sensors

MDPI

Article

Target-Specific Action Classification for Automated Assessment of Human Motor Behavior from Video

Behnaz Rezaei [1], **Yiorgos Christakis** [2], **Bryan Ho** [3], **Kevin Thomas** [4], **Kelley Erb** [2], **Sarah Ostadabbas** [1] and **Shyamal Patel** [2,*]

[1] Augmented Cognition Lab (ACLab), Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA; brezaei@ece.neu.edu (B.R.); ostadabbas@ece.neu.edu (S.O.)

[2] Digital Medicine & Translational Imaging Group, Pfizer, Cambridge, MA 02139, USA; Yiorgos.Christakis@pfizer.com (Y.C.); MichaelKelley.Erb@pfizer.com (K.E.)

[3] Neurology Department, Tufts University School of Medicine, Boston, MA 02111, USA; bho@tuftsmedicalcenter.org

[4] Department of Anatomy & Neurobiology, Boston University School of Medicine, Boston, MA 02118, USA; kipthoma@bu.edu

* Correspondence: Shyamal.Patel@pfizer.com

Received: 30 August 2019; Accepted: 28 September 2019; Published: 1 October 2019

Abstract: Objective monitoring and assessment of human motor behavior can improve the diagnosis and management of several medical conditions. Over the past decade, significant advances have been made in the use of wearable technology for continuously monitoring human motor behavior in free-living conditions. However, wearable technology remains ill-suited for applications which require monitoring and interpretation of complex motor behaviors (e.g., involving interactions with the environment). Recent advances in computer vision and deep learning have opened up new possibilities for extracting information from video recordings. In this paper, we present a hierarchical vision-based behavior phenotyping method for classification of basic human actions in video recordings performed using a single RGB camera. Our method addresses challenges associated with tracking multiple human actors and classification of actions in videos recorded in changing environments with different fields of view. We implement a cascaded pose tracker that uses temporal relationships between detections for short-term tracking and appearance based tracklet fusion for long-term tracking. Furthermore, for action classification, we use pose evolution maps derived from the cascaded pose tracker as low-dimensional and interpretable representations of the movement sequences for training a convolutional neural network. The cascaded pose tracker achieves an average accuracy of 88% in tracking the target human actor in our video recordings, and overall system achieves average test accuracy of 84% for target-specific action classification in untrimmed video recordings.

Keywords: action classification; human motor behavior; computer vision; deep learning; pose tracking

1. Introduction

Clinical assessment of human motor behavior plays an important role in the diagnosis and management of medical conditions like Parkinson's Disease (PD) [1]. However, such assessments can only be performed intermittently by trained clinical examiners, which limits the quantity and quality of information that can be collected to understand the impact of disease in the real-world setting. To address these limitations, significant efforts have been made to develop wearable sensing technologies that can be used for continuously monitoring various types of motor symptoms and behaviors [2–4]. While data collected using wearable sensors are well suited for detecting and measuring basic movements (e.g., arm or leg movements, tremor) and actions (e.g., sitting, standing, walking), they are ill-suited when it comes to complex activities (e.g., cooking, grooming) and behaviors (e.g., personal habits, routines)—particularly if they involve the interpretation of environmental interactions (e.g., with other humans, animals, or objects). Understanding the various factors that influence physical behavior can help clinicians better understand the impact of motor and non-motor symptoms on the daily life of patients with PD [5].

Recently, artificial intelligence (AI) assisted classification of human behavior using computer vision has received newfound attention among researchers in machine learning and pattern recognition communities for applications spanning from automatic recognition of daily life activities in smart homes to monitoring the health and safety of elderly and patients with mobility disorders in their homes/hospitals [6–12]. However, in contrast to wearable devices, vision-based approaches pose a greater risk to privacy and security of an individual [13]. Vision-based assessment of human behavior enables us to automate the detection and measurement of the full range of human behaviors. As illustrated in Figure 1, the taxonomy of human behaviors can be viewed as a four-level hierarchical framework with basic movements at the bottom (e.g., movement of body segments) and complex behaviours (e.g., personal habits and routines) at the top. Automatic recognition at any level requires that actions and/or behaviors at the level below it are also recognized. For example, in order to recognize walking, we first need to assess if the pose is upright, the arms are swinging and legs are moving. At the first level (motion), recognition deals with tasks such as movement detection or background extraction/segmentation in video recordings of the target [14–16]. These techniques try to locate the moving objects in a scene by extracting a silhouette of the object in a single frame or over a few consecutive frames. However, segmentation algorithms without any further processing provide only very basic pose estimation of the object with little to no temporal information. At the second level (action), human movements along with environmental interactions are classified in order to recognize what the target is doing over a period of seconds or minutes [17]. At the third level (activity), the recognition task is focused on identifying activities as a combination of sequence of actions and environmental interactions over a period of minutes to hours. Finally, at the fourth level (behavior), sequence of activities and environmental interactions along with information about their temporal dependencies are used to recognize complex human behaviors.

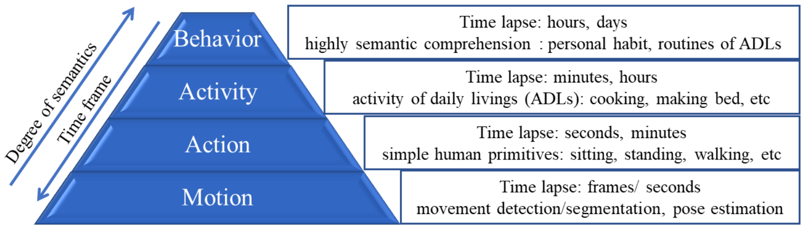

Degree of semantics
Time frame
Behavior
Activity
Action
Motion
Time lapse: hours, days
highly semantic comprehension : personal habit, routines of ADLs
Time lapse: minutes, hours
activity of daily livings (ADLs): cooking, making bed, etc
Time lapse: seconds, minutes
simple human primitives: sitting, standing, walking, etc
Time lapse: frames/ seconds
movement detection/segmentation, pose estimation

Figure 1. Taxonomy of human behaviors with different levels of semantics and complexity. Recognition of each level requires most of the underlying tasks to be recognized [6].

1.1. Our Contributions

Automated assessment of human behavior in multi-person video (i.e., when several people are present in the video) requires the tracking and classification of a sequence of actions performed by a target (e.g., patient). Therefore, accurate temporal tracking of the target is an essential requirement for this application, along with robust feature extraction that can be used for classifying human behaviors at different levels of complexity. In this paper, we present a hierarchical target-specific action classification method, which is illustrated as a block diagram in Figure 2. Detection of different actions performed by the target is done using pose evolution feature representation. We define pose evolution as a low-dimensional embedding of a sequence of posture movements that are required to perform an action (e.g., walking). In order to find the pose evolution feature representation corresponding to the target, we present a cascaded target pose tracking algorithm that receives multi-person pose estimation results from an earlier stage and tracks the target pose throughout the video. Our main contributions in this paper are: (1) development of a robust hierarchical multiple-target pose tracking method to facilitate action recognition in videos recorded in uncontrolled environments in the presence of multiple human actors; (2) introducing pose evolution, an explicit body movement representation, as complementary information to the appearance and motion cues for robust action recognition; and (3) a novel target-specific action classification architecture applied to untrimmed video recordings of patients with PD.

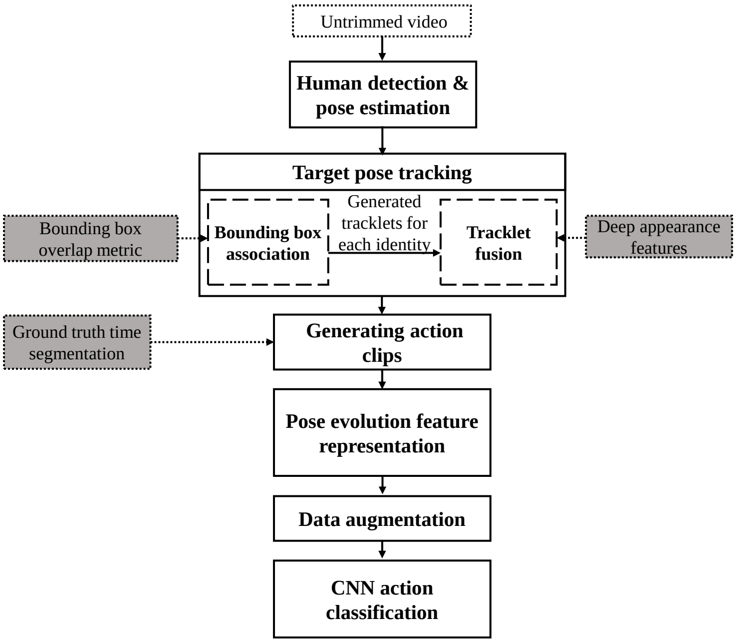

Untrimmed video
Human detection & pose estimation
Target pose tracking
Bounding box overlap metric
Bounding box association
Generated tracklets for each identity
Tracklet fusion
Deep appearance features
Ground truth time segmentation
Generating action clips
Pose evolution feature representation
Data augmentation
CNN action classification

Figure 2. Overview of the proposed multi-stage method for human behavior phenotyping in untrimmed videos. At the first stage, human detection and pose estimation are applied to the recorded video. At the second stage, the regressed bounding boxes for each detected person and corresponding keypoints are used for tracking the identities in the video. Tracking is done in an incremental process incorporating both appearance and time information. Outputs of tracking the target identity along with ground-truth time segmentation are used for generating a compact representation of the target actor pose evolution in time for each action clip. Finally, the augmented pose evolution representation is fed to a convolutional neural network (CNN)-based action classification network to recognize actions of interest.

1.2. Related Works

Assigning a single action label to a multi-person video clip dilutes the specificity of information and makes it less meaningful. For many real-world applications such as video-based assessment of human behavior, there is a need for person-centric action recognition, which assigns an action label to each person in a multi-person video clip. One of the challenges in person-centric action recognition is robust tracking of the target in long-term videos. Tracking is challenging because there are many sources of uncertainty, such as clutter, occlusions, target interactions, and camera motion. However, most of the research studies on human activity classification have typically dealt with videos with a single human actor or video clips with ground-truth tracking provided [18], with the exception of few that performed human-centeric action recognition [8,19]. Girdhar, et al. [19] re-purposed an action transformer network to exclude non-target human actors in the scene and aggregated spatio–temporal features around the target human actor. Chen, et al. [8] presented human activity classification using skeleton motions in videos with interference from non-target objects aimed at supporting applications in monitoring frail and elderly individuals. However, neither work provided details on how they addressed non-target filtering in their human action classification pipelines.

Beside the importance of dealing with non-target objects in providing a well-performing real-world human action recognition system, creating robust and discriminating feature representations for each video action clip plays an important role in detecting different human activities [20]. Most of the state-of-the-art action recognition architectures process appearance and motion cues in two independent streams of information, which are fused right before the classification phase or a few stages before the classification stage in a merge and divide scheme [21,22]. Others have used 3D spatio-temporal convolutions to directly extract relevant spatial and temporal features [23–25]. However, human pose cues, which can provide low-dimensional interpretations for different activities, have been overlooked in these studies. Most recently, Choutas, et al. [26] and Mengyuan, et al. [27] used temporal changes of pose information with two different representations for boosting action recognition performance. In [27], authors claim that if there are multiple people in the scene, pose motion representation does not need the time associations of the joints to work but they did not address how their proposed method can handle multiple human actors in a video.

In general, convolutional neural network (CNN) based action recognition approaches can be divided into three different categories based on their underlying architecture: (1) spatio-temporal convolutions (3-dimensional convolutions), (2) recurrent neural networks, and (3) two stream convolutional networks. The benefit of multi-stream networks is that different modalities can be aggregated in the network to improve performance of the final action classification task. In this paper, we addressed the problem of person-centric action recognition by long-term tracking of the target human actor. In addition, our method provides a novel pose evolution representation of the target human actor rather than the common spatio-temporal features extracted from raw video frames to the classification network. It is worth mentioning that our pose-based action recognition stream can be used to augment the current multi-stream action classification networks.

The rest of the paper is organized as follows. In Section 2, we describe the proposed method for tracking target human actor in untrimmed videos in order to extract appropriate pose evolution features from actions performed in a video. In Section 3, we describe the subsequent stages for action classification (illustrated in Figure 2), which include pose evolution feature representation and classification network. We present our experimental setup and performance evaluation results of the proposed method in Section 4. Finally, we discuss the results in Section 5 and conclude our paper in Section 6.

2. Target Pose Tracking

Diverging from the common approach of learning spatio-temporal features from videos for action classification, pose-based action classification methods have shown promising results by providing a compact representation of human pose evolution in videos [26–29]. The temporal evolution of pose can be used as the only discriminating feature information for classification of human actions that involve different pose transitions (e.g., walking). This approach can further be combined with spatio-temporal features to improve the performance of context-aware action classification in the case of more complex behaviors (e.g., moving an object from one place to another).

The primary task in pose-based action classification in untrimmed videos is locating the target. This requires a robust estimation and tracking of human body poses by addressing the challenges associated with long-term videos recorded for assessment of human motor behavior. These challenges include partial to complete occlusion, change of scene, and camera motion. In this section, we propose a cascaded multi-person pose tracking method using both time and appearance features, which will be used in later steps to generate pose evolution feature representations for action classification.

2.1. Human Pose Estimation

In order to extract human pose information in each video frame along with their associated bounding boxes as the first step in our system, we used a 2D version of the state-of-the-art human pose estimation method proposed in [30]. The pre-trained model performs efficient frame-level multi-person pose estimation in videos using the Mask R-CNN network [31]. This model was initialized on ImageNet [32] and then trained on the COCO keypoint detection task [33]. The Mask R-CNN network was then fine-tuned on the PoseTrack dataset [34]. The architecture of this pose estimation network is illustrated in Figure 3. The network uses ResNet-101 [35] as the base convolutional network for extracting image features. Extracted features are then fed to a region proposal network (RPN) trained to highlight regions that contain object candidates [36]. Candidate regions of the output feature map are all aligned to a fixed resolution via a spatial region of interest (ROI)-align operation. This operation divides feature maps that may have different sizes depending on the size of detected bounding boxes to a fixed number of sub-windows. The value for each sub-window is calculated by finding a bi-linear interpolation of four regularly sampled locations inside the sub-window. The aligned features are then fed into two heads, a *classification head* responsible for person detection and bounding box regression, and a *keypoint head* for estimating the human body joints defined as a human pose in each detected bounding box. The outputs of this pose estimation network are seventeen keypoints associated with various body joints and a bounding box surrounding each person.

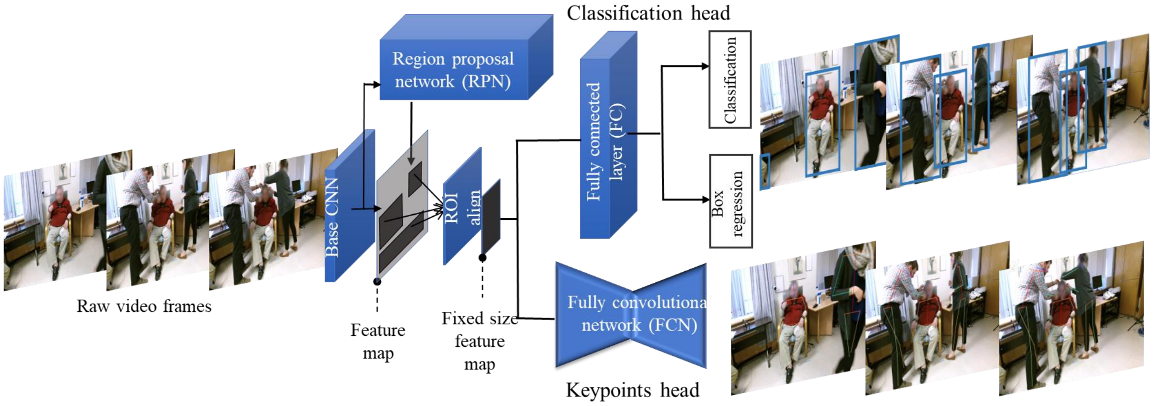


Figure 3. Architecture of the pose estimation network. Each video frame is fed separately to the base network (ResNet 101) for feature extraction. A region proposal network is applied on the output feature map to find the areas with the highest objectness probability. The fixed size features for proposed regions are then given to the classification and pose estimation heads to find the human bounding boxes and their corresponding keypoints.

2.2. Cascaded Pose Tracking

In many real-world settings where a person has to be tracked across videos recorded from different cameras located in different environments, a single tracker is unable to track the person throughout the video and all of them fail when the target leaves one environment and appears into another environment or is occluded from the camera view and then reappears in the camera's field of view [37]. In order to address this problem of tracking people in videos recorded in multiple environments (in our case different rooms and hallways) various person, re-identification methods have been proposed [38–40]. Most of the existing re-identification (re-id) methods are supervised with the assumption of availability of large manually labeled matching identity pairs. This assumption does not hold in many practical scenarios (such as our dataset) where the model has to be generalizable for any person and providing manually labeled identity matches is not feasible. Unsupervised learning for person re-id has become important in various scenarios where the system needs to be adapted to new identities such as video surveillance applications [41,42]. In this work, we have adapted the idea of person re-id, which is used for the matching the identities among non-overlapping cameras for tracking the target throughout the non-overlapping videos. This would address challenges such as changing environments or turning away from the camera, which can be treated as the case of re-identification across different non-overlapping cameras. In the traditional re-id problem, we typically have a gallery of images containing the images taken using different cameras for different identities. Given a probe image, the aim is to match the probe identity with the images in the gallery that belong to the same identity as the probe. In our problem of long-term tracking of the target human (patient) in videos, we have a set of tracklets and a given probe (an image of the target) and the aim is to fuse all the tracklets in the set which belong to the same identity as the probe in order to find the single track of the patient throughout the video. In contrast to the re-identification problem, multiple tracklets are generated because of the failure in the tracking of the target throughout the video because of the occlusions, change of environment and abrupt camera motions.

In order to continuously track the pose of the target (i.e., the subject in our dataset) in video recordings, we propose a two-step procedure based on the estimated bounding boxes and keypoints provided by the pose estimation network in Section 2.1. As illustrated in Figure 4, in the first stage (short-term tracking, Section 2.2.1) we use a lightweight data association approach to link the detected bounding boxes in consecutive frames into tracklets. Tracklets are a series of bounding boxes in consecutive frames associated with the same identity (person). In the next stage (long-term tracking, Section 2.2.2), we fuse tracklets of the same identity using their learned appearance features to provide continuous tracking of the target actor across the entire video recording. The implementation details are described in Section 2.2.2.

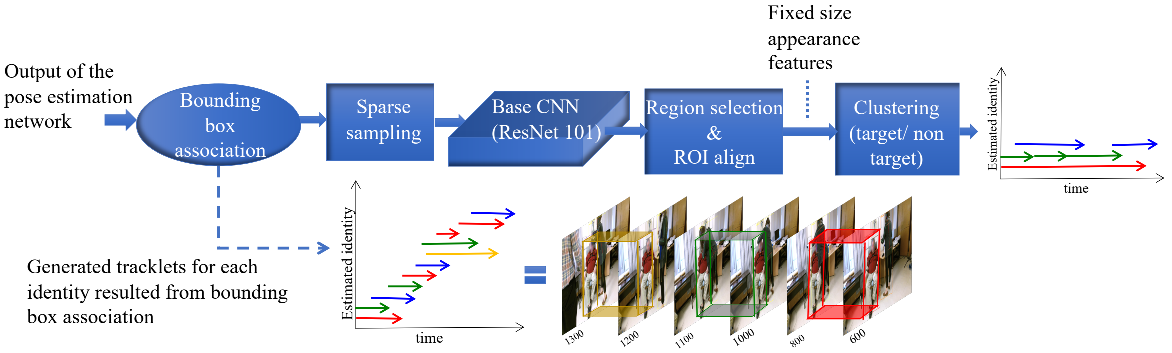

Output of the pose estimation network
Bounding box association
Sparse sampling
Base CNN (ResNet 101)
Region selection & ROI align
Fixed size appearance features
Clustering (target/ non target)
Estimated identity
time
Generated tracklets for each identity resulted from bounding box association
Estimated identity
time
1300
1200
1100
1000
800
600

Figure 4. Hierarchical pose tracking using temporal and appearance features. Tracking starts by associating detected bounding boxes in each pair of consecutive frames using the intersection over union metric. Output of this step is a number of different tracklets for each identity. At the next step generated tracklets are pruned based on their length, and pose estimation confidence followed by sparse sampling. Finally, generated tracklets which belong to the target identity are merged according to their appearance similarity to create the endpoint track for the target human actor (best viewed in color).

2.2.1. Short Term Tracking Based on Temporal Association

Given the detected bounding boxes for each person in the video, we link the bounding boxes that belong to the same identity in time to create pose tracklets. Assuming that there is no abrupt movement in the video, tracklets are generated by solving a data association problem with similarity measurement defined as the intersection over union between the currently detected bounding boxes and the bounding boxes from the previous frame. Like [30,43], we formulate the task as a bipartite matching problem, and solve it using the Hungarian algorithm [44]. We initialize tracklets on the first frame and propagate the labels forward one frame at a time using the matches. Any box that does not get matched to an existing tracklet instantiates a new tracklet. This method is computationally efficient and can be adapted to any video length or any number of people. However, tracking can fail due to challenges such as abrupt camera motion, occlusions and change of scene, which can result in multiple tracklets for the same identity. For instance, as illustrated in Figure 4, short term tracking generates 3 distinct tracklets for the target in just 700 consecutive frames (23 s).

2.2.2. Long Term Tracking using Appearance based Tracklet Fusion

Given the large number of tracklets generated from the previous stage (i.e., short term tracking), we fuse tracklets that belong to the same identity to generate a single long-term track for the target. As illustrated in Figure 4, in order to merge the generated tracklets belonging to the same identity throughout the video, we first apply sparse sampling by pruning the tracklets based on their length and the number of estimated keypoints, and then selecting the highest confidence bounding box from each tracklet. Finally, we merge the tracklets into a single track based on their similarity to the reference tracklet. The affinity metric between the tracklet T_i, and the reference tracklet T_{ref}, is calculated as:

$$P_a(T_i, T_{ref}) = ||f_i^{t'} - f_{ref}^{t}||_2, \tag{1}$$

where $f_i^{t'}$ is feature vector of the sampled detection in tracklet T_i at time t', and f_{ref}^{t} is feature vector of the sampled detection in reference tracklet T_{ref} at time t. Affinity metric, $P_a(.)$ is the Euclidean distance between the above feature vectors. In order to extract deep appearance features, we feed every sampled detection of each tracklet to the base network of a Mask R-CNN (i.e., ResNet-101), which has been trained on the PoseTrack dataset for pose estimation [31,35]. The extracted feature map is then aligned spatially to a fixed resolution via ROI-align operation. It is worth mentioning that we do not pay an extra computational cost for learning the features for merging the associated tracklets of the target into one track. In order to show the importance of the target tracking in the performance of the action classification network we trained the action classification network on the pose evolution maps without any tracking involved, more details are provided in Section 4.3.

3. Action Classification Based on Pose Evolution Representation

After locating the target, providing a compact yet discriminative pose evolution feature representation for each action clip plays an essential role in recognizing different actions. To achieve this, we first provide a compact spatio-temporal representation of the target's pose evolution in Section 3.1 for each video clip inspired by PoTion pose motion representation introduced in [27]. Then, we use the tracked pose evolution to recognize five categories of human actions: sitting, sit-to-stand, standing, walking, and stand-to-sit in Section 3.2.

3.1. Pose Evolution Representation

By using pose of the target for each frame of the video clip provided by the pose tracking stage, we create a fixed-size pose evolution representation by temporally aggregating these pose maps. Pose tracking in preceding stages gives us locations of the body joints of the target (i.e., the subject in our case) in each frame of the video clip. We first generate joint heatmaps from given keypoint positions by locating a Gaussian kernel around each keypoint. These heatmaps are gray scale images showing the probability of the estimated location for each body joint. The pose evolution representations are created based on these joint heatmaps.

As illustrated in Figure 5, in order to capture the temporal evolution of pose in a video clip, after generating pose heatmaps for the target actor in a video frame, we colorize them according to their relative time in the video. In other words, each gray scale joint heatmap of dimension $H \times W$ generated for the current frame at time t is transformed into a C-channel color image of $C \times H \times W$. As indicated in Equation (2), this transformation is done by replicating the original heatmaps C times and multiplying values of each channel with a linear function of the relative time of the current frame in the video clip.

$$Je_i(j, x, y) = \frac{\sum_{t=0}^{T-1} JH_i^t(x, y) \times oc_j(t)}{max_{x,y} \sum_{t=0}^{T-1} JH_i^t(x, y) \times oc_j(t)} \tag{2}$$
$$\text{for } i \in \{1, 2, ..., 14\}, \ \ j \in \{1, ..., C\}$$

where $JH_i^t(x,y)$ designates the estimated joint heatmap for joint number i of the target in a given frame number t. $oc_j(t)$ is the linear time encoding function for channel j evaluated at time t. Je_i is the joint evolution representation for each joint i. The final pose evolution representation, Pe is derived by concatenating all calculated joint evolutions, as $Pe = concatenate(Je_1, Je_2, ..., Je_{14})$, where we have 14 joints given by reducing the head keypoints to one single keypoint.

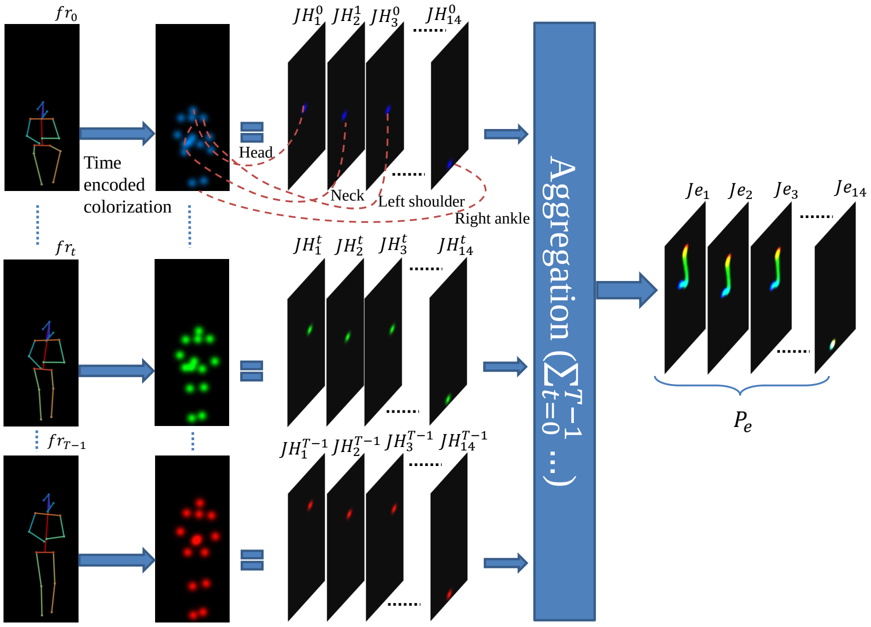

fr_0
Time
encoded
colorization
Head
Neck
Left shoulder
Right ankle
JH_1^0 JH_2^1 JH_3^0 JH_{14}^0
fr_t
JH_1^t JH_2^t JH_3^t JH_{14}^t
fr_{T-1}
JH_1^{T-1} JH_2^{T-1} JH_3^{T-1} JH_{14}^{T-1}
Aggregation ($\sum_{t=0}^{T-1}$ ...)
Je_1 Je_2 Je_3 Je_{14}
P_e

Figure 5. Illustration of the pose evolution feature representation in Figure 2 for the sit-to-stand task. Given the estimated keypoints of the target human actor from preceding stages in the first column, colorized joint heatmaps in the second column are generated using the time encoding function represented in Figure 6. The final pose evolution representation is generated by aggregating and normalizing the colorized joint heatmaps in time (best viewed in color).

In order to calculate the time encoding function for a C-channel pose evolution representation, the video clip time length T is divided into $C-1$ intervals with duration $l = \frac{T}{C}$ each. For each given frame at time t that sits in kth interval which $k = \lceil \frac{t}{T} \rceil$, $oc_j(t)$ is defined as follows:

$$oc_j(t) = \begin{cases} \frac{(-t + \frac{kT}{C-1})}{l}, & \text{for } j = k \\ \frac{(t - \frac{T(k-1)}{C-1})}{l}, & \text{for } j = k+1 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Figure 6 illustrates the time encoding functions that are defined based on the Equation (3) for 3-channel colorization used in our pose evolution representation. After creating the pose evolution representations, we augment them by adding white noise to our representation to train the action classification network.

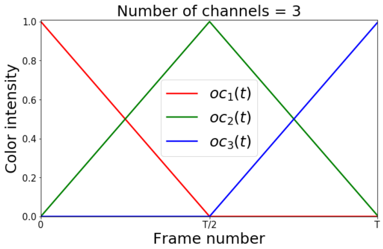


Figure 6. Demonstration of the time encoded colorization method utilized for creating body pose motion map representation. $oc_1(t)$, $oc_2(t)$, and $oc_3(t)$ show the time encoding function for each color channel.

3.2. Classification Network

We trained a CNN for classifying different actions using the pose evolution representations. Since pose evolution representations are very sparse and have no contextual information of the raw video frames, the network does not need to be very deep or pre-trained to be able to classify actions. We used the network architecture illustrated in Figure 7 consisting of 4 fully convolutional layers (FCN), and one fully connected layer (FC) as the classifier. The input of the first layer is the pose evolution representation of size 14 $C \times H \times W$, where 14 is the number of body joints that are used in our feature representation. In this work, we used $C = 3$ as the number of channels for encoding the time information into our feature representation. In Section 4.3, we explore the effect of number of channels on the performance of the action classification network.

The action classification network includes two blocks of convolutional layers, a global average pooling layer, and a fully connected layer with a Softmax loss function as the classification layer. Each block contains two convolution layers with filter sizes of $3 \times 3 \times 128$, and $3 \times 3 \times 256$, respectively. The first convolution layer in each block is designed with a stride of 2 pixels and a second layer with a stride of 1 pixel. All convolutional layers are followed by a rectified linear unit (ReLU), batch normalization, and dropout. We investigated the performance of several variations of this architecture on action classification in Section 4.

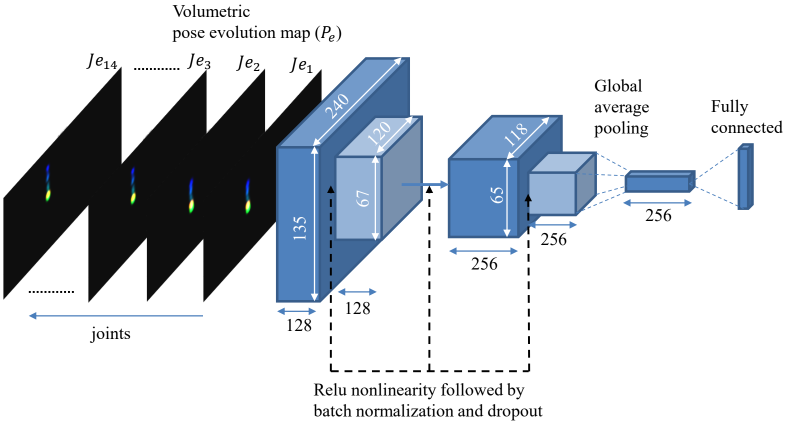


Figure 7. Architecture of the action classification network. This network takes the volumetric pose evolution map of the target human actor from a video clip as the input and classifies occurrence of an action in the video into one of the five predefined actions (best viewed in color).

4. Experiments

To evaluate the performance of the proposed approach, we used a real-world dataset collected in a neurology clinic. We provide an overview of the dataset in Section 4.1, and report on the performance of target tracking and action classification in Section 4.2 and Section 4.3 respectively.

4.1. Dataset

Our dataset consists of video recordings of 35 patients with Parkinson's disease (Age: 68.31 ± 8.03 (46–79) years; Sex: 23M/12F; Hoehn & Yahr I/II/III: 2/26/7; MDS-UPDRS III: 52.86 ± 16.03) who participated in a clinical study to assess changes in their motor symptoms before (OFF state) and after (ON state) medication intake. Individuals with a clinical diagnosis of PD between 30–80 years old, able to recognize wearing-off periods, with Hoehn & Yahr stage $\leq$ III and currently on L-dopa therapy were eligible to participate in this study. Exclusion criteria included the presence of other comorbidities (e.g., head injuries, psychiatric illness, cardiac disorders), recent treatment with investigational drugs, pregnant women and allergy to silicone or adhesives. The study had approval from the Tufts Medical

Center and Tufts University Health Sciences Institutional Review Board (study ID: HUBB121601) and all experimental procedures were conducted at Tufts Medical Center [45]. All subjects provided written informed consent.

The study protocol included two visits to the clinic; subjects were randomly assigned to be in the ON (after medication intake) or OFF (before medication intake) state for the first visit, and underwent the second visit in the other state. During each study visit, patients performed a battery of motor tasks including activities of daily living (e.g., dressing, writing, drinking from a cup of water, opening a door, folding clothes) and a standard battery of clinical motor assessments from the Movement Disorder Society's Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [46] administered by a trained clinician with experience in movement disorders. Each visit lasted approximately 1 h and most of the experimental activities were video recorded at 30 frames per second by two Microsoft Kinect$^{\mathrm{TM}}$ cameras (1080 × 1920-pixel resolution), one mounted on a mobile tripod in the testing room and another on a wall mount in the adjacent hallway. In total, the dataset consists of 70 video recordings (35 subjects × 2 visits per subject). The video camera was positioned to capture a frontal view of the subject at most times. Besides the subject, there are several other people (e.g., physicians, nurses, study staff) who appear in these video recordings.

Behaviors of interest were identified within each video using structured definitions and, their start and end times annotated using human raters as described elsewhere [47]. Briefly, each video recording was reviewed and key behaviors annotated by two trained raters. To maximize inter-rater agreement, each behavior had been explicitly defined to establish specific, anatomically based visual cues for annotating its start and end times. The completed annotations were reviewed for agreement by an experienced arbitrator, who identified and resolved inter-rater disagreements (e.g., different start times for a behavior). The annotated behaviors were categorized into three classes: postures (e.g., walking, sitting, standing), transitions (e.g., sit-to-stand, turning), and cued behaviors (i.e., activities of daily living and MDS-UPDRS tasks). In this manuscript, we focus on the recognition of postures (sitting, standing and walking) and transitions (sitting-to-standing and standing-to-sitting). Recognizing these activities in PD patients provide valuable context for understanding motor symptoms like tremor, bradykinesia and freezing of gait. Major challenges in recognizing activities of the target (i.e., subject) in this dataset were camera motion (when not on tripod), change of scene as the experimental activities took place in different environments (e.g., physician office, clinic hallway, etc.) and long periods occlusion (around a few minutes) due to interactions between the patient and the study staff.

4.2. Tracking Target Human and Pose

Given that video recordings involved the presence of multiple people, we first detected all human actors along with their associated keypoints in each video frame using the multi-person pose estimation method described in Section 2.1 (illustrated in Figure 3). This pose estimation network was pre-trained on the COCO dataset and fine-tuned on the PoseTrack dataset previously [30,33,34]. As illustrated in Figure 4, the output of this stage is a list of the bounding boxes for human actors detected in each video frame and the estimated locations of keypoints for each person along with a confidence estimate for each keypoint.

In order to recognize activities of the target, we first locate and track the subject (i.e., PD patient) in each frame. This was accomplished by using the hierarchical tracking method described in Section 2.2. Given all detected bounding boxes across all frames from the pose estimation stage, we first generate tracklets for each identity appearing in the video via short-term tracking explained in Section 2.2.1. Each tracklet is a list of detected bounding boxes in consecutive frames that belong to the same identity. In order to find the final patient track for the entire video, we use the long-term tracking method described in Section 2.2.2 to remove non-target tracklets (e.g., study staff, physician, nurse) and fuse the tracklets that belong to the patient using the appearance features. There is no supervision in tracking of the patient during the video except providing a reference tracklet, which is associated to with the target (i.e., subject) in the long-term tracking step.

To evaluate the performance of our target tracking method, we first manually annotated all tracklets generated by short-term tracking and then calculated accuracy of the long-term tracking method with respect to the manually generated ground-truth. Accuracy is calculated by treating the long-term tracker as a binary classifier as it excludes non-patient tracklets and fuses tracklets belonging to the target to find a single final patient track for the entire video recording. Considering patient tracklets as the positive class and non-patient tracklets as the negative class, our tracker achieved an average classification accuracy of 88% across 70 videos on this dataset.

4.3. Action Classification

In the last stage of our multi-stage target-specific action classification system, we trained a CNN to recognize the following five actions of interest: sitting, standing, walking, sitting-to-standing, and standing-to-sitting. After applying the target pose tracking system illustrated in Figure 4, we segmented the resulting long-term video into action clips based on ground-truth annotations provided by human raters. Although the action clips have variable lengths (ranging from a few frames to more than 10 min), each video clip includes only one of the five actions of interest. As a result, we ended up with a highly imbalanced dataset. In order to create a more balanced dataset for training and evaluating the action classification network, we first excluded action clips less than 0.2 s (too short for dynamic activities like walking) and divided the ones longer than four seconds into four-second clips. Assuming that four seconds is long enough for most activities of interest and below 0.2 s (lower than six frames) is too short to be used for recognizing an action [48]. This resulted in a total of 44,580 action clips extracted from video recordings of 35 subjects. We used 29 subjects (39,086 action clips) for training/validation set and the remaining 6 subjects (5494 action clips) were held out for testing. As shown in Figure 8, the resulting dataset is highly imbalanced with a significant skew towards the sitting class, which can result in over-fitting issues. To address this imbalance, we randomly under-sampled the walking, sitting, and standing classes to 4000 video clips each.

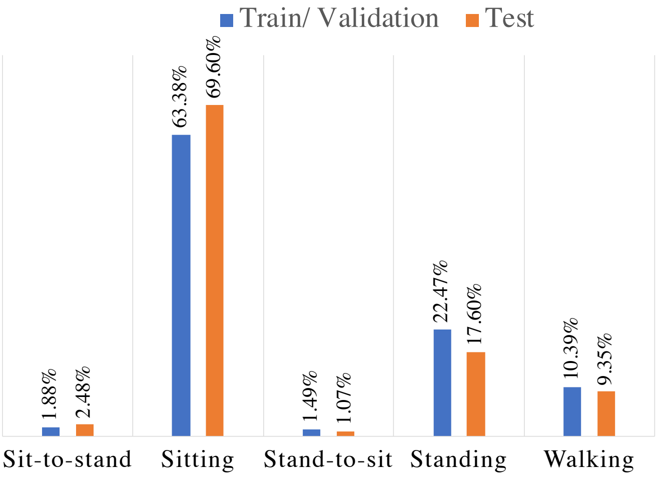


Figure 8. Distribution of the action clips based on the type of the actions for test and train/validation datasets. The distribution of the original set of action clips is highly imbalanced.

To prepare input data for the action classification network, we transformed each action clip into a pose evolution representation as described in Section 3.1. To create the pose evolution maps, we scaled the original size of each video frame (1080 × 1920) by a factor of 0.125 and chose 3 channels to represent the input based on training time and average validation accuracy in diagnostic experiments. The training dataset was also augmented by adding Gaussian noise. In addition, we tried data augmentation techniques like random translation and flipping during our diagnostic experiments, but the classification performance degraded by about 3%. Therefore, we only used additive Gaussian noise to randomly selected video frames as the only type of data augmentation.

We used 90% of the train/validation dataset for training the action classification network with architecture illustrated in Figure 7 and the rest for validation. The network training started with random weight initialization and we used the Adam optimizer with a base learning rate of 0.01, a batch size of 70 and a dropout probability of 0.3. We experimented with several variants of the network architecture proposed in Section 3 by increasing the number of the convolution blocks to three and changing the number of filters in each block to 64, 128, 256, and 512. Based on the performance on the validation set and training loss, Figure 7 provided the best performance while avoiding over-fitting to the training data. In addition, we investigated the impact of using a different number of channels for representing the temporal pose evolution on the performance of action classification. Figure 9 illustrates the accuracy of the classification network with different representations as input. We chose 3 channels for our representation because adding more channels would only increase the computational cost without any significant improvement in accuracy. The trained action classification model achieved a best-case weighted classification accuracy of 83.97% on the test dataset. In order to demonstrate the importance of target tracking on the performance of the action classification network, we conducted another experiment without using any tracking on the recorded videos. The results show that while the best case weighted overall accuracy for the validation set was slightly better (84.04%), it dropped to 63.14% on the test set. This is an indication that the model is not able to generalize well, because the quality of training data degrades without target tracking. More details of the classification performance including per class accuracy in the test and validations phase can be found in Table 1.

Table 1. Best case classification accuracy (%) per action class on the validation and test set with and without long-term tracking. Weighted overall accuracy was calculated to account for the class imbalance. Mean and standard deviation (std) of the weighted average accuracy (last column) were calculated by training the network 10 times using the same hyper-parameters but with different initialization and evaluating it on the validation split.

	Sit	**Sit-to-Stand**	**Stand**	**Walk**	**Stand-to-Sit**	**Weighted Overall Accuracy**	**Mean ± Std. of Average Accuracy**
With long-term tracking							
Validation	92.8	68.1	81.5	78.9	70.7	82.00	79.85 ± 2.38
Test	91.6	75.0	85.7	81.0	78.6	83.97	-
Without long-term tracking							
Validation	90.9	88.1	91.0	71.8	75.8	84.04	71.42 ± 10.32
Test	72.6	63.9	81.6	51.7	16.3	63.14	-

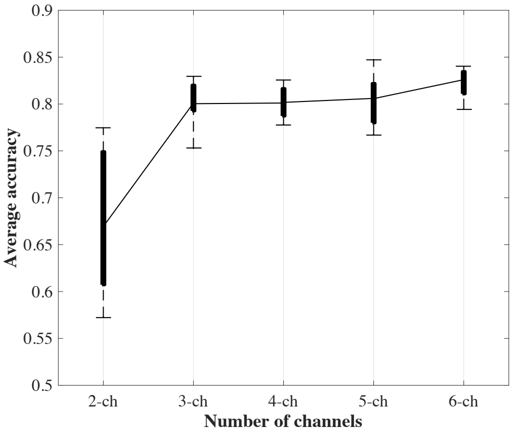


Figure 9. Average classification accuracy with respect to the number of channels of input pose evolution representations.

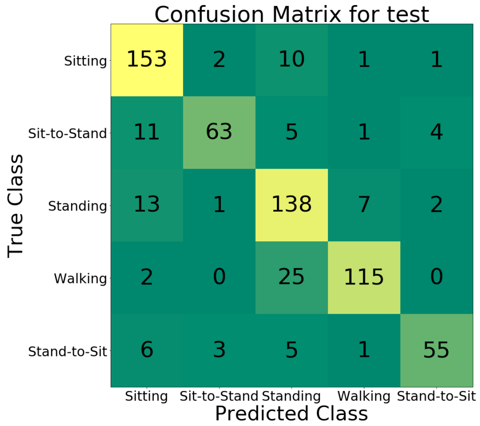
5. Discussion

Real-world assessment of motor behaviour can provide valuable clinical insights for diagnosis, prognosis and prevention of disorders ranging from psychiatry to neurology [49,50]. In this paper, we propose a new approach for automated assessment of target-specific behavior from video recordings in the presence of other actors, occlusions and changes in scene. This approach relies on using temporal relationships for short-term tracking and appearance-based features for long-term tracking of the target. Short-term tracking based on temporal relationships between adjacent frames resulted in 1466 ± 653 tracklets per video, which were then fused by using appearance-based features for long-term tracking. Using this approach, we were able to identify the target track throughout the video recording with an accuracy of 88% in our dataset of 70 videos belonging to 35 targets (i.e., PD patients). However, one of the limitations of our dataset was that the target's appearance did not change significantly (except for a brief period when the subject put a lab coat on to perform a task) over the duration of the recording. This is unlikely in the real-world as we expect appearance to change on a daily basis (e.g., clothing, makeup) as well as over weeks and months (e.g., age or disease-related changes). Therefore, the proposed method requires further validation on a larger dataset collected during daily life and would benefit from strategies for dealing with changes in appearance.

The second aspect of our work focused on classification of activities of daily living. Activities like sitting, standing, sit-to-stand, stand-to-sit and walking are basic elements of most of the tasks that we perform during daily life. To train the activity classification model, we used pose evolution representations to capture both temporal and spatial features associated with these activities. While this model achieved a classification accuracy of 84%, as we can see in Figure 10, a significant source of error was the misclassification of 18% (25/142) of walking as standing. This could be attributed to two factors. Firstly, video recordings of the walking activity were performed with a frontal view of the subject, which limits the ability of pose evolution representations to capture features associated with spatial displacement during walking. As a result, pose evolution representations of walking and standing look similar. This would be challenging to deal with in real-world scenarios because the camera's field of view is typically fixed. This limitation highlights the need for developing methods that are robust to changes in feature maps associated with different fields of views. Secondly, the activity transition period from standing to walking was labeled as walking during the ground-truth annotation process. As a result, when the action classification network is applied to short action clips, those containing such transitions are more likely to be misclassified as standing. Examples of the aforementioned misclassification are illustrated in Figure 11. In Figure 11a the subject takes a couple of steps to reach for a coat hanging on the rack and in Figure 11b the subject is about to start walking from a standing position. In both cases, the ground-truth annotation was walking but the video clip was classified as standing. This is a potential limitation of a video-based action recognition approach as its performance will be dependent on factors like the camera view. By comparison, approaches using one or more wearable sensors (e.g., accelerometers and gyroscopes on the thigh, chest and ankle [51]) are relatively robust to such problems as their measurements are performed in the body frame of reference, which results in high classification accuracy (>95%) across a range of daily activities.

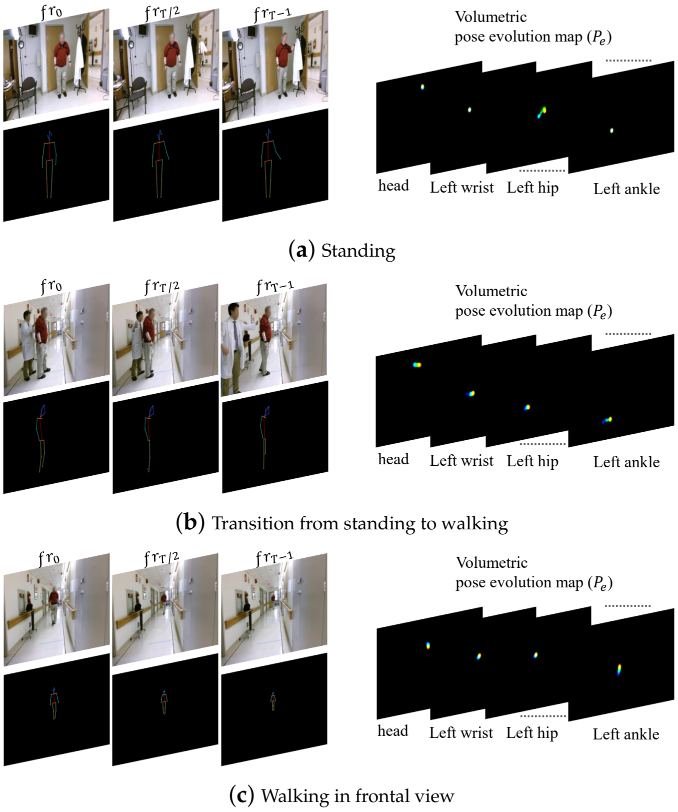
Another source of error which impacts overall performance is the error propagated from pose tracking (~12%) and pose estimation stages. Pose estimation error can be tolerated to some degree by aggregating the colorized joint heatmaps in the pose evolution feature representation. However, since the output of the pose tracking is directly used for generating pose evolution maps, any error in tracking the patient throughout the video would negatively impact the action classification performance. One approach for tolerating the error from pose tracking stage is to incorporate raw RGB frames as the second stream of information for action classification and using attention maps based on the tracking outcome rather than excluding non-target persons from the input representations.

Vision-based monitoring tools have the distinct advantage of being transparent to the target, which would help with issues of compliance associated with the use of wearable devices. Also, unlike wearable devices, vision-based approaches can capture contextual information, which is necessary for understanding behavior at higher level. However, this also comes at an increased risk to privacy for the target as well as other people in the environment. The proposed approach can potentially mitigate this concern by limiting monitoring to the target (e.g., patient) and transforming data at the source into sparse feature maps (i.e., pose evolution representations).



	Sitting	Sit-to-Stand	Standing	Walking	Stand-to-Sit
Sitting	153	2	10	1	1
Sit-to-Stand	11	63	5	1	4
Standing	13	1	138	7	2
Walking	2	0	25	115	0
Stand-to-Sit	6	3	5	1	55

Figure 10. Confusion matrix of the action recognition network evaluated on the test dataset.



(**a**) Standing

(**b**) Transition from standing to walking

(**c**) Walking in frontal view

Figure 11. An example of the misclassification of walking as standing. (**a**–**c**) The first, middle, and last frame of three action video clips along with the corresponding pose estimations and pose evolution maps. During the manual annotation process (**a**) was labeled as standing, whereas (**b**,**c**) were labeled as walking. The action classification network classifies (**a**,**b**) as standing because they have a very similar pose evolution map (best viewed in color and zoomed in).

6. Conclusions and Future Work

In this paper, we have presented an AI-assisted method for automatic assessment of human motor behavior from video recorded using a single RGB camera. Results demonstrate that the multi-stage method, which includes pose estimation, target tracking and action classification, provides an accurate target-specific classification of activities in the presence of other human actors and is robust to changing environments. The work presented herein focused on the classification of basic postures (sitting, standing and walking) and transitions (sitting-to-standing and standing-to-sitting), which commonly occur during the performance of many daily activities and are relevant to understanding the impact of diseases like Parkinson's disease and stroke on the functional ability of patients. This has laid the foundation for future research efforts that will be directed towards detecting and quantifying clinically meaningful information like detection of emergency events (e.g., falls, seizures) and assessment of symptom severity (e.g., gait impairments, tremor) in patients with various mobility limiting conditions. The proposed method is mainly intended for offline processing of video recording. To provide real-time detection of serious events (e.g., falls), future research efforts should focus on enabling real-time target tracking and action classification by developing more computationally efficient approaches. In addition, achieving high-resolution temporal localization of actions will be necessary to ensure accurate assessment of clinical events of interest (e.g., duration of a seizure) for certain medical applications. Lastly, the code and models developed during this work are being made available for the benefit of the broader research community (https://github.com/brezaei/PoseTrack_ActionClassification).

Author Contributions: Conceptualization: B.R., S.O. and S.P.; data curation: B.R., Y.C., B.H., K.T. and K.E.; formal analysis: B.R.; investigation, B.H. and K.E.; methodology: B.R., S.O. and S.P.; project administration: K.E.; resources: B.H.; software: B.R.; supervision: S.O. and S.P.; writing—original draft: B.R. and S.P.; writing—review and editing: B.R., Y.C., B.H., K.T., K.E., S.O. and S.P.

Funding: Funding for this work was provided by Pfizer, Inc.

Acknowledgments: The authors would like to acknowledge the BlueSky project team for generating the data that made this work possible. Specifically, we would like to acknowledge Hao Zhang, Steve Amato, Vesper Ramos, Paul Wacnik and Tairmae Kangarloo for their contributions to study design and data collection.

Conflicts of Interest: S.P., Y.C. and M.K.E. are employees of Pfizer, Inc. The remaining authors declare no conflict of interest.

References

1. Post, B.; Merkus, M.P.; de Bie, R.M.; de Haan, R.J.; Speelman, J.D. Unified Parkinson's disease rating scale motor examination: Are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable? *Mov. Disord. Off. J. Mov. Disord. Soc.* **2005**, *20*, 1577–1584. [CrossRef] [PubMed]
2. Espay, A.J.; Bonato, P.; Nahab, F.B.; Maetzler, W.; Dean, J.M.; Klucken, J.; Eskofier, B.M.; Merola, A.; Horak, F.; Lang, A.E.; et al. Movement Disorders Society Task Force on Technology. Technology in Parkinson's disease: Challenges and opportunities. *Mov. Disord.* **2016**, *31*, 1272–1282. [CrossRef] [PubMed]
3. Thorp, J.E.; Adamczyk, P.G.; Ploeg, H.L.; Pickett, K.A. Monitoring Motor Symptoms During Activities of Daily Living in Individuals With Parkinson's Disease. *Front. Neurol.* **2018**, *9*, 1036. [CrossRef] [PubMed]
4. Lara, O.D.; Labrador, M.A. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 1192–1209. [CrossRef]
5. van Nimwegen, M.; Speelman, A.D.; Hofman-van Rossum, E.J.M.; Overeem, S.; Deeg, D.J.H.; Borm, G.F.; van der Horst, M.H.L.; Bloem, B.R.; Munneke, M. Physical inactivity in Parkinson's disease. *J. Neurol.* **2011**, *258*, 2214–2221. [CrossRef] [PubMed]
6. Chaaraoui, A.A.; Climent-Pérez, P.; Flórez-Revuelta, F. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Syst. Appl.* **2012**, *39*, 10873–10888. [CrossRef]
7. Vrigkas, M.; Nikou, C.; Kakadiaris, I.A. A review of human activity recognition methods. *Front. Robot. AI* **2015**, *2*, 28. [CrossRef]
8. Chen, Y.; Yu, L.; Ota, K.; Dong, M. Robust Activity Recognition for Aging Society. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1754–1764. [CrossRef]
9. Li, M.H.; Mestre, T.A.; Fox, S.H.; Taati, B. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *J. Neuroeng. Rehabil.* **2018**, *15*, 97. [CrossRef]

10. Brattoli, B.; Buchler, U.; Wahl, A.S.; Schwab, M.E.; Ommer, B. LSTM Self-Supervision for Detailed Behavior Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6466–6475.
11. Song, S.; Shen, L.; Valstar, M. Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 158–165.
12. Schmitt, F.; Bieg, H.J.; Herman, M.; Rothkopf, C.A. I see what you see: Inferring sensor and policy models of human real-world motor behavior. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
13. Chen, A.T.; Biglari-Abhari, M.; Wang, K.I. Trusting the Computer in Computer Vision: A Privacy-Affirming Framework. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1360–1367.
14. Rezaei, B.; Ostadabbas, S. Background Subtraction via Fast Robust Matrix Completion. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy, 22–29 October 2017; pp. 1871–1879.
15. Rezaei, B.; Huang, X.; Yee, J.R.; Ostadabbas, S. Long-term non-contact tracking of caged rodents. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 1952–1956.
16. Rezaei, B.; Ostadabbas, S. Moving Object Detection through Robust Matrix Completion Augmented with Objectness. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 1313–1323, doi:10.1109/JSTSP.2018.2869111. [CrossRef]
17. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
18. Dawar, N.; Ostadabbas, S.; Kehtarnavaz, N. Data Augmentation in Deep Learning-Based Fusion of Depth and Inertial Sensing for Action Recognition. *IEEE Sens. Lett.* **2018**, *3*, 1–4. [CrossRef]
19. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–21 June 2019; pp. 244–253.
20. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2019**, *19*, 1005. [CrossRef] [PubMed]
21. Li, N.; Huang, J.; Li, T.; Guo, H.; Li, G. Detecting action tubes via spatial action estimation and temporal path inference. *Neurocomputing* **2018**, *311*, 65–77. [CrossRef]
22. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, ON, Canada, 8–13 December 2014; pp. 568–576.
23. Zhou, Y.; Sun, X.; Zha, Z.J.; Zeng, W. MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 449–458.
24. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6450–6459.
25. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
26. Liu, M.; Yuan, J. Recognizing Human Actions as the Evolution of Pose Estimation Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
27. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. PoTion: Pose MoTion Representation for Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
28. Cherian, A.; Sra, S.; Gould, S.; Hartley, R. Non-Linear Temporal Subspace Representations for Activity Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2197–2206.
29. Zolfaghari, M.; Oliveira, G.L.; Sedaghat, N.; Brox, T. Chained Multi-stream Networks Exploiting Pose, Motion, and Appearance for Action Classification and Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

30. Girdhar, R.; Gkioxari, G.; Torresani, L.; Paluri, M.; Tran, D. Detect-and-Track: Efficient Pose Estimation in Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 350–359.
31. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
32. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014, pp. 740–755.
34. Andriluka, M.; Iqbal, U.; Milan, A.; Insafutdinov, E.; Pishchulin, L.; Gall, J.; Schiele, B. Posetrack: A benchmark for human pose estimation and tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5167–5176.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June 26–1 July 2016; pp. 770–778.
36. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
37. Gou, M.; Wu, Z.; Rates-Borras, A.; Camps, O.; Radke, R.J.; others. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 523–536.
38. Gou, M.; Camps, O.; Sznaier, M. Mom: Mean of moments feature for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1294–1303.
39. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
40. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
41. Li, M.; Zhu, X.; Gong, S. Unsupervised person re-identification by deep learning tracklet association. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 737–753.
42. Lv, J.; Chen, W.; Li, Q.; Yang, C. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7948–7956.
43. Pirsiavash, H.; Ramanan, D.; Fowlkes, C.C. Globally-optimal greedy algorithms for tracking a variable number of objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Springs, CO, USA, 20–25 June 2011; pp. 1201–1208.
44. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist.* **2005**, *52*, 7–21. [CrossRef]
45. Erb, K.; Daneault, J.; Amato, S.; Bergethon, P.; Demanuele, C.; Kangarloo, T.; Patel, S.; Ramos, V.; Volfson, D.; Wacnik, P.; et al. The BlueSky Project: Monitoring motor and non-motor characteristics of people with Parkinson's disease in the laboratory, a simulated apartment, and home and community settings. In Proceedings of the 2018 International Congress, Hong Kong, China, 5–9 October 2018; Volume 33, p. 1990.
46. Goetz, C.G.; Tilley, B.C.; Shaftman, S.R.; Stebbins, G.T.; Fahn, S.; Martinez-Martin, P.; Poewe, W.; Sampaio, C.; Stern, M.B.; Dodel, R.; et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord. Off. J. Mov. Disord. Soc.* **2008**, *23*, 2129–2170. [CrossRef] [PubMed]
47. Brooks, C.; Eden, G.; Chang, A.; Demanuele, C.; Kelley Erb, M.; Shaafi Kabiri, N.; Moss, M.; Bhangu, J.; Thomas, K. Quantification of discrete behavioral components of the MDS-UPDRS. *J. Clin. Neurosci.* **2019**, *61*, 174–179. [CrossRef] [PubMed]
48. Barrouillet, P.; Bernardin, S.; Camos, V. Time constraints and resource sharing in adults' working memory spans. *J. Exp. Psychol. Gen.* **2004**, *133*, 83. [CrossRef]

49. Insel, T.R. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA* **2017**, *318*, 1215–1216. [CrossRef] [PubMed]
50. Arigo, D.; Jake-Schoffman, D.E.; Wolin, K.; Beckjord, E.; Hekler, E.B.; Pagoto, S.L. The history and future of digital health in the field of behavioral medicine. *J. Behav. Med.* **2019**, *42*, 67–83. [CrossRef] [PubMed]
51. Attal, F.; Mohammed, S.; Dedabrishvili, M.; Chamroukhi, F.; Oukhellou, L.; Amirat, Y. Physical Human Activity Recognition Using Wearable Sensors. *Sensors* **2015**, *15*, 31314–31338. [CrossRef] [PubMed]