The Fourth IEEE RAS/EMBS International Conference
on Biomedical Robotics and Biomechatronics
Roma, Italy. June 24-27, 2012

Vision-Based Posture Assessment to Detect and Categorize Compensation During Robotic Rehabilitation Therapy

Babak Taati, Rosalie Wang, Rajibul Huq, Jasper Snoek, and Alex Mihailidis

***Abstract*— A vision-based posture assessment system for real-time monitoring of upper-limb robotic rehabilitation therapy is developed. The system is capable of automatically detecting and categorizing compensatory movements during robotic exercises and could be used in prompting the patient into the correct pose. A consumer depth camera and skeleton tracking algorithms were used to track the pose of the patient in real-time, and to extract a set of discriminating features which correlated with various posture modes. A multi-class classifier capable of incorporating temporal dynamics was trained to identify and categorize the most common types of compensation at high accuracy (86% per frame). A simple multi-stage active learning strategy was used to minimize the amount of manual annotation needed in providing the classifier with training data.**

I. INTRODUCTION

A. Motivation

According to the World Health Organization, about 15 million people suffer stroke worldwide each year. It has been estimated that up to 65% of stroke survivors have difficulty using their upper limbs in daily activities [1]. To maximize its benefit, post-stroke rehabilitation should start as early as possible, and be intensive and individualized for the stroke survivor. However, such therapy requires therapists to spend a significant amount of time guiding stroke survivors through repetitive exercises and functional activities. The application of rehabilitation robots may alleviate some of the burden on therapists and increase patient access to rehabilitation [2]. A prototype intelligent haptic robotic rehabilitation system has been developed to augment upper-limb mobility treatment and to reduce the amount of supervision required by a therapist during treatment sessions. The work thus far includes hardware and mechatronics design of the haptic interface, the design of intelligent and adaptive software to control the robot, and preliminary clinical evaluation [3], [4].

The work presented in this paper tackles a new aspect of the problem and focuses on monitoring and correcting the pose of a post-stroke patient as they perform rehabilitation exercises using the robot. The upper body and the upper-limbs contain a large number of degrees of freedom (DOF). In robotics terminology, the upper body is a highly redundant structure. For instance, moving an end-effector hand hold of a robot in a specific location, i.e. reaching to a point with the hand, is often possible through an infinite number of upper-limb configurations and trajectories. Compared to healthy individuals, when reaching for a point in space to grab/push/hold something, stroke survivors with mobility impairments tend to accommodate the same goal through a different motion profile by taking advantage of unaffected DOFs to overcome the limitations caused by the affected DOFs [5]. From a purely functional point of view, obtaining the goal in any manner, even via additional compensatory DOFs, may be desirable as it allows the person to function as independently as possible. However, relying excessively on compensation prohibits progress and might contribute to settling into undesirable limb movement synergies, joint contractures, or inefficiencies in limb use [6], [7].

The concern for detecting compensatory movements was identified during preliminary therapist evaluation of the prototype and the follow up therapist focus group and individual feedback sessions [3]. To ensure that the system enables stroke survivors to complete their exercises without ongoing therapist supervision, a tool is necessary to detect compensatory upper-limb positions. Feedback (e.g. auditory or visual) can then be delivered through the system interface to the stroke survivor to correct his or her body position.

The work presented here takes advantage of state-of-the art computer vision hardware (depth sensors) and algorithms (skeleton tracking) in order to develop and experimentally validate a posture assessment technology that does not require placing additional markers or sensors on or around the person's body and is effective in identifying compensation while using the robot.

B. Previous Work

Much of the previous work on detecting posture in rehabilitation has relied on wearable sensors, such as accelerometers [8], sensing garments [9], torso harnesses [10], or other sensors attached to the patient for ambulatory detection and monitoring of gross body changes in posture. Other approaches, have used sensors placed around the patient. These include photo resistors placed on the back of the chair to detect trunk motion compensation while performing reaching exercises [4]. Such sensor-based approaches often require careful engineering and a design specific to the setup of the robotic device at hand. Torso harnesses may also be uncomfortable to wear.

Recent advancements in the field of computer vision and real-time tracking offer an opportunity to develop posture monitoring systems that do not require placing additional sensors on or around the body of the patient [11]. Moreover, vision-based systems allow for posture categorization via

This work was supported by the NSERC-CIHR Collaborative Health Research Program, the MITACS NCE Strategic Project, and Quanser Consulting Inc.

The authors are with the Intelligent Assistive Technology and Systems Lab (IATSL), University of Toronto, and affiliated with the Toronto Rehabilitation Institute, Toronto, Ontario, Canada

978-1-4577-1200-5/12/$26.00 ©2012 IEEE

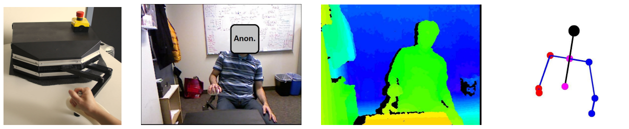
software processing. That is, whereas the number of sensors (i.e. hardware) required to detect and identify various forms of compensation increases with the number of posture or compensation categories, computer vision systems employ a single sensor (a camera or a depth sensor) and use artificial intelligence techniques to analyze the posture in the live video feed. State-of-the-art computer vision and machine learning algorithms enable such processing in real-time. Since the adjustment of software parameters is easier than changing the hardware equipments and setup, the use of computer vision allows for the development of more generic systems which could be adapted to new robotic devices with less effort.

The literature on vision-based human pose estimation and skeleton tracking is vast, and its reviewing is beyond the scope of this article. Furthermore, the work presented here is built on top of an existing pose estimator and applies skeleton tracking to posture classification. Previous work on gait and posture assessment for clinical or health-related applications include video-based analysis of the standing balance [12] among older adults to identify the statistics with significant correlations with the risk of falling; and real-time video analysis in images captured by a ceiling mounted camera with a wide-angle lens for automatic fall detection and emergency response [13]. A review of markerless motion capture algorithms for biomechanical applications is presented in [11].

C. Contributions

The goal of the present study was to explore the possibility of using vision-based monitoring for automatic detection of compensatory movements during robotic upper-limb rehabilitation. This was conducted by (i) developing real-time vision-based pose assessment algorithms for detecting and categorizing the most common types of compensatory movement observed in stroke survivors with upper-limb disability and (ii) evaluating the prediction outcomes of the system against their true labels. A consumer depth sensor (a Microsoft Kinect) was used to capture live video and depth streams and track major skeletal joints as each human subject performed repetitive rehabilitation exercises using the robot. A simple active learning strategy was used (and empirically validated) in order to minimize the need for manual annotation and a multi-class classifier that exploits temporal dependencies and considers the inherent ambiguity in true posture labels during transition periods was used in categorizing compensation. While the release of the Kinect sensor has generated enthusiasm for its application in rehabilitation (e.g. [14], [15]), to the best of the authors' knowledge, this is the first study that applied the capabilities of the sensor to the real-time detection of compensatory postures during robotic rehabilitation therapy.

The rest of this paper is organized as follows: §II explains the experimental setup and §III explains the dataset used in the experiments. §IV provides an overview of the developed system and algorithms and the experimental results are reported in §V. §VI concludes the paper with the take home message and future work.



(a) The Robot (b) Color Image (c) Depth Image (d) Skeleton

Fig. 1. The Robot and Sample Captured Data by the Kinect.

II. EXPERIMENTAL SETUP

The robot used here is a 2-DOF planar haptic interface, developed in collaboration with Quanser Inc, Fig. 1(a). The manipulator is specifically designed for intelligent and adaptive post-stroke rehabilitation exercises. More information about the mechatronics of the device can be found in [3]. The robot is placed on a tabletop, with the end-effector extending out and the subject interacts with the robot while sitting in front of it, Fig. 1(b). A Microsoft Kinect sensor was placed at roughly 90 cm behind the robot and 60 cm above the tabletop, facing the subjects as they interacted with the robot. The Kinect captures color and depth image frames in VGA resolution at 30 frames per second, Figs. 1(b)-1(c), and its accompanying 3-D skeleton tracking library [16] processes the depth information and allows for the tracking of the major body joints of a person in real-time, Fig. 1(d). While it is possible to calibrate the sensor to increase both the depth accuracy and the color-depth registration, the default factory setting calibration proved sufficient for the purpose of posture categorization and was used here.

The sensor was mounted on a small tripod and was placed on the tabletop, behind the robot, without any particular attempt at securing it against shakes or vibrations. The exact placement and height of the tripod varied slightly across the experiments. This, compounded by the large height variations among the subjects, resulted in some variations in the relative position and orientation of the sensor with respect to the subject (e.g. ∼15-20 cm and 10-15°). These variations in the viewing angle, as well and the shakes and vibrations represented realistic scenarios. It was decided that rather than trying to secure a more rigid mount for the sensor and to fix its position with respect to the robot, which might be difficult to reproduce in clinical settings, the system should be developed in such a way that it would be robust with respect to such deviations. While the experiments (§V) confirmed the robustness of the overall system with respect to the variations in the position of the sensor and shakes, it was noted that the skeleton tracking worked best when the sensor was placed at a slightly higher height than the subject's head and looked down from the above at a small angle, e.g. ∼ 10°. This prevented the arms and the hands from being partially occluded by the robot.

III. DATASET

Seven healthy adult subjects without mobility impairments were recruited and asked to simulate a series of compensations during short sequences of interaction with the robotic

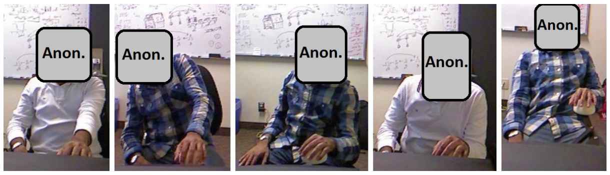


Fig. 2. (left to right) Sample Frames in *No-Comp*, *Shoulder-Hike*, *Trunk-Rotation*, *Lean-Forward*, and *Slouch* Postures.

device in order to record a dataset for evaluating the proposed methods in vision-based detection of compensation during robotic rehabilitation therapy. The interaction involved sitting in front of the robot, holding the end-effector with the right hand, and moving it forward and backward five times. The subjects were asked to move the robot back and forth in a straight line on the sagittal plane (as in [17]), but their performance was not judged or annotated based on the straightness of their motion path, or based on the traversal speed. To evaluate the developed algorithms for pose assessment during subject-robot interaction, assistive and resistive haptic forces of the robot were kept off during the experiment. Joint frictions were minimal and the 2-D motion of the robot limited to the horizontal plane was not affected by gravity. The subjects were therefore able to move the end-effector with a relatively small force, not significantly larger than that of when moving their arm in free space. Compensation movements and poor postures were limited to a subset of some of the most common types (according to [5], [6]), namely hiking the shoulder to ease the extension of the elbow joint, rotating the trunk or leaning forward to take advantage of the stronger body muscles when reaching forward, or slouching. Each subject was recorded simulating each of these four scenarios separately and also once interacting with the robot without any compensations, amounting to a total of five videos per subject or 35 videos (23,782 captured frames) in total. Given the set of three simulated compensations, the one simulated poor posture (slouch), and also the videos with correct posture, the set of possible frame labels was enumerated as $L = \{$*No-Comp, Shoulder-Hike, Trunk-Rotation, Lean-Forward, Slouch*$\}$. Throughout the rest of the paper, the term *posture mode* is used to refer to the five aforementioned categories and the term *compensation mode* is liberally used to refer any of the three compensation postures or slouching. Fig. 2 illustrates a sample frame from each of the five categories.

IV. ALGORITHMS

The NITE skeleton tracking library [16] was used to track the main upper body skeletal points at each frame. These included the 3-D position of the *head*, *neck*, *torso*, and the left and right *shoulder*, *elbow* and *hand* points at each frame, all expressed in the coordinate system of the depth camera. Because of the variations in the sensor mount pose, a coordinate change was required in order to express the skeletal points in a canonical frame, independent of the relative pose between the subject and the depth sensor. To do so, at the beginning of each sequence, the subject was asked to sit still in front of the robot for a short instance. A frame during this period was taken as the *rest pose* and was used to define a subject-centered coordinate frame as follows: The unit vector $\vec{j}$ along the line segment connecting the neck to the head point defined the up and down coordinate, i.e. the y axis in the canonical frame. The component of the 3-D vector connecting the left shoulder to the right shoulder which was perpendicular to $\vec{j}$ produced the unit vector along the x axis, or $\vec{i}$, indicating the left and right direction. The constraints of an orthonormal frame determined the z axis and its unit vector $\vec{k}$ along the back and forth direction.

In order to categorize the subjects' posture at a given frame, a multi-class classifier trained on labeled data operated on a collection of features extracted from the position of the skeletal joints at each frame. The following sub-sections explain the details of extracting the features, obtaining the frame labels for training, and the choice of classifiers.

A. Features

The features consisted of the 3-D orientation of line segments connecting relevant skeletal points, expressed as Euler angles. As the subjects interacted with the robot with their right arm (targeting right-sided hemiparesis), the movements of the left arm had no bearing on the correct posture and were thus discarded. The segments used here include the *head-neck*, *neck-torso*, *left-shoulder-right-shoulder*, *neck-right-shoulder*, *right-shoulder-right-elbow*, and *right-elbow-right-hand* segments.

It is possible to extend the feature set beyond the orientation of the skeletal segments, e.g. to include motion or raw positional information. However, preliminary experiments in augmenting the angle features with raw positional or velocity information did not result in any significant classification improvements overall. When dealing with raw skeletal positions, the coordinates were normalized to keep the shoulder width, i.e. the distance between the left and right shoulder joints, at unit length, in order to eliminate the effect of variations in body size. Similarly, additional motion features representing the dominant modes of motion over a short period of time, e.g. through using Principal Components Analysis (PCA) over concatenated velocity vectors, did not significantly improve the overall posture categorization results. For brevity, only the experimental results with the orientation features are reported here.

B. Multi-Class Classifiers to Identify Compensation Modes

Once the features are computed at a given frame, a multi-class classifier is used to categorize the posture of the subject into one of five classes, $l \in L$. The experiments were performed using three different linear kernel Support Vector Machine (SVM) classifiers. These included a structured SVM for simple frame-by-frame multi-class classification (SVM^{multiclass}), and two classifiers that considered the temporal dependency of subsequent frames, namely the Hidden-Markov SVM (HM-SVM) and a modified version of it (HM-SVM$^{\circ}$) which accounts for the ambiguity of the training labels near the transition boundaries [18].

The SVM^{multiclass} classifier is instantaneous, i.e. it produces an estimate for a frame label solely based on the features extracted from that frame and independent of the previous estimates. That is, the estimated frame label l_i^v, for the i^{th} frame of a given video v, is determined solely via the feature vector of the frame, $\vec{f_i^v}$. The other two HM-based classifiers take into account the temporal context prior to the current frame, i.e. the predicted labels of the adjacent frames, via a Hidden Markov Model (HMM) assumption. Here the Markov states are the five compensation labels (L) and transitioning from one state to another is equivalent to changing the posture mode, e.g. from *No-Comp* to *Lean-Forward*.

Since each of the sequences in the collected dataset (§III) simulates at most one compensation mode, the dataset contains no instances of switching directly from one compensatory mode to another without transiting through a *No-Comp* phase. In other words, whereas the dataset contains frequent transitions to and from any of the four poor postures to a correct pose, it contains no example of a direct passage from one poor posture to another. Training an HMM-based classifier based on such training sequences thus limits the set of possible state transitions to *No-Comp→X* and *X→No-Comp*, where *X* represents any of the four compensation modes. At first sight, this might appear as an unrealistic assumption and a shortcoming in the collected dataset. However, the system will be used to prompt the subject into the correct pose as soon as it detects any deviations. Therefore, as long as the subject conforms to the given prompts, it will be highly unlikely that the subject will shift from one poor posture, e.g. *Trunk-Rotation*, directly to another, e.g. *Slouched*.

In labeling the posture modes along a sequence, it is often clear when the subject is fully in the correct pose vs. when they are significantly compensating or are in a severely poor posture. The labels however can be ambiguous near the transitions between the successive posture modes. Even a human labeler is generally uncertain at exactly which frame one mode, e.g. *No-Comp*, ends and another, e.g. *Slouch* begins. Such ambiguities in the training labels can be misleading for the standard formulation of the HM-SVM classifier. The modified version considers this inherent ambiguity and assigns lower weights to the penalty term computed for mislabeling the frames close to transition boundaries [18]. The objective of the optimization algorithm used during the training phase is to minimize the cumulative sum of the penalty terms over the lengths of the training sequences. Modifying the objective, or the loss function in the SVM terminology, in this manner therefore reduces the effect of the ambiguous labels without sacrificing the computational efficiency of the training phase. It has been shown that this modification improves the overall classification performance and reduces the Dynamic Time Warping (DTW) [19] distance between the sequence of estimated labels and the sequence of ground truth labels.

C. Training Labels

Training a classifier requires access to a set of labeled data; in this case complete per-frame compensation labels in a series of training videos. Manual annotation of every single frame throughout multiple videos is time consuming. A simple 3-stage active learning strategy was used here to provide per-frame labels offline for all the training videos with the minimal amount of manual annotation.

In the first stage, for a given video (v) consisting of N^v frames, the feature vectors of all the frames ($\vec{f_1^v} \ \ldots \vec{f_{N^v}^v}$) are clustered into a small number (n_{c1}) of representative poses. A human manual annotator (a *rater*) was then asked to manually mark the true posture label for a number (n_{c1}) of frames in the video which were the closest to the cluster centers in the feature space. The rater assigned one of the five labels in L or, through a user-friendly GUI, passed the ambiguous frames for which they could not easily identify the posture. The output of the first stage was thus a sparse set of manual labels (l_{m1}) for only a very small subset of frames in each video. The clustering ensures that manual labels were collected for representative frames that span the variety of poses in each video as much as possible.

The second stage identified a few additional frames in each video that most required manual annotations and prompted the rater for marking them. Since each video simulated a single compensation strategy, it contained at most two labels throughout its playback. (*No-Comp* videos contained a single label.) Given the handful of manual labels (l_{m1}) in each video, it was therefore possible to train a binary classifier that could estimate the labels for the remaining frames. The frames in which the classifier was least confident indicated areas where further manual annotation could best improve the results. A simple linear classifier (logistic regression) was used here as it was both easy to train and use and it associated confidence levels on its estimates. Following the linear binary classification, the features of the video were once more clustered into a number (n_{c2}) of groups ($n_{c2} > n_{c1}$) and the top n_u clusters which showed the highest uncertainty were once more annotated manually, as in the first stage, to provide an augmented set of sparse manual labels, l_{m2} ($l_{m1} \subset l_{m2}$). Here the k-means algorithm [20] was used for clustering, and the parameter values were set empirically at $n_{c1} = 10$, $n_{c2} = 100$, and $n_u = 20$. The second round of clustering ensured that the augmented set of manual labels were dispersed along the length of each video and did not request for redundant manual annotations. Simply sorting all the frames which obtained the lowest confidence in the binary classification was likely to return adjacent frames that all belonged to a single cluster.

The final stage consisted of training a binary classifier given the augmented set of manual labels (l_{m2}) and estimating a label for all the frames throughout each video $l_c(t)$. A classic binary Support Vector Machine (SVM) classifier with a linear kernel was used here as confidence values were no longer required and the SVM outperformed a simple linear classifier without regularization. These complete labels were

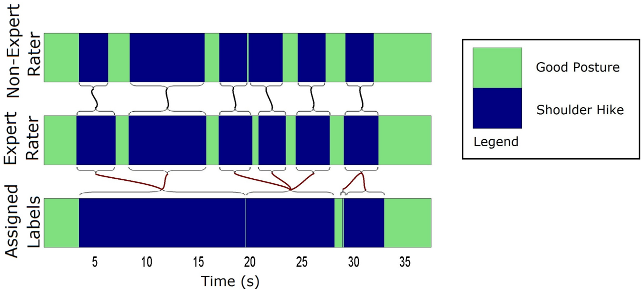


Fig. 3. Comparing the Assigned Labels with the Human Raters.

empirically evaluated against manual annotations (§V-A).

The output of each of the aforementioned stages, combined across a series of videos, provides a possible training set for training the multi-class classifier used in categorizing compensation postures. It is natural to expect that the augmented manual annotations (l_{m2}) contains a richer training set over the original manually annotated set (l_{m1}) and would train better performing classifiers for categorizing postures. It is also natural to expect that the complete annotations (l_c) outputted by the third stage would be richer yet and would result in the best performance overall. These hypotheses were confirmed experimentally (§V-B) using each of the three multi-class classifiers discussed in §IV-B.

To avoid a possible confusion between the assigned labels via the three-stage labeling process (l_{m1}, l_{m2}, and l_c) used to train the multi-class classifiers, and the estimated labels provided by either of the three multi-class classifiers, the former are referred to as the training or the *assigned* labels and the latter are referred to as the *estimated* labels throughout the remainder of the text.

V. EXPERIMENTAL RESULTS AND DISCUSSION

A. Assigned Labels

The *No-Comp* label was assigned to all the frames within the seven videos, one of each subject, in which the subject operated the robot in the correct pose and without simulating any poor posture or compensation. The three stage process described in §IV-C was applied to all of the remaining $(4 \times 7 =)$28 videos and significantly reduced the burden of assigning labels for the purpose of training. In total, $(28 \times (10 + 20) =)$ 840 frames were examined visually during the first two stages, resulting in a manually assigned label or a pass, amounting to ∼4.3% of the total frame count across all the 28 simulation videos.

In order to evaluate the assigned per-frame labels obtained via the third stage, complete frame-by-frame manual annotation was collected for all four simulated compensation videos of one of the subjects, chosen at random. The *No-Comp* video was excluded as it did not require any manual annotation. These four videos were each fully annotated twice, once by a trained occupational therapist (expert rater) and once by another person (non-expert), denoted as $l_{gt}(t)$ and $l_{\hat{gt}}(t)$ respectively.

The Cohen's kappa coefficient [21], a statistical measure of inter-rater agreement, was $\kappa = 0.88$, indicating a very good agreement between the two raters across all the four videos. All of the deviations occurred in frames near the transition boundaries, in line with the premise that inherent ambiguities exist near the state transition boundaries even for human raters (as noted in [18]). That is, both the expert and non-expert raters annotated all of the individual incidents of compensation or poor posture (17 in total) identically, but deviated from one another only in marking the exact time at which each incident started or ended. For instance, in one case, when the subject was in the correct pose and then started to hike her shoulder, the expert rater marked that transition as starting at five frames earlier than when the non-expert rater indicated. The average absolute value of the time difference was less than a third of a second (0.30 s) across all 34 cases (17 × 2, referring to the beginning and end of each incident).

The non-expert rater showed more leniency when annotating the *Shoulder-Hike* postures by marking several of the ambiguous frames as the correct pose. By contrast, when marking the *Lean-Forward* and *Trunk-Rotation* compensations, the expert rater tended to be slightly more lenient and accepted more of the ambiguous situations as the correct pose. The expert and the non-expert raters produced virtually identical results in marking the *Slouch* posture. The inter-rater agreement puts an upper bound on the accuracy levels to be expected of the final multi-class posture categorization. That is, an automated posture assessment system would operate at human level if it could duplicate the same accuracy level and deviate from the expert labeled ground truth labels only close to the state transition boundaries.

The kappa coefficients quantifying the agreement between the assigned labels ($l_c(t)$) and the expert and non-experts annotations were $\kappa = 0.85$ and $\kappa = 0.88$ respectively, indicating very good agreements. When considering the *Lean-Forward*, the *Slouch*, and the *Trunk-Rotation* videos, the assigned labels concurred with the ground truth expert (as well as non-expert) annotations in detecting all the individual incidents of poor posture and compensation, 11 in total, without reporting any false positives. In these three scenarios, deviations between $l_c(t)$ and $l_gt(t)$ (as well as $l_{\hat{gt}}(t)$) occurred only near the transition boundaries. In the *Shoulder-Hike* video, the human raters identified six intermittent instances of hiking the shoulder (when extending the shoulder to push the robot forward), separated by five short (avg. length 1.1 s) segments of correct posture (when retracting the elbow to pull the end-effector). The assigned labels on the other hand, partitioned the same sequence as containing longer poor-posture segments with fewer discontinuities (Fig. 3). While this was not ideal, it did not raise a serious concern as a second look through the video revealed that the subject indeed had a slightly hiked shoulder during the intermittent "good posture" segments that the human raters identified. In other words, while the raters accepted those short segments as having the correct pose, they were visibly different from the "normal" pose recorded in the *No-Comp* videos. Moreover, since the primary purpose of categorizing postures is to prompt the subject into the correct pose, *grouping* intermittent poor-posture segments

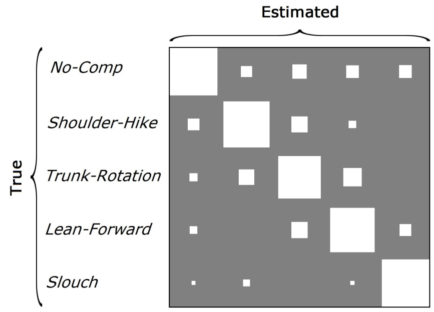


Fig. 4. Normalized Hinton Diagram for HM-SVM° Trained with $l_c(t)$.

into longer ones will not result in any missed prompt so long as all the instances of poor posture are identified.

B. Categorizing Compensations

Leave-one-subject-out cross validation was used to evaluate all three classifiers mentioned in §IV-B (i.e. SVM^{multiclass}, HM-SVM, and HM-SVM°), each trained with the output of the first (l_{m1}), the second (l_{m2}), or the third (l_c) phase of the three-phase approach described in §IV-C. The leave-one-subject-out strategy meant that when testing the categorization performance on each subject, only the videos from the remaining six subjects were used in training the multi-class classifier. This ensured that the classifiers were not trained on the idiosyncrasies of the posture and the movement patterns of a person and instead learned the broader patterns associated with good vs. poor posture. Table I summarizes the overall per-frame accuracy across all the 35 videos and for all the nine cases (3 classifiers, 3 training sets). As expected, richer training data resulted in better overall categorization. Similarly, considering the temporal dynamics improved the accuracy over instantaneous classification. The best performance (accuracy $\approx$ 86%) was acheived with HM-SVM°, trained on the output of the third phase labeling, l_c. While the accuracy levels obtained with HM-SVM and HM-SVM° were similar, as expected, HM-SVM° resulted in lower average DTW distance between true and predicted sequences. With the l_c training data for instance, the average DTW distance via HM-SVM° was 10.1% lower than that obtained with HM-SVM. Fig. 4 illustrates the per-frame confusion matrix, represented as a normalized Hinton diagram, in categorizing the five posture modes. The Cohen's kappa coefficient of $\kappa = 0.80$ indicates a very good agreement between the assigned and the estimated labels.

TABLE I

ACCURACY LEVELS (%) OBTAINED WITH EACH MULTI-CLASS CLASSIFIER AND USING DIFFERENT SETS OF TRAINING DATA.

	Training Data		
Classifier	Phase I (l_{m1})	Phase II (l_{m2})	Phase III (l_c)
SVM^{multiclass}	74.1	78.5	80.0
HM-SVM	78.6	81.0	84.9
HM-SVM°	77.6	80.6	85.9

While the correlation coefficients between the true and estimated labels (or the avg. per-frame accuracy levels) provide useful information on the performance of the classifiers in detecting compensations, for reasons discussed earlier (i.e. the ambiguity near the transition boundaries and the grouping of close by segments), they are not an ideal metric for evaluating temporal segmentation results along a time series. The DTW distance and the modified loss function in HM-SVM° ([18]) offer alternative measures, but even those metrics are not ideal. The modified loss function was designed for computational efficiency and is still a per-frame metric and the DTW distance is susceptible to misaligning the segments. While allowing for some slack is a desirable property near the transition boundaries, in the extreme, the DTW distance could accept completely erroneous predictions as correct so long as the order of the state transitions in the estimated time series remain identical (or close) to that of the true sequence. Adjusting the parameters of the metric (e.g. the relative weight between the penalty for switching a state to another vs. the penalty for delaying a transition) can handle some of these undesirable characteristics, but requires too much manual tuning and the hard coding of carefully selected settings. An ideal metric would provide a basis for evaluating and quantifying the detection and miss rates of individual incidents (i.e. compensation modes) along the time axis rather than aggregating the measures on individual frames or time instances. This would require extending the binary classification metrics such as precision and recall (= true positive), or the multi-class metrics such as the confusion matrix and the various inter-rater agreement statistics, to the classification scenarios when the aim is to detect incidents along a time series.

While the development of such a metric is the subject of future work, for the sake of completeness and in order to provide further insight into the performance of the developed system, some further statistics regarding the hit and miss rates of the individual incidents of compensation are provided here. Table II presents the total number of detected and missed compensation incidents, when compared to the true sequence of segments. Here, a temporal "segment" refers to a contiguous series of frames, i.e. subsequent frames with an identical label, and a compensation "incident" refers to a segment corresponding to a poor or compensatory posture. Missed incidents refer to the cases where a compensation was either not detected or was miscategorized as another compensation mode.

TABLE II

NUMBER OF DETECTED AND MISSED COMPENSATION INCIDENTS.

	True Incidents	
Compensation Mode	Detected	Missed
Shoulder-Hike	10	4
Trunk-Rotation	28	12
Lean-Forward	44	12
Slouch	7	1

Table III presents the number of true positives and the number of false positives in each compensation mode, when compared to the sequence of detected segments. The false positives indicate the cases where the subject was in the correct pose but an incident of compensation was reported. Note that the numbers in the first columns of Tables II and

III differ in three out of the four cases. This is due to the fact that the segments in Table II are divided according to the state transitions in the assigned labels (l_c) while the segments in Table III are based on the estimated labels. These variations correspond to the grouping of the adjacent events, when they are separated by short intervals of good posture, and as discussed earlier.

TABLE III

NUMBER OF TRUE AND FALSE POSITIVES.

	Reported Incidents	
Compensation Mode	True Positive	False Positive
Shoulder-Hike	7	0
Trunk-Rotation	27	11
Lean-Forward	32	4
Slouch	7	0

Tables II and III show that most of the compensation incidents were detected and a low number of false positives were reported. In total, less than 25% of the compensation incidents were missed and false positives constituted less than 15% of the overall reported incidents.

It should be noted that these statistics under represent the performance of the overall system in detecting compensations. As noted in §V-A, grouping two adjacent identical compensation segments separated via short good-posture segments is not nearly as consequential as entirely missing a compensation segment, but the stated statistics report the two scenarios as the same and, in the former case, report the second compensation segment among the missed events. Similarly, detecting a long compensation segment as two subsequent shorter segments separated by a few frames of good posture is not significant but the stated statistics report the second segment among the false positives.

The experimental results and the analysis provided here (per-frame accuracy levels, per-frame confusion matrix, incident detection and miss counts, and incident true and false positive counts) provide ample evidence illustrating the performance of the developed system in detecting compensations during robotic rehabilitation therapy. The stated shortcomings in evaluating temporal classification results however highlight the need for developing a unified and standardized metric for evaluating the detection of incidents along a time series.

VI. CONCLUSIONS AND FUTURE WORK

This work focused on the automatic assessment of posture during robotic rehabilitation therapy. The developed system employed a consumer depth camera and real-time skeleton tracking libraries to build a working prototype capable of identifying compensatory postures in real-time, and with good accuracy and a low false positive rate. A three-phase labeling mechanism was used to minimize the need for manual annotation during the development of the system. The need for the development of a unified metric for evaluating the identification of incidents along a time series was highlighted through the analysis of the experimental results.

Future work will study the effectiveness of providing real-time alert to prompt the subjects into a good posture. Future research will also attempt at correlating the tracked posture data and the detected compensations with recorded clinical assessments over time, in order to monitor the progress (or degradation) in mobility and study the efficacy of posture tracking as a diagnostic tool.

REFERENCES

[1] B. H. Dobkin, "Clinical practice, rehabilitation after stroke," *New England J Medicine, 352*, pp. 1677–1684, 2005.

[2] S. E. Fasoli, H. I. Krebs, and N. Hogan, "Robotic technology and stroke rehabilitation: Translating research into practice," *Topics in Stroke Rehabilitation*, vol. 11, no. 4, pp. 11–19, 2004.

[3] E. Lu, R. Wang, R. Huq, D. Gardner, P. Karam, K. Zabjek, D. Hebert, J. Boger, and A. Mihailidis, "Development of a robotic device for upper limb stroke rehabilitation: A user-centered design approach (to appear)," *Paladyn. J. Behavioral Robotics*, 2012.

[4] P. Kan, R. Huq, J. Hoey, R. Goetschalckx, and A. Mihailidis, "The development of an adaptive upper-limb stroke rehabilitation robotic system," *J. Neuroengineering and Rehabiliation, 8(33)*, 2011.

[5] M. C. Cirstea and M. F. Levin, "Compensatory strategies for reaching in stroke," *Brain: A journal of neurology, 123(5)*, pp. 940–953, 2000.

[6] C. Gowland, P. Stratford, and M. W. et al., "Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment," *Stroke, 24(1)*, pp. 58–63, 1993.

[7] P. Lum, S. Mulroy, R. Amdur, P. Requejo, B. Prilutsky, and A. Dromerick, "Gains in upper extremity function after stroke via recovery or compensation: Potential differential effects on amount of real-world limb use," *Topics in Stroke Rehabilitation, 16(4)*, pp. 237–253, 2009.

[8] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, and P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly," *IEEE Trans Biomedical Engineering, 50(6)*, pp. 711–723, 2003.

[9] A. Tognetti, F. Lorussi, R. Bartalesi, S. Quaglini, M. Tesconi, G. Zupone1, and D. D. Rossi, "Wearable kinesthetic system for capturing and classifying upper limb gesture in post-stroke rehabilitation," *J. Neuroengineering and Rehabiliation, 2(8)*, 2005.

[10] S. Conroy, J. Whitall, L. Dipietro, L. Jones-Lush, M. Zhan, M. Finley, G. Wittenberg, H. K. HI, and C. Bever, "Effect of gravity on robot-assisted motor training after chronic stroke: A randomized trial," *Archives of Physical Medicine and Rehabilitation, 92*, pp. 1754–1761, 2011.

[11] L. Mundermann, S. Corazza, and T. P. Andriacchi, "The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *J. NeuroEngineering and Rehabilitation, 6(3)*, 2006.

[12] S. J. Allin, C. Beach, A. Mitz, and A. Mihailidis, "Video based analysis of standing balance in a community center," *IEEE EMBC*, pp. 4531–4534, 2008.

[13] M. Belshaw, B. Taati, J. Snoek, and A. Mihailidis, "Towards a single sensor passive solution for automated fall detection," *IEEE EMBC*, pp. 1773–1776, 2011.

[14] E. Stone and M. Skubic, "Evaluation of an inexpensive depth camera for in-home gait assessment," *J Ambient Intelligence and Smart Environments*, pp. 349–361, 2011.

[15] E. Auvinet, F. Multon, and J. Meunier, "Gait analysis with multiple depth cameras," *IEEE EMBC*, pp. 6265–6268, 2011.

[16] *Prime Sensor NITE^{TM}1.3 Algorithms Notes*, PrimeSense Inc., 2010.

[17] R. Huq, P. Kan, R. Goetschalckx, D. Hebert, J. Hoey, and A. Mihailidis, "A decision-theoretic approach in the design of an adaptive upper-limb stroke rehabilitation robot," *IEEE Intl Conf Rehab Robotics*, pp. 589–596, 2011.

[18] B. Taati, J. Snoek, and A. Mihailidis, "Towards aging-in-place: Automatic assessment of product usability for older adults with dementia," *IEEE-HISB*, 2011.

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Tran Acoustics, Speech and Signal Processing, 26(1)*, pp. 43–49, 1978.

[20] G. A. F. Seber, *Multivariate Observations*. Hoboken, NJ: John Wiley & Sons Inc., 1984.

[21] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics, 33(1)*, pp. 159–174, 1977.