypoints (**Figure 5A**), 21 keypoints for both hand (**Figure 5B**), and 70 keypoints for the face. These keypoints are the (x, y)-pixel coordinates of the skeleton joints as shown in **Figure 5C**. In the presence of multiple persons in the captured frame, we measure the area of the estimated upper body skeleton and consider the person with the largest area. We find that this simple pre-filtering approach worked in all cases within our test set to gives us the actual patient's keypoints. In the SARAH system, the side camera used in this experiment always focuses on the impaired arm and the non-impaired arm is obscured. Therefore, we exclude the estimated keypoints from the non-impaired limb

FIGURE 5 | (A) Twenty-five body keypoints that OpenPose generates following the COCO dataset; the circles indicate the set of keypoints used in the proposed analysis framework for the right hand impaired patient (red circle) and the left-hand impaired patient (blue circle). **(B)** Twenty-one hand keypoints. **(C)** OpenPose extracted upper body skeleton overlapped on the actual frame. **(D)** Detected bounding box and object label using Faster RCNN.

in our analysis framework. Since we are only interested in the upper body keypoints, we only consider keypoints above the lower torso line (keypoints 9, 8, and 12). Also, because of the non-standardized placement of the camera in this experiment, the face was not always within the frame and therefore, we excluded the face keypoints from further analysis. After all the exclusions, we use the two sets of keypoints as indicated in **Figure 5A**. The set of keypoints inside the red circle is for right-hand impaired patients and the blue one for the left-hand impaired patients. All keypoints are normalized with respect to the image frame.

3.1.2. Object Detection and Tracking

It was also imperative to obtain high quality object identification and location data for every frame, as the type of object being used and the relationship between the objects and the keypoints extracted from the subjects play an important role in determining segment classification. To this effect, we fine-tuned an object detection algorithm which classifies the objects in the frame, places a bounding box around the classified object, and provides the object/bounding box location relative to the overall frame. We considered a Faster-RCNN model (39) pre-trained on MS COCO (40) dataset for our experiments. To fine tune this object detection model we first labeled all objects being used across the 12 performed SARAH tasks from a small number of videos using CVAT, an open-source annotation tool. In total we had a training set of around 23,253 training images and we validated it on 11,112 images with a mean IOU score of 0.67. We then used the open source detectron 2 framework (41), built on top of PyTorch (42) to fine-tune the model.

As stated in section 2.1, there is a high variability in the training data, owing to issues such as use of a single camera view, non-standardized camera angles, variability in the patient's movement, unconstrained ways of grasping the objects, errors in transportation etc. As a consequence, the object detection model misses certain objects in certain frames due to factors such as occlusions or dropped objects, and sometimes even misclassifies certain objects due to partial visibility. However, as described in section 1, the task being attempted in each captured video is known to the algorithm and each task is associated with specific

objects. We use this knowledge to filter out misclassifications. To further improve the smoothness of the trajectories of the objects, we use a simple Kalman filter based tracking algorithm SORT (43). Given the fine-tuned object detection model, the pipeline constructed is as follows. First, we extract all the frames in the video as images. Second, the trained Faster-RCNN is run on each of the frames thus collecting per-frame object data. Third, misclassifications are corrected automatically and the trajectories of the tracked objects are smoothed.

3.1.3. Denoising Pose Keypoints and Object Locations

The activity space in different videos has high variability. As the camera angle and patient position are different for each captured session, the ratio of the activity space to the frame resolution changes. Low camera resolution and low frame-rate (used so as to allow efficient capture of long sessions), variable focus (cameras were set up by non-experts without standardized instructions) and varied lighting conditions (typical of therapy sessions happening in different contexts), create challenges for pose-estimation methods from the raw frames. As a result, the keypoint detection and bounding box estimation accuracy can suffer, sometimes even completely missing all keypoints and not detecting any bounding box in a frame. To reduce the missing data problem we first detect the outliers. We calculate the z-score for each sample using,

$$ zscore_i = \frac{x_i - mean(x)}{std(x) + \epsilon}; i = 1, 2, \ldots, N \qquad (1) $$

where x_i is the keypoints or object locations for the ith frame (an $\epsilon > 0$ is used to guard against pathological situations). If the z-score value is higher than a threshold, we consider the sample as an outlier. For our case, using a threshold of 2.5 gives good performance. Then we use spline interpolation to fill the missing values or replace the outliers.

3.1.4. Normalization of Input Data

Denoising and normalizing the input data is a vital step in achieving good performance through machine learning. We

experimented with three different techniques to find the best techniques for normalizing both the patient keypoints data and the object locations—(i) global normalization with respect to the image frame, (ii) normalization with respect to the activity mat area, and (iii) normalization using a computed homography matrix.

Global-Norm: Here, we just divide the raw patient keypoints and object location data with the image frame's width and height, thus mapping the data to a $[0, 1]$ range. In practice due to the high amount of noise in the data, and variability in inter-patient performance, this global normalization fails to achieve good results and was therefore not applied (see **Table 6**).

Mat-Norm: Here, we normalize the data with respect to the ratio of the area of the activity mat to the total image area. This is done by labeling the activity mat in one of the frames per subject per capture session. For every capture session the camera position is set at the beginning. We could thus precompute the area of the mat for all recorded tasks of each capture session. This normalization with regards to the mat area removes a good degree of variability and thus performs better than global normalization.

Homography-Norm: Since the camera angles across subjects and across attempts vary considerably,it is important to reduce the impact of this variation. To achieve this, we apply a simple homography-based re-projection of the object and keypoint data. First, a homography matrix is pre-computed that transforms data from image frame to the coordinate frame defined by the activity mat. This transformation matrix is then applied to the object and keypoint data. Even though this achieves high invariance, it also removes discriminative features and thus there is not any observably significant improvement in the performance of the transformer models.

3.2. HMM Features and Algorithmic Pipeline

3.2.1. Kinematic Feature Extractor

Using the denoised keypoints and object locations, we calculate 20 kinematic features that, based on our prior work (5), can be combined to successfully assess functionality and movement quality in upper extremity stroke rehabilitation. These features can be used to analyze the movement of the object, the affected limb and torso, and the patterns of human-object interaction. These 20 features are summarized in **Table 1**. The variability in stroke survivor movement and the low quality capture resulted in 7 of the 20 features being too noisy to support automated analysis. We therefore only used the 13 less noisy features for further analysis in the paper. In addition to these 13 features, we also calculate derivatives of some of the features. The derivatives introduce oscillations unique to segments and by learning those patterns the efficacy of the HMM in per-frame segmentation improves. A detailed description of the kinematic features are given in the **Supplementary Material**.

3.2.2. Hidden Markov Model With Prior Transitions

A hidden Markov model (HMM) (44) consists of a set of states, a transition model, and an observation model. Denote $S = S_1, S_2, \ldots, S_N$ as the N states. Evolution between states is modeled

TABLE 1 | List of kinematic features extracted from patient 2D keypoints and object co-ordinates.

Patient features	Object features
Limb location	Object location
Limb velocity	Object velocity
Limb absolute velocity	Object absolute velocity
Distance between limb and Torso Line	Distance between two Objects
Pronation and supination	Object angular velocity
Relative angle between torso line and limb	Distance between object and limb
Hand openness	Distance between object and torso line
Torso sway, rotation, and lean	
Elbow extension	
Jerkiness	
Absolute limb angle	

by the transition probability conditioned on the previous state. The transition probability between states is represented by a matrix, $A = a_{ij}$; where $a_{ij} = Pr(S_j(t + 1)|S_j(t))$. For a given observation, $O = O_1, O_2, \ldots, O_k$, the emission probability distribution is represented by a matrix $B = b_j(k)$; where $b_j(k)$ is the probability of generating observation O_k when the current state is S_j, where k is the total number of observation symbols.
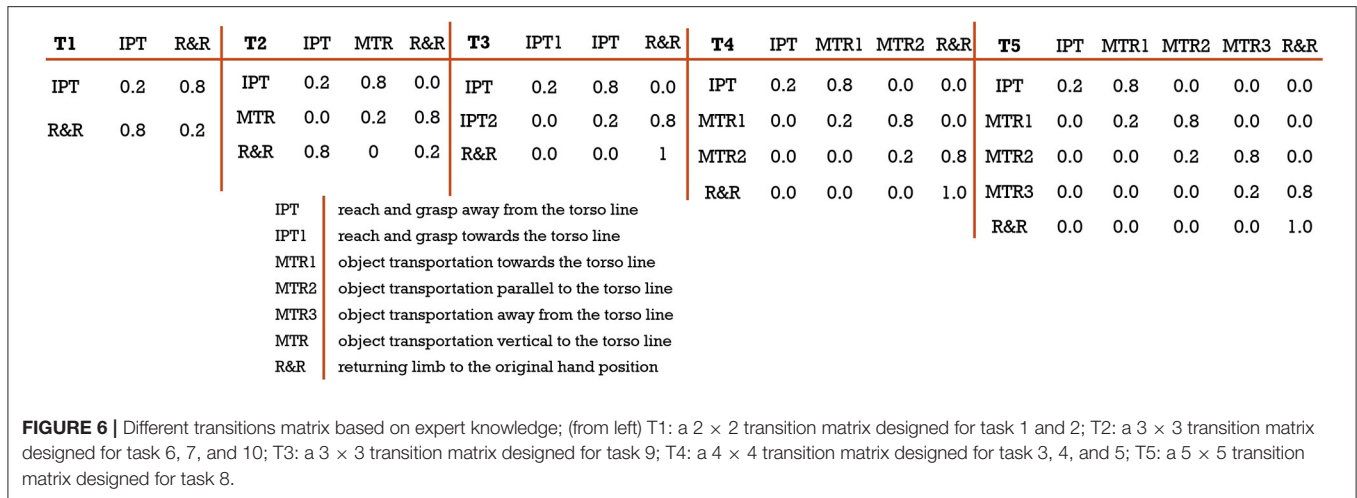
HMMs have been used in movement quality assessment because of their good performance in detecting subtle inconsistencies in the movement (45–47). However, the topological structure of the HMM in most cases cannot be automatically determined due to the highly variant and small dataset (47). We therefore use expert knowledge based design constraints to model the topological structure of the HMM for different exercises. We have modeled 5 different transition matrices demonstrated in **Figure 6** that are based on the number of unique segments present in the 12 SARAH tasks. These matrices provide a strong prior as they are modeled after the therapist's approach to parsing movement and focusing attention on key composite movement features per type of segment.

To train the HMM, we first need to define the initialization for each state in the transition matrix. For example, to initialize an HMM with T4 transition matrix, we need four states: $S = \{IPT, MTR1, MTR2, R\&R\}$. Each state is initialized with a normal distribution of mean, $\mu = 0.5$, and standard deviation, $\sigma = 0.1$. We choose the normal distribution because it produced the lowest chi-square error after intensive experimentation with different distributions like beta, exponential, gamma, log-normal, normal, Pearson3, triangular, uniform, and Weibull. The distribution parameters can be randomly initialized, and we find that the model is not sensitive to random initialization. The training of the HMM tunes the parameters which are then compactly represented as,

$$\lambda_{T,F} = (\pi, A_T, B_{T,F}), where \ T = 1, 2, 3, 4, 5 \ and \ F = 1, 2, \ldots, F \tag{2}$$

Here, the subscript T and F refer to different transition matrices and features and π is the initial state distribution and is defined

T1	IPT	R&R
IPT	0.2	0.8
R&R	0.8	0.2

T2	IPT	MTR	R&R
IPT	0.2	0.8	0.0
MTR	0.0	0.2	0.8
R&R	0.8	0	0.2

T3	IPT1	IPT	R&R
IPT	0.2	0.8	0.0
IPT2	0.0	0.2	0.8
R&R	0.0	0.0	1

T4	IPT	MTR1	MTR2	R&R
IPT	0.2	0.8	0.0	0.0
MTR1	0.0	0.2	0.8	0.0
MTR2	0.0	0.0	0.2	0.8
R&R	0.0	0.0	0.0	1.0

T5	IPT	MTR1	MTR2	MTR3	R&R
IPT	0.2	0.8	0.0	0.0	0.0
MTR1	0.0	0.2	0.8	0.0	0.0
MTR2	0.0	0.0	0.2	0.8	0.0
MTR3	0.0	0.0	0.0	0.2	0.8
R&R	0.0	0.0	0.0	0.0	1.0

IPT	reach and grasp away from the torso line
IPT1	reach and grasp towards the torso line
MTR1	object transportation towards the torso line
MTR2	object transportation parallel to the torso line
MTR3	object transportation away from the torso line
MTR	object transportation vertical to the torso line
R&R	returning limb to the original hand position

FIGURE 6 | Different transitions matrix based on expert knowledge; (from left) T1: a 2 × 2 transition matrix designed for task 1 and 2; T2: a 3 × 3 transition matrix designed for task 6, 7, and 10; T3: a 3 × 3 transition matrix designed for task 9; T4: a 4 × 4 transition matrix designed for task 3, 4, and 5; T5: a 5 × 5 transition matrix designed for task 8.

as $\pi = \pi_i$ where π_i is the probability of state S_i being the initial state. Based on the tasks, $\pi = [1, 0, \dots]$. Since all the tasks start with the state IPT, the first state probability is always 1 followed by zeros for other states present in the transition matrix. This is another strong design prior we leverage, in the form of supervision of the labels with state names. In this case, maximum likelihood estimation (MLE) estimates the emissions from data partitioned by the labels and the transition matrix is calculated directly from the adjacency of labels. Then, various transition matrices and their initialization are further tuned with patient data. As we provide more training data, the final models are expected to become more generalizable across different types of tasks. Another prior we use for structuring our HMM relates to the use of distributions of individual kinematic features. Therapists establish through experience an expected distribution per feature per segment (i.e., Gaussian velocity profile for IPT). Therefore, to learn the distribution parameters for each of the kinematic features, it is a requisite to train separate HMMs for each of those features.

We use the open source toolbox Pomegranate (48) for the training and estimation of the state probabilities for each feature. Then, we formulate an objective function on the state probabilities of the kinematic features to estimate one set of state probabilities. For this purpose we use a mean squared error based objective function defined as,

$$J_i = \sum_{j=1}^{N} \left(\hat{p}_{ij} - \frac{\sum_{k=1}^{F} w_i^k p_{ij}^k}{\sum_{k=1}^{F} w_i^k} \right)^2 ;$$
$$j = 1, 2, \dots, N; i = 1, 2, \dots, T_{max}; k = 1, \dots, F \quad (3)$$

where, $\hat{p}$ denotes the per-frame categorical labels and p is the predicted state probabilities. N is the total number of observations in the training set, T_{max} is the maximum observable frame number, F represents the total number of features, and w is the weight for optimization. We optimize the objective function by adding two constraints on the weights. Firstly, restricting the range of weights $w \in [0, 1]$ and secondly, constraining the weights to unit sum: $\sum_{k=1}^{F} w_i^k = 1$. The objective function was minimized using Sequential Least Square Programming (SLSQP) (49) due to its ease of implementation and because our objective function naturally is an instance of this type of problem.

3.3. Transformer Pipeline

In addition to HMMs, we also explore deep-learning based transformers (35), which have attracted a large amount of interest in the natural language processing community due to their strengthes in modeling long term dependencies while being computationally efficient and avoiding problems such as vanishing gradients in other deep-learning based time-series approaches like long short-term memory (LSTM) based approaches. Since our goal is video-segmentation, and per-frame classification, we model this problem as a Sequence-to-Sequence problem (50), where the input is a multivariate time-series determined by body keypoints and object data. Given this data, we output discrete segment labels for every timestep (frame). The pipeline is as follows: (i) concatenate normalized pre-processed zero-padded keypoints and object location data to obtain $X \in \mathbb{R}^{N \times F \times T_{max}}$ where N is the number of videos, F is the total number of feature and T_{max} is the maximum observable frame number including the zero-padding, (ii) partition the data X into training X^{tr}, validation X^{val}, and test X^{te} sets randomly, with a constraint that no patient-task pair should be repeated in train and test sets, (iii) train the transformer with the objective to find a function g that maps X^{tr} to Y, where $Y^{tr} \in \mathbb{R}^{1 \times Ts}$ is the per-frame label vector, by minimizing cross entropy loss between $f(g^{tr})$ and Y^{tr}. We train the network with different augmentations on the input data such as random drift, Gaussian noise etc using (51) to avoid overfitting.

3.3.1. Transformer Architecture

We adopt the transformer architecture from (35), with minor modifications to adapt it for timeseries (52): (i) we replace the embedding layer with a generic layer; and (ii), we apply a window on the attention map to focus on short-term trends because the segments are short. Other hyperparameters are as follows: number of encoder-decoder pairs are set to 2, number of heads as

4, the dimension of key query and value and an attention window of 100 along with sinusoidal positional encoding.

3.4. MSTCN++ Pipeline

Temporal convolutional networks (TCNs) (53) are shown to outperform methods such as LSTMs on various time-series modeling problems. Building on the success of TCNs, (36) recently developed MS-TCN a multi-stage variant, by stacking multiple TCNs. This multi-stage approach is shown to outperform TCNs for video segmentation. We have thus used MS-TCN (36) and its variants MS-TCN++ (54) for our experiments. While our transformer and HMM based methods operate on carefully constructed features from keypoints and objects, our TCN approach can directly use features extracted by a feature-learner approach such as a pre-trained video classification network, I3D (55). I3D is trained on the Kinetics dataset (56), which contains videos of human actions, with classes such as single person actions, person-person actions, and person-object actions. Since the I3D model that has been trained on this large scale dataset, features extracted though the model can capture representations of complex human activity, including human-object interaction, thus making the model a fitted choice for composite feature extraction for our analysis. We extract features on our data from the I3D (55) model trained on Kinetics 400 dataset (56). We then train MS-TCN++ on these extracted features with an objective to minimize the weighted combination of (i) cross entropy loss between the predicted segment classes and ground truth, and (ii), a penalty for over segmentation. We fix the multiplier to the penalty term, which is a hyperparameter at 0.15 as mentioned in (54). We train the network for 300 epochs with an ADAM optimizer (57), learning rate of 10^{-3} and step-wise reduction in learning rate by a factor of 0.8 for every 30 epochs.

3.5. Weighted Average Ensemble

Finally, we fuse the outputs of the data driven models with the expert knowledge constrained models. The HMM algorithm uses kinematic features and design priors that are constrained by the expert knowledge. The predictions from the HMM encode information about the segments and task level of the movement analysis hierarchy shown in **Figure 3**. On the other hand, techniques like Transformer and MSTCN++ use kinematic and composite features to perform better in the frame level but fail to encode information about the segments and the tasks. To combine the strengths of each approach, we fuse the analysis outcomes of the different algorithms. We employ a standard weighing scheme to fuse these models i.e., $\hat{y} = \sum_i W_i P_i$ where W_i is the weight assigned to each model's prediction P_i. To determine the weights W_i, we have conducted a grid search for each of them in the range of $[0.1, 0.2, 0.3, \ldots, 1.0]$. Thus, for a three model ensemble, we run the evaluation $10 \times 10 \times 10 = 1,000$ times for each combination of weights. We experimented with four ensemble schemes to find the combination that provides the most robust results: (i) HMM and Transformer, (ii) HMM and MSTCN++, (iii) Transformer and MSTCN++, and (iv)HMM, Transformer, and MSTCN++.

3.6. Block Based Task Completion Assessment

To perform task completion assessment we need to organize the per-frame segment labels into segment blocks. We again leverage therapist expertise for grouping segment samples into segment blocks (as embedded in the SARAH system design), to automatically extract segment blocks and assess the order, duration, and continuity of the blocks. To calculate segment blocks, we first denoise the ensemble predictions. The per-frame segment predictions resulting from the ensemble model have two types of noise: false transitions and missing transitions.
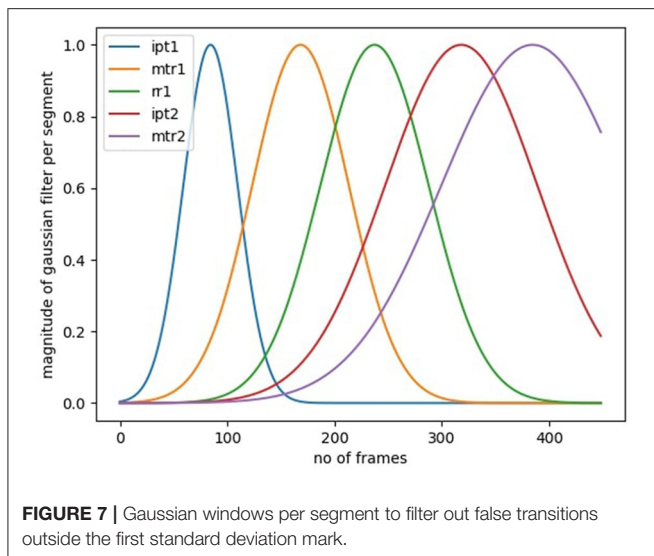
3.6.1. Denoising of Ensemble Predictions

In the SARAH system, the performance of all tasks is realized through a sequence of a small number of segments. These segments are performed with a defined set of objects, and within a defined activity space. Further, they are executed at a speed that is constrained by human biomechanics and potentially constrained by stroke induced impairment. We can filter some of the false predictions through a Gaussian filter for each transition model that implements these priors. In **Figure 7**, we illustrate the per-segment filters for a 3×3 transition matrix model. The filters can be mathematically expressed as $G_s(\mu_s, \sigma_s)$, where μ_s and σ_s denotes to the mean duration and standard deviation of each segment, s calculated from the training data. For a test data of length, N, we calculate the posterior distribution and use a 75% confidence cutoff to get the upper and lower bounds for each segment. We only accept a predicted transition, if it falls within the bounds. In addition, we pose a 20 frame window constraint on the predictions since all the task segments in the SARAH system require more than 1 s to complete. Therefore, any transitions within 20 frames is discarded. This filtering technique reduces the majority of the false predictions.

3.6.2. Rule-Based Binary Decision Tree (RBBDT)

To correct the remaining false predictions and find the missing segment transitions, we use a rule-based binary decision tree (RBBDT) to encode the process that the experts (therapists) use to segment the patient movement. As observed during the development of the SARAH system with therapists, instead of classifying each frame of movement, therapists utilize a few key events to find transition point candidates between segments. These events are primarily based on three relationships: (i) patient–object interaction, (ii) patient–activity space relationship, and (iii) object–activity space relation. People execute functional tasks differently (i.e., the trajectory of raising a glass to our mouths to drink is slightly different for each person). These differences are much more pronounced for stroke survivors as they have different types of impairment. Therefore, therapists organize the patient activity space into a few generalizable regions that are robust to variation. Standardized stroke rehabilitation assessment tests (i.e., WMFT, ARAT) rely on such generalizable regions and the regions have accordingly been adopted in the design of the SARAH system (see section 1).

To model relationships between the object(s), patient, and the generalizable activity regions, we recreate the activity regions as eight bounding boxes on each frame of the video. These

FIGURE 7 | Gaussian windows per segment to filter out false transitions outside the first standard deviation mark.

TABLE 2 | List of RBBDT features calculated using patient keypoints, object co-ordinates and center of eight bounding boxes.

Index	Feature notation	Definition
1	d_{Fl}	Distance between upper torso, F, and the limb
2	d_{Gl}	Distance between lower torso, G, and the limb
3	d_{El}	Distance between head, E, and the limb
4	d_{hl}	Distance between limb and hand bounding box, h
5	$d_{Al}, d_{Bl}, d_{Cl}, d_{Dl}$	Distance between limb and bounding box A, B, C, and D
6	d_{Fo}	Distance between upper torso, F, and the object
7	d_{Go}	Distance between lower torso, G, and the object
8	d_{Eo}	Distance between head, E, and the object
9	d_{ho}	Distance between object and hand bounding box, h
10	$d_{Ao}, d_{Bo}, d_{Co}, d_{Do}$	Distance between object and bounding box A, B, C, and D
11	d_{lo}	Distance between limb and object
12	d_{oo}	Distance between two objects
13	$\frac{d}{dt}()$	First derivative of the above features

bounding boxes are illustrated in **Figure 2**. Our approach is similar to region based object detection (39) techniques, where the bounding boxes are created based on the relationships between different objects and the space. Our bounding boxes, divide the body into the regions of head, upper and lower torso, since movements of the impaired arm toward the head (i.e., for feeding) have different functionality than movement of the impaired limb toward the torso (i.e., dressing). Moving the end point of the impaired upper limb toward the upper and lower torso presents different challenges for different patients. The tabletop activity space is divided into an ipsilateral and contralateral area since the engagement of these two spaces requires different coordination of joints and muscles. The ipsilateral and contralateral areas are further divided into proximal and distal areas, since different extension patterns of the arm are needed to engage the distal space (see **Figure 2**). We use these eight bounding boxes to calculate the space-patient, space-object, and patient-object relationships that appear in **Table 2**.

We use previously calculated object and end point position and velocity to mark the object and limb end point stopping time. We can then calculate multiple distance features like limb-bounding box distance, object-bounding box distance, and limb-object distance. These simple distance features do not exhibit a simple statistical pattern (e.g., Gaussian) across different patients for each type of segment and are therefore not easily amenable to training the HMM segmentation algorithm. However, under the RBBDT framing they did indeed become usable for detecting segments blocks because they encode the features used by experts to define segment blocks as shown in **Table 2**.

In our previous work (32), we used a standard classification and regression tree (CART) for modeling therapist decision processes in movement assessment. Data driven tuning of the parameters of these decision trees is sensitive to noisy data. Since the dataset for this experiment is small and noisy, we developed an approach for manually tuning the thresholds and split branches, and selecting the order of features of the RBBDT. Our manually tuning is supported by training data.

The goal of the tuning is to place the features in descending order of observability and error for each type of transition prediction with the most observable and accurate feature coming first in the decision sequence. The exercises of the SARAH system have different combinations of segments and variable object locations. Therefore, the RBBDT based prediction of each segment transition, the transition between two segment blocks, requires a unique order of features. In the appendix section, we show the order of features used for each transition of the different SARAH tasks based on the mathematical notation from **Table 2**. If a candidate sample of a feature stream meets the threshold condition, the confidence value of that candidate as a transition point increases.

Let's consider the previous example of exercise three and four where the patient transports the object close to their mouth to simulate drinking action. To get the candidates for the IPT-MTR1 transition, the highest observable feature is the location of the object and limb as a phase change occurs at the beginning of MTR1. The next most observable feature is the directionality of the object's movement in relation to the bounding boxes. The process for calculating the confidence value of the candidates for the above IPT-MTR1 transition is illustrated in **Table 3**. In the shown example, RBBDT generates a possible candidate when two or more conditions are true for the sample. Therefore, we set the confidence threshold to 0.5. It is important to note, that the RBBDT does not aim for accuracy at the per-frame classification level as this accuracy is secured through the ensemble model presented above. The RBBDT aims to only find transition points organizing the data stream into segments blocks that are feasible (could have been performed by a stroke patient), while minimizing missed transitions and false transitions between blocks.

TABLE 3 | Confidence calculations based on binary decision from the rules.

Object phase	Limb phase	Bounding box directionality	Confidence, C
0	0	0	0
0	0	1	0.25
0	1	0	0.25
1	0	0	0.25
1	0	1	0.5
1	1	0	0.5
0	1	1	0.5
1	1	1	0.75

TABLE 4 | Per frame segmentation results using the Hidden Markov Model.

Split index	Frame wise accuracy	Precision	Recall
1	75.52	77.94	76.04
2	78.95	79.21	79.08
3	80.75	80.17	80.96
4	74.08	74.69	74.72
5	79.82	81.01	80.49
Mean	77.82	78.60	78.26
STD	2.88	2.47	2.76

3.6.3. Task Completion Assessment

For a given task, the segment blocks will inform us about the continuity and order of the segments. In addition, we can calculate the duration of the completed task. Once we have calculated the segment blocks and duration for each task, we can compare that calculation to the expected sequence of segments and duration for each task for a coarse assessment of task completion. The algorithm checks for three conditions:

- Is the task completed with all the necessary segment blocks?
- Is the order of the segment blocks correct?
- Is the task performed within allowable time?

If a task satisfies all three conditions, the algorithm gives a score of 3. If the first two are satisfied then it generates a score of 2. If either of the first two are not satisfied then it generates a score of 1. Finally, a score of 0 is assigned if the task is not attempted. Although these scores are similar to the scores assigned by therapists they do not take into account issues of movement quality. Therapists may assign a score of 2 to a task that is completed within the allowable time because of issues of movement quality. Since our methodology connects different types of movement features to assessment of segment and task execution, in the future we plan to use actual therapists ratings of patient videos to further train our algorithms so they can automatically assess task completion and movement quality.

4. EXPERIMENTAL SETTINGS

Our initial dataset consists of 610 captured videos. The experimental dataset includes 404 videos after exclusion of videos where severe limb impairment caused multiple object dropping or multiple segment occurrence leading to an incomplete exercise. Including these videos would severely skew the training sets at this early stage. However, some of these videos were included in the test sets for task completion and in the future these videos can be included in the training sets too. In our dataset each patient performed each task multiple times in a single session. Therefore, we design random split experiments based on three factors: patient ID, session ID and task number.

For any task number, the same patient ID with the same session ID is included in either of the training or test set but not in both. We choose different random seeds and create 5 random experiments for fairness based on this selection method. The average training and test size of the 5 splits are 370 and 34, respectively. We implement the deep learning experiments on two NVIDIA RTX 2080 Ti GPUs. To evaluate the segmentation performance of the experimented models, we calculate frame-wise accuracy, precision, and recall. The calculation formula for the matrices are given in the appendix section. In the **Supplementary Material**, we also report two additional metrics for all our models- IOU scores and F1 scores at different overlap thresholds. For each of the split experiments, we evaluate the matrices independently and calculate mean and standard deviation of the results for all the algorithms.
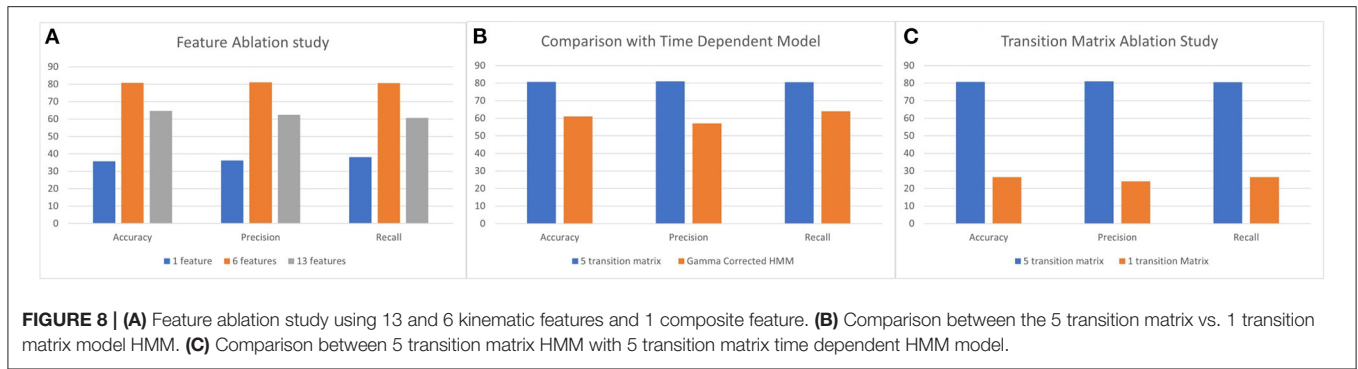
5. RESULTS

5.1. HMM Segmentation Results

Table 4 shows the segmentation results using the proposed five-transition-matrix HMM. The average per-frame accuracy is $77.82 \pm 2.88\%$ meaning around 78% of the frames are labeled correctly. The precision and recall values are $78.60 \pm 2.47\%$ and $78.26 \pm 2.76\%$, respectively. Additionally, we have performed the following ablation studies to demonstrate the efficacy of the proposed five-transition-matrix HMM.

5.1.1. Number of Features

We experimented with the proposed HMM using various numbers of features. The feature selection is an important factor as the combination of correct features captures the unique patterns of different transitions. To understand the effect of feature selection, we perform three experiments: (i) 13 kinematic features; (ii) six kinematic features including four derivatives; and (iii) one composite feature. To get the composite feature from 13 kinematic features, we experimented with different dimensionality reduction techniques like PCA, NMF, LDA, and RP (58–61). Based on the performance, we choose PCA to reduce the dimension of 13 features into one and produce the composite feature. In **Figure 8A**, we illustrate the comparative results for the above three experiments. As evident from the illustration, the best performing result was achieved when a combination of

FIGURE 8 | (A) Feature ablation study using 13 and 6 kinematic features and 1 composite feature. **(B)** Comparison between the 5 transition matrix vs. 1 transition matrix model HMM. **(C)** Comparison between 5 transition matrix HMM with 5 transition matrix time dependent HMM model.

TABLE 5 | Per frame segmentation Results using Transformer with three different types of activity space normalization for 5 splits.

Normalization	Metrics	1	2	3	4	5	Mean	STD
Homography-Norm	ACC	70.07	74.29	75.85	65.34	72.37	71.59	3.67
	Precision	65.03	61.68	68.69	58.37	63.20	63.40	3.43
	Recall	61.84	61.45	68.08	55.55	60.54	61.49	3.99
Global-Norm	ACC	69.60	78.55	75.29	75.36	76.20	75.00	3.29
	Precision	68.62	69.43	72.58	72.40	70.83	**70.77**	1.76
	Recall	65.37	68.88	72.77	69.51	68.73	**69.05**	2.63
Mat-Norm	ACC	73.43	78.82	78.55	73.67	77.24	**76.34**	2.34
	Precision	70.20	72.35	73.49	66.72	67.74	70.10	2.59
	Recall	68.71	70.55	74.19	61.94	66.43	68.36	4.10

The bold number represent the best performing values.

the raw kinematics and the derivatives were used as input to the HMM.

5.1.2. Transition Matrix

The proposed HMM model uses five transition matrices. We also experimented with one transition matrix model to understand the capability of the HMM to learn different variations of segment transitions across all types of tasks. A 6×6 transition matrix is necessary to represent all the segment transitions across all performed tasks in our data set. In **Figure 8C**, we show the comparison between the one transition matrix HMM and the proposed five transition matrix HMM (with each transition matrix calibrated to particular types of tasks).

5.1.3. Time Dependency

We also experimented with time dependent models and wanted to exploit the duration of segments to correct the predictions from the HMM. To model time dependency in the HMM predictions, we use the gamma correction (62) on the predictions of the HMM. We calculate the gamma probabilities using:

$$\gamma(x, \alpha, \beta) = x^{\alpha-1} \frac{\beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)} \qquad (4)$$

where, α and β is calculated per segment using mean (μ) and standard deviation (σ) of the segments in the training set. Since the dataset has a very high standard deviation, the gamma function [$\Gamma(\alpha)$] results in a large value and thus

Equation (4) generates probabilities < 0.01. Therefore, the effect on the HMM predictions are negligible. In **Figure 8B**, we illustrate the comparison with the proposed five transition matrix HMM model.

5.2. Transformer Segmentation Results

Like the HMM, the transformer was also tested on five random splits. In **Table 5**, we demonstrate the segmentation results for the five splits. As described in section 3.1.2, we experimented with two types of normalization technique to remove activity space variance. We present the result for both normalization techniques in **Table 5**. We conducted two additional studies to understand the dynamics of our transformer pipeline by ablating on (i) the number of features used as input; and (ii) the amount of labeled training data needed. The former ablation study attempts to identify the features that are beneficial, while the latter answers the reduction in time and cost needed to label the data by hand.

5.2.1. Number of Features

We experimented with different number of keypoints while always using the object features as input to the transformer based pipeline. We decreased the number of keypoints from 8, 4, 2, and then to 1. We selected the keypoints for each reduction based on our prior work (63) that defined the relative importance of different key points in characterizing upper limb impaired movement. The best performing result was achieved using just the wrist keypoints and object locations. This conforms to the observations made in (63). The results are given in **Table 6**.

TABLE 6 | Ablation results on transformers using different number of keypoints.

No. of Keypoints	ACC		Precision		Recall	
	Mean	STD	Mean	STD	Mean	STD
1	**76.34**	2.34	**70.10**	2.59	**68.36**	4.10
2	69.65	1.66	55.34	3.67	53.87	4.02
4	71.30	3.20	67.51	4.23	66.14	2.89
8	67.80	2.53	60.19	7.37	56.47	6.15

Using only wrist keypoint outperforms the rest.
The bold number represent the best performing values.

TABLE 7 | Transformer performance with varying sizes of training data.

No. of Samples	ACC		Precision		Recall	
	Mean	STD	Mean	STD	Mean	STD
100	65.00	1.85	57.23	3.12	55.12	3.05
200	71.97	3.00	64.63	1.79	62.85	2.17
300	75.54	2.57	69.49	2.54	68.11	3.88
404(all data)	**76.34**	2.34	**70.10**	2.59	**68.36**	4.10

As expected, more data prevents overfitting thus explaining the better performance.
The bold number represent the best performing values.

TABLE 8 | Per frame segmentation result using MSTCN++ for 5 splits.

Split	1	2	3	4	5	Mean	STD
ACC	79.75	85.96	83.26	79.96	80.60	81.91	2.38
Precision	76.84	85.29	82.05	74.30	75.08	78.71	4.25
Recall	77.64	84.04	82.24	74.68	75.51	78.82	3.70

This method outperforms HMM and transformers for the task of per-frame segmentation.

5.2.2. Amount of Training Data

We experimented with reducing the amount of training data to the transformers to better study the generalization properties of our pipeline in low data scenarios. This is a practical setting, as the amount of time required to label the data is very high. However, as transformers contain large number of parameters, it is important to study the over fitting trends as well. We report average accuracies on the same test data for fair comparison while changing the size of input training data. As can be seen from the **Table 7**, if the amount of data available is too small, unsurprisingly the performance is very poor due to overfitting; however, as size of available data increase, the performance improves significantly plateauing after sometime. It is encouraging that even with 25% reduction in training data(404 to 300) the drop in performance is only 2% thus showing the robustness of the model.

5.3. MSTCN++ Segmentation Results

The results for the five split experiment using the MS-TCN++ is shown in **Table 8**. This is the best standalone model in the pipeline as evident from the values. All coordinates were normalized with respect to the activity mat area as described in section 3.1.3 refereed to as Mat-Norm. As can be seen from **Table 8**, MS-TCN++ achieves the best performance in terms of per-frame classification error. However, as will be shown

in section 5.7, even though it achieves better classification error, it performs poorly in standalone estimation of the completion of the task. As MS-TCN++ is a completely data driven model it fails to encode any design priors or expert knowledge (i.e., expected order of segments). Thus, the few errors that MS-TCN++ makes are highly harmful. As an example, while every task should start with an "IPT" segment, the MS-TCN++ based model labels the first few frames of some tasks as "MTR" thus affecting the automated task completion performance. Thus, for higher level automated decisions (assessing completion of segments and tasks by patients) it is imperative to fuse the MS-TCN++ with other models that encode prior knowledge such as our implementations of the HMM and transformer based models.

5.4. Ensemble Model Segmentation Results

We experiment with three combinations of model ensembles. In **Table 9**, we illustrate the results of three ensemble models. The best performance is achieved when all three models are combined and that finding is consistent through all the splits. This is rather unsurprising because, the models are all complementary in nature with MS-TCN++ being completely data driven, while our transformer and HMM based models encode certain amounts of prior knowledge.

TABLE 9 | Results using different ensembles.

	Ensemble 1		Ensemble 2		Ensemble 3	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**
ACC	81.01	1.99	83.76	2.77	85.08	2.14
Precision	77.28	4.38	81.40	3.92	84.34	2.61
Recall	77.88	4.46	81.57	3.66	84.6	2.68

(i) Ensemble1: Transformer and HMM, (ii) Ensemble2: HMM and MS-TCN++, and (iii) Ensemble3: Transformer, HMM, and MS-TCN++. We observe that ensemble3 outperforms the rest.

TABLE 10 | Per-sample inference time for each method.

	Transformer	**HMM**	**MS-TCN++**
Inference time (1e-3s)	10.1	1.85	8.3

5.5. Analysis of Running Time

In **Table 10**, we provide time taken for inference by each of our algorithm for each video. As can be seen, Transformer takes the maximum amount of time due to its large amount of parameters compared to HMM. Results are reported on a Intel Xeon Cpu and after averaging the run-times for three different runs.

5.6. Qualitative Analysis

In **Figure 9**, we demonstrate the qualitative performance of the ensemble model. **Figure 9A** shows the predicted vs ground truth comparison for exercise 1 which is one of the easiest task among the 15 tasks in the SARAH system. As evident from the illustration, the proposed ensemble model is almost 97% accurate in this example. However, in **Figure 9B**, the model fails to predict one of the transportation segment. In this case the patient drops the object at the end of "MTR2" segment resulting in artifacts in the features. However, after incorporating the block-based segmentation using the design constraint denoising and RBBDT, we can recover more than 50% of the "MTR2" segment as indicated by striped blue region in **Figure 9B**.

5.7. Task Performance Assessment Scores

In **Table 11**, we demonstrate the results of the block based segmentation. As evident from the table, with the incorporation of the proposed denoising and RBBDT, around 99% of the segment blocks can be labeled correctly. We also calculate task completion accuracy based on the continuity, and order of the segment blocks and duration of the tasks. A task is completed if it has an appropriate order of segments. As shown in **Figure 10A**, with the highest as 100% for split 4 and lowest as 87.5% for split 3, the average task completion accuracy is 92%. This means 92% of the tasks in the test case that were tagged as completed by the therapists are also labeled as "completed" by our algorithm. These are the tasks that have a score of either a 3 or 2. The algorithm generates a score of 2 only when a patient takes longer time to complete the task correctly. In **Figure 10B**, we show the percentage of correctly predicted 2 s using the segment blocks for three experiments. For these experiments we varied the range of standard deviation (std). As seen from the figure, our algorithm
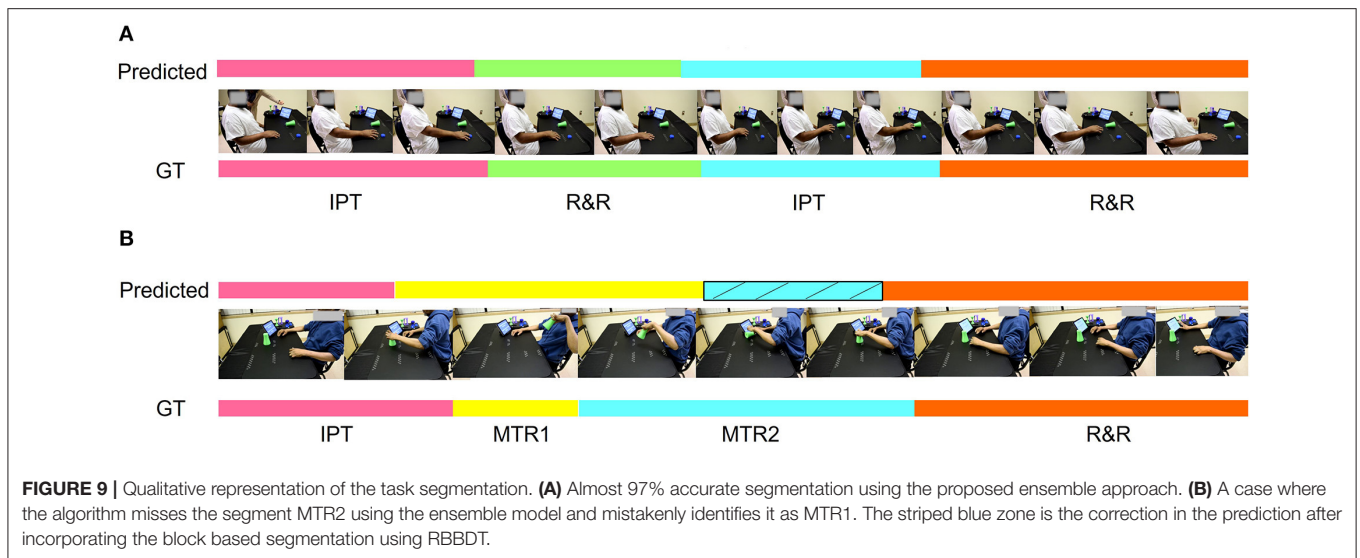
predicted maximum 22% of the 2 s among the videos that are rated 2 s by the therapist. All these tasks are labeled as completed both by the algorithm and the therapist.

6. DISCUSSION

In this paper, we propose a hierarchical model for automatically segmenting stroke survivor's movements and generating task performance assessment scores during rehabilitation training. The hierarchical model fuses expert knowledge-biased approaches with data-driven techniques. The expert knowledge is more observable in the higher layers of the hierarchy (task and segment) and therefore more accessible to algorithms incorporating high level constraints relating to activity structure (i.e., type and order of segments per task). The lower layers need to be addressed primarily through data driven techniques. By developing a sequence for combining complimentary algorithms that effectively encode the information from the different layers, we produce robust segmentation and task assessment results driven by noisy, variable, and limited data.

The MSTCN++ (54) is a data driven technique that relies primarily on the RGB data layer of our hierarchy and on composite kinetic features (the features extracted through the pre-processor) that are trained through generic videos of human activity. Therefore, the composite features are not fully observable and are not directly related to upper extremity functional tasks. Furthermore, the algorithm cannot incorporate the higher layer constraints resulting from the expert driven design of the system (segment vocabulary, composite features related to each segment, order of segments per task). As the algorithm is most sensitive to the lower layers of the hierarchy, it performs the best in terms of per-frame segment classification, but because it does not incorporate higher layer constraints, is prone to errors in the order of segments. For example it can classify the beginning frames of a task as belonging to an MTR segment when all the SARAH tasks start with IPT. Because of the ordering errors the MSTCN cannot be used in standalone mode for assessing segment completion and task completion, since task completion is based on assessment of segment completion and segment order.

The Transformer utilizes raw kinematic features of the torso, upper limb, and object. We experimented with 8, 4, 2, and 1 keypoints from open pose combined with object movement features. We selected the keypoints representing joints that have been show, in our previous work, to have prominence in characterizing upper limb impaired movement (32, 34). It can

FIGURE 9 | Qualitative representation of the task segmentation. **(A)** Almost 97% accurate segmentation using the proposed ensemble approach. **(B)** A case where the algorithm misses the segment MTR2 using the ensemble model and mistakenly identifies it as MTR1. The striped blue zone is the correction in the prediction after incorporating the block based segmentation using RBBDT.

TABLE 11 | Segmentation block results.

Split	Segment wise accuracy	Precision	Recall
1	99.33	99	98.5
2	98.48	98.18	97.04
3	98.66	98	97
4	100	100	100
5	98.66	98.5	97.5
Mean	99.03	98.73	98.01
STD	0.56	0.71	1.13

thus be said that some form of prior expert knowledge is used to constrain the Transformer. The best performing Transformer was with only the wrist key point and object locations. The object location is an important input feature as a lot of the segment transitions are dependent on the object achieving specific translations in the activity space (i.e., movement from bounding box C to bounding box D and back). In prior work, we show that the object-limb end point (wrist or hand) interaction data can be sufficient for the coarse analysis of task completion in upper extremity rehabilitation (32). We also further proposed that in these scenarios, the limb could be considered as a dynamic system that can be sufficiently characterized by the behavior of the end point (32). Therefore, through the Transformer we codify the long term relationship between the object and the patient limb for better predictions of the segment labels. However, because of the limited and noisy data set, and the limited data points being used for the Transformer, this model has the worst per-frame classification performance of the three segmentation algorithms.

The hidden Markov model (HMM) utilizes kinematic features that are prominent in characterizing upper limb functional movement (5) and also incorporates segment and task layer information through the customized transition matrices. We experimented with multiple training schemes for the HMM. The

ablation study of the HMM provides three key insights. First, the HMM is sensitive to input features. The best performance achieved used a six raw kinematic feature HMM model combined with the first derivatives of four kinematic features. The derivatives have more oscillation compared to the raw features as shown in **Figure 11** and follow the Gaussian distribution. Second, the proposed HMM requires five transition matrix as a prior. The transition probabilities are selected based on the expert knowledge. The size of the transition matrix depends on the state machine of the exercise. However, we experimented with one transition matrix model to represent all the exercises and as shown in **Figure 8B**. The performance is poor compared to a 5 model HMM since the expert knowledge of segment-task relation is lost. Lastly, we compared the proposed HMM with a time dependent HMM model but the variance in duration of performance among stroke survivors with different levels of impairment reduces the performance of the time dependent HMM. The HMM has limited segment ordering issues as those are constrained by the transition matrices. However, the reduced sensitivity of the HMM to nuances in the data layer (the HMM model would perform better if the differences in the data between states was significant) combined with the variance in the data cause a higher number of frame level misclassifications. We also varied the batch size to understand the sensitivity of the HMM. In **Table 12**, the comparative results for different batch sizes are shown. It is evident that with the increase in data size the performance of the HMM increases. Since the HMM tunes its transition matrix and distribution parameters based on the training data, the more data the better the tuning, and we can thus expect to see continuous improvement as we add data in the future.

For all the splits, the ensemble model with three algorithms performs the best in the frame level classification. This proves that a fusion of algorithms that have different sensitivities to different layers of the hierarchy outperforms individual models. Without the HMM, the ensemble model predictions have a higher misclassification rate and incorrect sequence orders. For
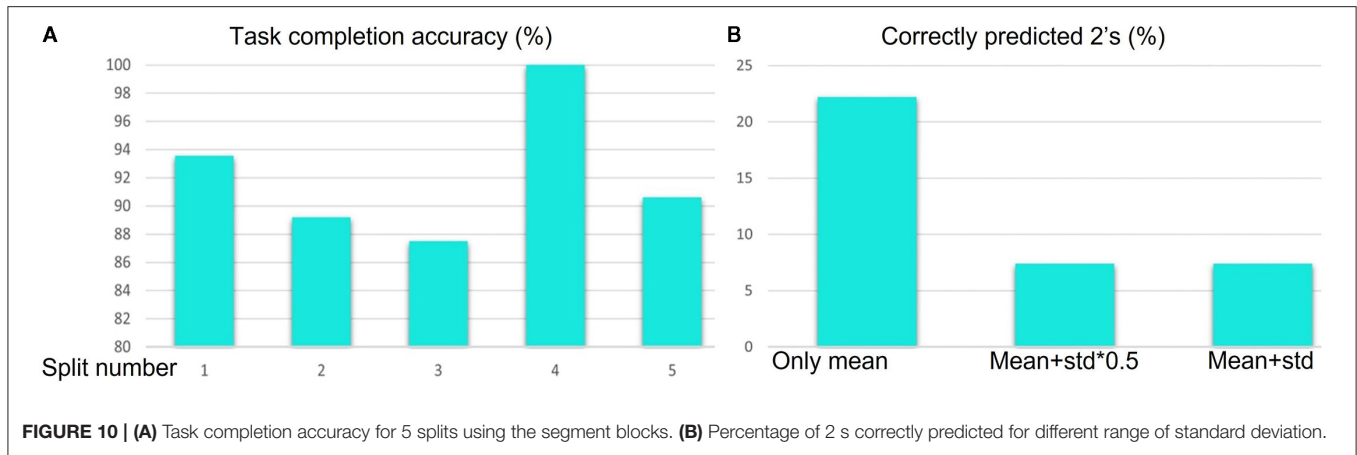
FIGURE 10 | (A) Task completion accuracy for 5 splits using the segment blocks. **(B)** Percentage of 2 s correctly predicted for different range of standard deviation.
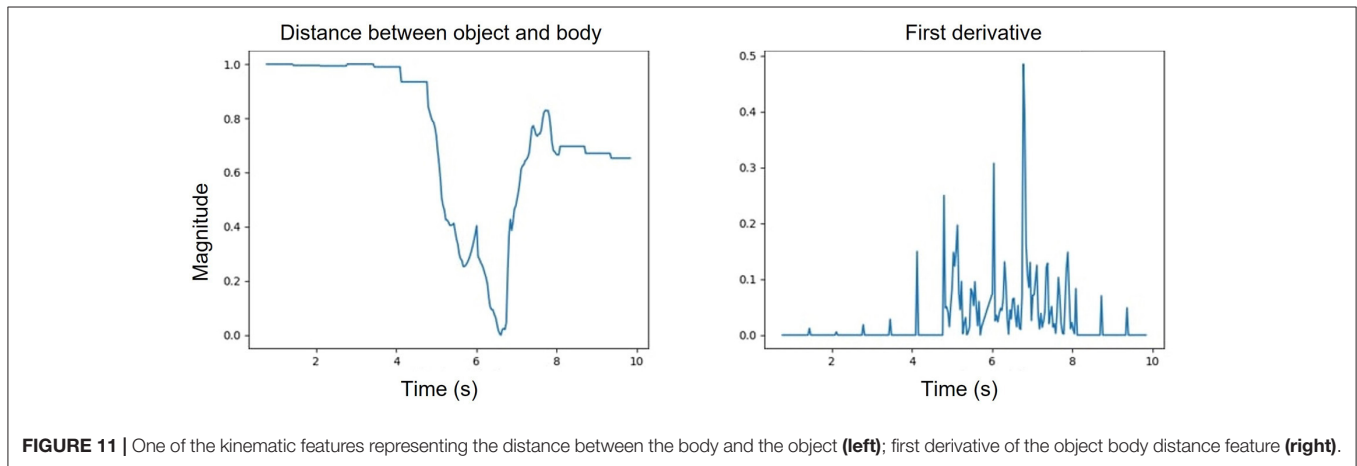
FIGURE 11 | One of the kinematic features representing the distance between the body and the object **(left)**; first derivative of the object body distance feature **(right)**.

TABLE 12 | Results using different batch sizes for HMM input: (i) Experiment 1: batch size of 1, (ii) Experiment 2: batch size of 9 and (iii) Experiment 3: all data.

	Experiment 1		Experiment 2		Experiment 3	
	Mean	**STD**	**Mean**	**STD**	**Mean**	**STD**
ACC	55	10.83	73.15	6.8	**77.82**	2.88
Precision	60.5	10.15	75.02	6.72	**78.60**	2.47
Recall	57	10.21	75.20	5.83	**78.26**	2.76

The bold number represent the best performing values.

example, without the HMM a prediction can start with MTR or end with IPT. But based on our design prior, we know that all exercises start with IPT and end with *R&R*. Since this prior is incorporated into the HMM, the predicted order of segment blocks are correct and the per-frame segment label prediction accuracy increases. As is evident from **Table 9**, without the MSCTN the ensemble model has significant deterioration in the frame level classification.

To automatically assess the completion of a task, we need to assess the completion and order of segment blocks. This cannot be done solely through the ensemble model as the ensemble model has many false positive blocks and missed transitions. The connection of the segmentation layer to the task assessment layer requires an algorithm that is heavily biased by expert knowledge toward behaviors characterizing transition between

segments, as well as the effect of stroke related impairment on these behaviors. The RBBDT algorithm successfully incorporates key unimpaired and impaired movement priors for segment block transition points in one integrated decision sequence. For example, a phase transition is expected for both object and limb end point for a simulated drinking motion (task three) when the object gets close to the head. However, not all stroke survivors are able to lift the object all the way to their mouth, so the phase transition can sometimes happen away from the head. Thus, the phase transition becomes the first decision node with the place of the transition in the activity space the second node. The RBBDT is highly sensitive to transition point features but not sensitive to per-frame data patterns across the whole task. Thus, the RBBDT produces low quality performance in per-frame classification but higher quality performance in identifying

transition candidates. When integrated with the ensemble model, the correct identification of segment blocks rises above 90%.

To perform the task completion assessment, we use a very simple algorithm relying only on comparison of the type and order of the segment blocks given by the integrated ensemble and RBBT predictions, with the expected type and order of segments in tasks established by the expert therapists who designed the SARAH tasks. By adding this codification of prior task knowledge to our analysis, we are able to correctly classify completed and uncompleted tasks over 90% of the time. We are also able to classify completed tasks that took longer than the performance time allowed by the therapist for impaired performance. These results further enhance the validity of the use of a hierarchical model of automated analysis combining algorithms with different sensitivities to the different layers of the hierarchy. This includes layers primarily incorporating observable expert knowledge (higher layers), layers primarily incorporating computer analyzed data (lower layers) and layers integrating data and expert knowledge (middle layers; see **Figure 3**).

The use of ∼400 videos for training/testing is enough to overcome the issues of overfitting that we demonstrate when using 200 or less videos. As we progressed to 300 and then 400 videos, we show steady improvement for the performance of all the algorithms we were using. It is thus realistic to assume that as we collect more data, the performance of the hierarchical model will continue to improve at all layers.

7. CONCLUSION AND FUTURE WORK

Low cost and low intrusion long-term automated rehabilitation at the home is expected to produce low quality and high variability data. A hierarchical model fusing expert knowledge-biased approaches with data-driven techniques can produce robust results in segmentation and task completion assessment in rehabilitation even when utilizing low quality and high variability data. The hierarchical model produces over 90% performance in assessment of segment completion and task completion. Even with this limited information, long term, semi-automated rehabilitation at the home using our SARAH system is feasible. The system can monitor whether the patient has completed the assigned exercises for the day and whether exercises were completed correctly and with ease. The system can use this information to provide coarse feedback to the patient to reward accurate performance or remind the patient of the correct structure of the task if errors are detected. The system can also make simple decisions including when to encourage the patient to repeat an exercise (if they are improving but still have more room for improvement), or move to another simpler or harder exercise (if they have repeated the exercise enough times or are not making progress on the current exercise).

To realize this type of interactive training at the home the system needs to have the capacity to assess each performed task within 10 s from its completion. During these 10 s patients rest and self-asses their performance of the previous task. Then the system provides the automated assessment and the prompt for the performance of the next task. The results of our run-time comparison (section 5.5) show that with some further

optimization of our algorithms we can potentially automatically segment video recorded tasks and assess task and segments completion within 10 s from completion of each performed task.

Our team is currently working to develop a hybrid approach where we will combine the vision-based and IMU based techniques to fully capture movement patterns of stroke survivors. We will use four IMU sensors that will record accelerometer data of both hands, one index finger, and the pelvis. The ring sensor on the finger will capture the hand openness that will differentiate between different grips and release patterns. Our team is currently developing techniques to reconstruct 3D skeleton from accelerometer data using the other three sensors.

Our computational hierarchy, since it is based on computational data, realizes the hierarchical analysis in a bottom up manner (features conditioning segment frames, conditioning segment blocks, conditioning task completion). Therapists utilize a similar hierarchical analysis in a top down manner (task conditioning segments, conditioning composite features, conditioning raw features) so as to leverage their heuristics about functional task performance. Therefore, the SARAH system can send a daily summary of training results to a remote therapist. Since the computational results will be highly compatible with the therapist assessment heuristics, the therapist will be able to quickly use this summary to structure the next day's therapy and send a message to the patient guiding their training the next day. The therapist will also be able to select any analysis result and review the corresponding video thus further informing their assessment and therapy adaptation. The feasibility of low-cost semi-automated rehabilitation at the home using the SARAH system will allow the collection of many more videos used to further train our algorithms. We are currently planing to use the SARAH system in a pilot study in the homes of stroke survivors in the Spring of 2022.

Our team also developed an intuitive assessment interface allowing expert therapists to rate videos of therapy tasks in a top down hierarchical manner, complimenting their assessment approach: rate the overall task, then the segments, then the composite features per segment, and then return to a final assessment of the task) (26). As the videos collected through the SARAH system are rated by expert therapists, we can use this detailed rating (which is highly compatible with our computational approach) to further inform our segmentation and task assessment algorithms and evolve these algorithms to also automatically analyze movement quality and the relation of movement quality to functionality. The automatic rating of the videos will also address the limited availability of rated videos by experts since therapists have overloaded work schedules and do not have time to rate videos. Automated assessment will also allow therapist to focus more time on providing treatment. To receive the benefits of automated assessment that is compatible with their practice, therapists may be willing to quickly review automatic ratings and comment on potential issues. These annotations could then be used to improve algorithm performance. Once a significant number of videos have been rated, this data can be distributed within the rehabilitation community as a basis for quantitatively exploring the relation of movement changes to functional recovery. This information can then be used to refine and optimize therapy customization.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because as per instruction from the institutional review board (IRB), the authors are not allowed to disclose the raw data. The processed features supporting the conclusions of this article will be made available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Emory University IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fneur.2021.720650/full#supplementary-material

REFERENCES

1. Haghi M SR Thurow K. Wearable devices in medical internet of things: scientific research and commercially available devices. *Healthc Inform Res.* (2017) 23:4–15. doi: 10.4258/hir.2017.23.1.4

2. Pantelopoulos A, Bourbakis NG. A survey on wearable sensor-based systems for health monitoring and prognosis. *IEEE Trans Syst Man Cybern Part C.* (2010) 40:1–12. doi: 10.1109/TSMCC.2009.2032660

3. Kleim J, Jones T. Kleim JA, Jones TAPrinciples of experience-dependent neural plasticity: implications for rehabilitation after brain damage. *J Speech Lang Hear Res.* (2008) 51:S225–39. doi: 10.1044/1092-4388 (2008/018)

4. Lang CE, Bland MD, Bailey RR, Schaefer SY, Birkenmeier RL. Assessment of upper extremity impairment, function, and activity after stroke: foundations for clinical decision making. *J Hand Ther.* (2013) 26:104–14; quiz: 115. doi: 10.1016/j.jht.2012.06.005

5. Chen Y, Duff M, Lehrer N, Sundaram H, He J, Wolf S, et al. A computational framework for quantitative evaluation of movement during rehabilitation. *AIP Conf Proc.* (2011) 6:1371. doi: 10.1063/1.3596656

6. Duff M, Attygalle S, He J, Rikakis T. A portable, low-cost assessment device for reaching times. *Annu Int Conf IEEE Eng Med Biol Soc.* (2008) 2008:4150–3. doi: 10.1109/IEMBS.2008.4650123

7. Lehrer N, Attygalle S, Wolf S, Rikakis T. Exploring the bases for a mixed reality stroke rehabilitation system, Part I: a unified approach for representing action, quantitative evaluation, and interactive feedback. *J Neuroeng Rehabil.* (2011) 8:51. doi: 10.1186/1743-0003-8-51

8. Picha K, Howell D. A model to increase rehabilitation adherence to home exercise programmes in patients with varying levels of self-efficacy. *Musculoskel Care.* (2017) 4:16. doi: 10.1002/msc.1194

9. Donelan K, Hill CA, Hoffman C, Scoles K, Feldman PH, Levine C, et al. Challenged to care: informal caregivers in a changing health system. *Health Aff.* (2002) 21:222–31. doi: 10.1377/hlthaff.21.4.222

10. Carilion Clinic. *Carilion Clinic Home Care FY14 Annual Report.* Carilion Clinic (2015). Available online at: https://issuu.com/carilionclinic/docs/home-care-fy14_annualreport

11. Deaver B, Nelson T, Turner C. Using a mobile application to assess knee valgus in healthy and post-anterior cruciate ligament reconstruction participants. *J Sport Rehabil.* (2019) 28:532–5. doi: 10.1123/jsr.2018-0278

12. Hanley B, Tucker C, Bissas A. Differences between motion capture and video analysis systems in calculating knee angles in elite-standard race walking. *J Sports Sci.* (2018) 36:1250–5. doi: 10.1080/02640414.2017.1372928

13. Reinkensmeyer DJ, Blackstone S, Bodine C, Brabyn J, Brienza D, Caves K, et al. How a diverse research ecosystem has generated new rehabilitation technologies: review of NIDILRR's Rehabilitation Engineering Research Centers. *J Neuroeng Rehabil.* (2017) 14:109. doi: 10.1186/s12984-017-0321-3

14. Freedson PS, Melanson E, Sirard J. Calibration of the computer science and applications, Inc. accelerometer. *Med Sci Sports Exerc.* (1998) 30:777–81. doi: 10.1097/00005768-199805000-00021

15. Axelrod L, Fitzpatrick G, Burridge J, Mawson S, Smith PP, Rodden T, et al. The reality of homes fit for heroes: design challenges for rehabilitation technology at home. *J Assist Technol.* (2009) 3:35–43. doi: 10.1108/17549450200900014

16. Levin MF, Kleim J, Wolf S. What do motor recovery and compensation mean in patients following stroke? *Neurorehabil Neural Repair.* (2008) 23:313–9. doi: 10.1177/1545968308328727

17. Rabadi MH, Rabadi FM. Comparison of the action research arm test and the Fugl-Meyer assessment as measures of upper-extremity motor weakness after stroke. *Arch Phys Med Rehabil.* (2006) 87:962–6. doi: 10.1016/j.apmr.2006.02.036

18. Wolf S, Catlin P, Ellis M, Archer A, Morgan B, Piacentino A. Assessing wolf motor function test as outcome measure for research in patients after stroke. *Stroke.* (2001) 32:1635–9. doi: 10.1161/01.STR.32.7.1635

19. Chen Y, Baran M, Sundaram H, Rikakis T. A low cost, adaptive mixed reality system for home-based stroke rehabilitation. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* Boston, MA: IEEE Engineering in Medicine and Biology Society Conference. (2011). p. 1827–30.

20. Cirstea MC, Levin MF. Improvement of arm movement patterns and endpoint control depends on type of feedback during practice in stroke survivors. *Neurorehabil Neural Repair.* (2007) 21:398–411. doi: 10.1177/1545968306298414

21. Nordin A, Alt Murphy M, Danielsson A. Intra-rater and inter-rater reliability at the item level of the Action Research Arm Test for patients with stroke. *J Rehabil Med.* (2014) 6:46. doi: 10.2340/16501977-1831

22. Wolf S, Winstein C, Miller J, Taub E, Uswatte G, Morris D, et al. Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after stroke: the excite randomized

clinical trial. *JAMA*. (2006) 296:2095–104. doi: 10.1001/jama.296.17.2095

23. Reinkensmeyer DJ, Burdet E, Casadio M, Krakauer JW, Kwakkel G, Lang CE, et al. Computational neurorehabilitation: modeling plasticity and learning to predict recovery. *J NeuroEng Rehabil*. (2016) 13:42. doi: 10.1186/s12984-016-0148-3

24. Norman DA. *The Design of Everyday Things*. New York, NY: Basic Books (2002).

25. Kelliher A, Gibson A, Bottelsen E, Coe E. Designing Modular Rehabilitation Objects for Interactive Therapy in the Home. In: *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction, TEI '19*. New York, NY: Association for Computing Machinery (2019). p. 251–257. doi: 10.1145/3294109.3300983

26. Kelliher A, Zilevu S, Rikakis T, Ahmed T, Truong Y, Wolf SL. Towards standardized processes for physical therapists to quantify patient rehabilitation. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*. New York, NY: Association for Computing Machinery (2020). p. 1–13. doi: 10.1145/3313831.3376706

27. Teasell R, Mehta S, Pereira S, McIntyre A, Janzen S, Allen L, et al. Time to rethink long-term rehabilitation management of stroke patients. *Top Stroke Rehabil*. (2012) 19:457–62. doi: 10.1310/tsr1906-457

28. Woytowicz EJ, Rietschel JC, Goodman RN, Conroy SS, Sorkin JD, Whitall J, et al. Determining levels of upper extremity movement impairment by applying a cluster analysis to the Fugl-Meyer assessment of the upper extremity in chronic stroke. *Arch Phys Med Rehabil*. (2017) 98:456–62. doi: 10.1016/j.apmr.2016.06.023

29. Wolf SL, Thompson PA, Morris DM, Rose DK, Winstein CJ, Taub E, et al. The EXCITE trial: attributes of the Wolf Motor Function Test in patients with subacute stroke. *Neurorehabil Neural Repair*. (2005) 19:194–205. doi: 10.1177/1545968305276663

30. Rikakis T, Kelliher A, Choi J, Huang JB, Kitani K, Zilevu S, et al. Semi-automated home-based therapy for the upper extremity of stroke survivors. In: *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference, PETRA '18*. New York, NY: Association for Computing Machinery (2018). p. 249–56. doi: 10.1145/3197768.3197777

31. Clark J, Kelliher A. Understanding the needs and values of rehabilitation therapists in designing and implementing telehealth solutions. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21*. New York, NY, USA: Association for Computing Machinery (2021). doi: 10.1145/3411763.3451704

32. Venkataraman V, Turaga P, Lehrer N, Baran M, Rikakis T, Wolf S. Decision support for stroke rehabilitation therapy via describable attribute-based decision trees. *Annu Int Conf IEEE Eng Med Biol Soc*. (2014) 2014:3154–9. doi: 10.1109/EMBC.2014.6944292

33. Simon HA. *The Sciences of the Artificial*. 2nd Edn. Cambridge, MA: MIT Press (1981).

34. Venkataraman V, Turaga PK, Baran M, Lehrer N, Du T, Cheng L, et al. Component-level tuning of kinematic features from composite therapist impressions of movement quality. *IEEE J Biomed Health Inform*. (2016) 20:143–52. doi: 10.1109/JBHI.2014.2375206

35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. (2017). Available online at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

36. Farha YA, Gall J. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In: *CVPR, Computer Vision Foundation/IEEE*. (2019). p. 3575–84. doi: 10.1109/CVPR.2019.00369

37. Kelliher A, Choi J, Huang JB, Rikakis T, Kitani K. HOMER: an interactive system for home based stroke rehabilitation. In: *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '17*. New York, NY: Association for Computing Machinery (2017). p. 379–380. doi: 10.1145/3132525.3134807

38. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans Pattern Anal Mach Intell*. (2019) 43:172–86. doi: 10.1109/TPAMI.2019.2929257

39. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, editors. *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc. (2015). Available online at: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

40. Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, et al. *Microsoft COCO: Common Objects in Context*. (2014). Available online at: http://arxiv.org/abs/1405.0312

41. Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. *Detectron2*. (2019). Available online at: https://github.com/facebookresearch/detectron2

42. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: *NeurIPS*. Vancouver, BC (2019). p. 8024–35.

43. Bewley A, Ge Z, Ott L, Ramos F, Upcroft B. Simple online and realtime tracking. In: *2016 IEEE International Conference on Image Processing (ICIP)*. (2016). p. 3464–8. doi: 10.1109/ICIP.2016.7533003

44. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. (1989) 77:257–86. doi: 10.1109/5.18626

45. Osgouei RH, Soulsby D, Bello F. An objective evaluation method for rehabilitation exergames. In: *2018 IEEE Games, Entertainment, Media Conference (GEM)*. Galway (2018). p. 28–34.

46. Nguyen TN, Huynh HH, Meunier J. Skeleton-based abnormal gait detection. *Sensors*. (2016) 16:1792. doi: 10.3390/s16111792

47. Deters JK, Rybarczyk Y. Hidden Markov Model approach for the assessment of tele-rehabilitation exercises. *Int J Artif Intell*. (2018) 16:1–19.

48. Schreiber J. Pomegranate: fast and flexible probabilistic modeling in python. *J Mach Learn Res*. (2018) 18:1–6. doi: 10.1002/9781119557500.ch1

49. Nocedal J, Wright SJ. Sequential quadratic programming. *Numer Optim*. (2006) 529–62.

50. Sutskever I, Vinyals O, Le QV. Sequence to Sequence Learning with Neural Networks. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc. (2014). Available online at: https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

51. Arundo. tsaug. *GitHub*. (2019). Available online at: https://github.com/arundo/tsaug

52. Cohen M, Charbit M, Corff SL, Preda M, Noziere G. End-to-end deep metamodeling to calibrate and optimize energy loads. *ArXiv*. (2020) abs/2006.12390. doi: 10.1016/j.enbuild.2021.111218

53. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). p. 1003–12. doi: 10.1109/CVPR.2017.113

54. Li SJ, AbuFarha Y, Liu Y, Cheng MM, Gall J. MS-TCN++: Multi-stage temporal convolutional network for action segmentation. *IEEE Trans Pattern Anal Mach Intell*. (2020). doi: 10.1109/TPAMI.2020.3021756

55. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). p. 4724–33. doi: 10.1109/CVPR.2017.502

56. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. *arXiv preprint arXiv:170506950*. (2017).

57. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*. San Diego, CA (2015).

58. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdisc Rev*. (2010) 2:433–59. doi: 10.1002/wics.101

59. Choi S. Algorithms for orthogonal nonnegative matrix factorization. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Hong Kong (2008). p. 1828–32.

60. Balakrishnama S, Ganapathiraju A. *Linear Discriminant Analysis-A Brief Tutorial*. Institute for Signal and information Processing (1998). p. 1–8.

61. Bingham E, Mannila H. Random projection in dimensionality reduction: applications to image and text data. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining. Singapore (2001). p. 245–50. doi: 10.1145/502512.5 02546

62. Turaga P, Veeraraghavan A, Chellappa R. Unsupervised view and rate invariant clustering of video sequences. *Comput Vis Image Understand*. (2009) 113:353–71. doi: 10.1016/j.cviu.2008.08.009

63. Baran M, Lehrer N, Duff M, Venkataraman V, Turaga P, Ingalls T, et al. Interdisciplinary concepts for design and implementation of mixed reality interactive neurorehabilitation systems for stroke. *Phys Ther*. (2015) 95:449–60. doi: 10.2522/ptj.201 30581