IET Journals
The Institution of Engineering and Technology

A Novel 3D Motion Capture System For Assessing Patient Motion During Fugl-Meyer Stroke Rehabilitation Testing

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

N. Eichler[1,2] *H. Hel-Or*[1] *I. Shimshoni*[2] *D. Itah*[3] *B. Gross*[4,5] *S. Raz*[2]

[1] *Department of Computer Science, University of Haifa, Haifa, Israel*
[2] *Department of Information Systems, University of Haifa, Haifa, Israel*
[3] *Occupational Therapy Unit, Galilee Medical Center, Naharia, Israel*
[4] *Neurology Department, Galilee Medical Center, Naharia, Israel*
[5] *Azrieli school of Medicine, Bar Ilan University, Israel*
* *E-mail: eichler@outlook.com*

Abstract: We introduce a novel marker-less multi-camera setup that allows easy synchronization between 3D cameras as well as a novel pose estimation method that is calculated on the fly based on the human body being tracked, and thus requires no calibration session nor special calibration equipment. We show high accuracy in both calibration and data merging and is on par with equipment-based calibration. We deduce several insights and practical guidelines for the camera setup and for the preferred data merging methods. Finally, we present a test case that computerizes the Fugl-Meyer stroke rehabilitation protocol using our multi-sensor capture system. We conducted a Helsinki-approved research in a hospital in which we collected data of stroke patients and healthy subjects using our multi-camera system. Spatio-temporal features were extracted from the acquired data and Machine Learning based evaluations were applied. Results showed that patients and healthy subjects can be correctly classified at a rate of above 90%. Furthermore, we show that the most significant features in the classification are strongly correlated with the Fugl-Meyer guidelines. This demonstrates the feasibility of a low-cost, flexible and non-invasive motion capture system that can potentially be operated in a home setting.

1 Introduction

As part of a project on motor rehabilitation of stroke patients, we aim to automate the Fugl-Meyer stroke rehabilitation assessment protocol [1] (Figure 14). This involves capturing and tracking the motion of patients and later analysing the data as part of the diagnostic process. Thus, we require a non-invasive human motion capture system that is inexpensive, portable, (allowing use both in the hospital and at home during extended rehabilitation) and is very easy to use (even by the patient).

3D cameras (depth cameras) form a good basis for such a system, as they provide a depth map, namely, a depth value per pixel indicating the distance of the scene point from the camera. The resulting depth map is often referred to as 2.5D data. 3D cameras are primarily used for gaming devices and thus typically supply some form of Body Tracking (Human Motion-Capture). Typically, a representation in the form of a Body-Skeleton is used (Figure 1), which is calculated from the depth map captured by the camera sensors.

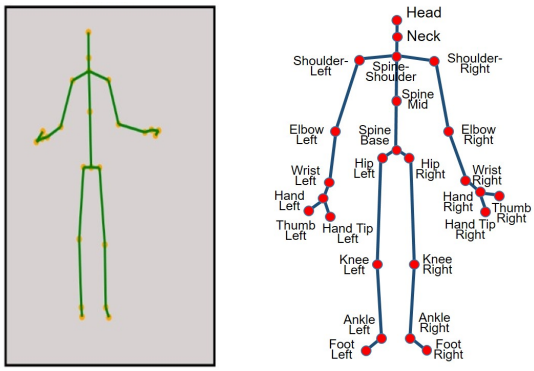

Fig. 1: Micorosft Kinect Skeleton (left) and Joint Map (right).

Several high-end motion capture systems are available (e.g. VICON [2], Optotrak[3], Ipi Soft[4]) that provide the state-of-the-art in human body tracking performance. However, these advanced technologies have several significant limitations:

- Very expensive
- Non-Portable
- Invasive - requires subjects to be marked with reflective markers.
- Complex to use, often requiring a professional trained operator.
- Requires system calibration typically in the form of a recording session using special dedicated accessories.
- To obtain a skeleton representation, requires additional manual measurements of subjects' body parts as part of the calibration.
- Often requires a large setup area, thus is inappropriate for home settings.

Our goal is to develop a human motion capture system that is inexpensive, portable, marker-less (non-invasive), requires no calibration equipment and performs at accuracy rates on par with the high-end advanced motion tracking systems. The developed system is targeted for easy home use and tele-medicine applications, but can also serve in other applications including security, object tracking, and more.

To accomplish the task, we chose to use inexpensive and portable consumer 3D cameras - namely the Kinect V2 [5]. These cameras supply a video stream of 2.5D data in the form of 3D point clouds, in addition to RGB data. Additionally, the system supplies a skeleton representation of the subject computed per frame from the point cloud [6, 7]. However, due to the low-cost requirement in these consumer cameras, the accuracy and robustness of the captured data and of the resulting skeleton is left to be desired [8–10]. These cameras have known issues inherent to the technology - they assume the subject is mostly in frontal pose to the camera plane in order to be accurately tracked by the sensor. Violation of this constraint results in "Low-Confidence", unreliable or complete failure of skeleton as well as erroneous or incomplete point clouds, as shown in Figure 2. To overcome these issues, and to improve accuracy and robustness

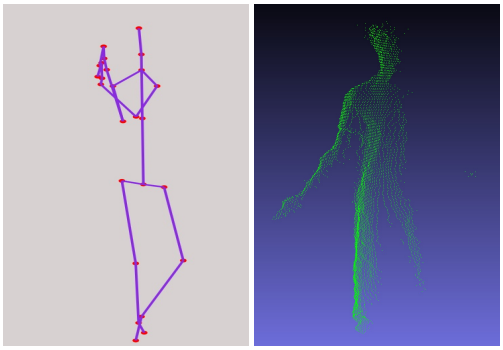

Fig. 2: Kinect acquisition failure examples.

of the data, we suggest using a multiple camera setup. The data acquired from different viewpoints can achieve greater coverage of the captured scene and can assist in error minimization. Furthermore, registration from several views can transform the data from the 2.5D representation into full 3D (see Figure 9).

In this paper we introduce a novel, inexpensive, portable and marker-less multi-camera setup that allows easy temporal synchronization between cameras. Most importantly we introduce an accurate pose estimation method that is calculated on the fly based on the human body being tracked, and thus requires no calibration session nor special calibration equipment. This is in stark contrast with other camera systems which require a dedicated session using specialized accessories such as a glowing marker [4], wand [2] or checkerboard [11] (Figure 3).

The proposed system performs run time merging of data, resulting in a more reliable and stable skeleton representation and as an added bonus, allows alignment and merging of the 3D point clouds to form a full 3D body representation.

As a test case, we applied the proposed system as part of our medical application for automating the Fugl-Meyer stroke rehabilitation assessment protocol (Figure 14) [1]. The Fugl-Meyer Assessment (FMA) is a stroke-specific, performance based impairment index. It is designed to assess motor functioning, balance, sensation and joint functioning in patients with post-stroke hemiplegia (weakness of one entire side of the body). Our system uses 2 cameras (Figure 4), ensuring each side of the patient's body is properly viewed, and produces a reliable and error free skeleton to be used in diagnosis. We conducted a Helsinki-approved research in a public hospital (ID: 0194-15-NHR, Galilee Medical Center) in which stroke patients and healthy subjects were filmed using the multiple Kinect system. The collected data was analyzed using machine learning techniques and compared with the FMA score provided by a medical professional, yielding very accurate predictions.

2 Related Work

Microsoft Kinect V2 was released in 2013 using time-of-flight technology and incorporating human tracking and skeleton representation [6, 7]. Since its release, the Kinect Camera V2 has been extensively studied for its noise statistics [8–10], tracking capabilities [12, 13], and compared with state-of-the-art and commercial human motion camera systems [12, 14–16]. The Kinect has been used in various applications such as medical applications including Parkinson Tracking [12, 17], Balance Disorders [18], rehabilitation

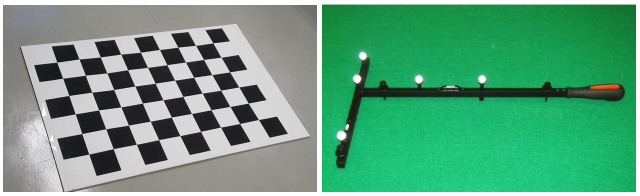

Fig. 3: Calibration accessories: Checker board (left) and VICON calibration wand (right)

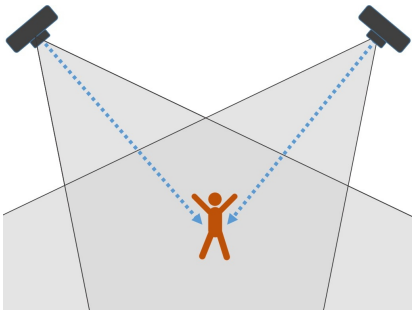

Fig. 4: A Two-Kinect camera setup.

[19], elderly monitoring [13] as well as sport and dance tracking and analysis [20, 21], and various other computer vision applications (see [22] for a review). Several studies have considered a multi-camera setup using Kinect cameras [18, 21, 23]. The sensors either showed interference (Kinect V1) or required special spatial layout [24], or assisted calibration [18].

In this study we aim to maximize the potential of a multi-3D camera setup beyond the capabilities shown previously and provide novel synchronization, calibration, and merging methods, as well as providing new insights and guidelines for multi-3D-camera setups and human motion trackers. We propose novel methods using both skeletal and 3D data for calibration, and using frontality and distance from camera to perform data merging.

Only a few studies attempted to implement automated Fugl-Meyer (FMA) tracking systems. In [25], the feasibility of automating FMA was tested, but the set up required a large space and was expensive (e.g., robotic arms, and EMG sensors), which is not suitable for a clinical setup. In another study [26], accelerometers were used for automated assessment, but its potential to quantify many FMA tests is low, due to the limited data acquired by the accelerometers [26]. Computerizing the FMA using depth sensors has been previously explored using a single Kinect V2 sensor [27] and Kinect V1 sensor [25]. However, as we show in our work, the multi-Kinect setup, is more reliable than these systems, since the FMA test requires evaluation of the patient body movements from both sides of the body and a single Kinect is limited in the sense that occlusion and non-frontality significantly reduces body tracking and skeleton formation.

3 The 3D multi-camera system

In this section we present our multi-camera tracking system which is non-invasive (no markers), inexpensive, portable and easy to use. It outputs a reliable skeleton and a merged point cloud.

As in any multi-camera system, several necessary challenges must be dealt with:

- Temporal Synchronization - ensures frames from different cameras are temporally aligned prior to merging their data.
- Camera Calibration - calculates scene to camera coordinate transformation.
- Pose Estimation (inter-camera alignment) - ensures alignment of data from different cameras into a uniform coordinate space, prior to merging.
- Data Merging - efficiently and reliably merges data from the different cameras into a single coherent representation.

3.1 Temporal and Data Synchronization

The first generation of Kinect (V1), designed with structured-light IR technology, raises challenges in a multiple-camera setup with scene overlap due to IR interference. In contrast, the Kinect V2, based on time-of-flight IR technology, allows using multiple depth sensors concurrently with only minor interference. However, the Microsoft Kinect PC Driver does not support connection of multiple sensors

into a single computer. Open source drivers such as the "OpenKinect" [28] allows connection of several sensors, however, at the cost of excluding the skeleton data. In our system, we developed a unique solution for handling several computers connected to multiple Kinects simultaneously. The temporal synchronisation between recordings is obtained using an NTP server. All data required for calibration and merging is transmitted to the server that runs the system algorithms. Since this requires only the skeleton data stream, which is of narrow bandwidth, the system runs in realtime over a LAN communication network producing a single fused skeleton stream (Figure 5).

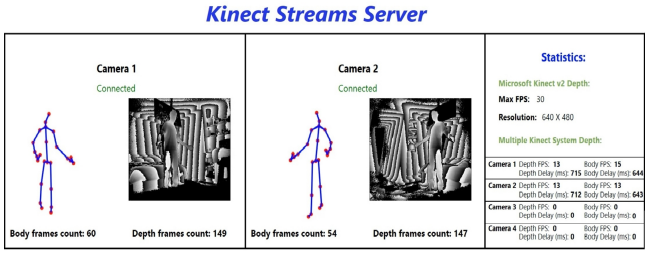

Fig. 5: Two Kinects during real-time network streaming

3.2 Pose Estimation - Inter Camera Calibration

Multiple camera setups require scene to camera calibration as well as pose estimation between cameras in order to merge their synchronized frames. Camera calibration, namely, mapping scene coordinates to camera coordinates is inherently given in 3D cameras. However, pose estimation or inter-camera calibration must still be calculated. There are several accurate techniques for calibrating multiple cameras using accessories such as Board [4], Flashlight [4], Marker [2]. and the Checker-board calibration [11] (Figure 3).

However, running calibration sessions, and using special calibration equipment is inappropriate in our case where simplicity of activation is necessary especially for home use of the system without professional operators. We propose a novel method for calibration. It can be run on-the-fly and requires no special calibration session. Rather, it relies on the body of the subject being captured in the scene (Figure 11). We exploit both the skeleton representation and the 3D point clouds captured by the camera to achieve accurate calibration in very little time.

The skeletal data consists of only a few points (Figure 1) and they are tagged so that paired matching points between cameras is easily obtained. However, the positions of the skeleton points are unreliable, noisy and at times missing. On the other hand the 3D cloud of points are of higher positional precision, however there are over 300,000 data points and they are not matched between the cameras. Additionally, perfect alignment between cameras is not always guaranteed. In our calibration approach we exploit the advantages of each type of data set. The calibration process involves 2 steps:

1. Determine the pose (rigid transformation) between cameras, based on the skeleton data.
2. Use the estimated pose from the skeleton as an initializer for estimating the pose using the 3D point clouds, obtaining a more precise transformation.

The multi-camera system consists of a number of cameras (typically 2-6) that must be calibrated. In the following we describe the calibration between a pair of cameras. The extension to a large number of cameras will be discussed later.

3.3 Skeleton based pose estimation

The transformation between cameras is initially calculated by aligning the body skeletons as shown in Figure 6. This can be described as finding the optimal 3D rigid transformation (rotation and translation) between two sets of corresponding 3D point data.

The skeleton points are streamed per frame per camera and each point is associated with: a name, 3D coordinates and a "Confidence of tracking" value which is exploited by our proposed algorithms because it directly indicates the joint's accuracy. The joint confidence is graded as: "well-tracked", "inferred" and "not-tracked". The skeleton is estimated per video frame based on the depth map acquired by the camera. There are frequent estimation errors caused both from the depth measurement error and due to body parts occluded from the camera's line of sight. Furthermore, since the Kinect was "trained" selectively for frontal views, it does not work as well on side or back poses (see Figure 2). One of our goals is to minimize these errors by combining body skeletons from multiple views and producing an improved body-skeleton stream. In the context of calibration, errors and noise in the skeleton joints must be taken into consideration during the process.

We use the Kabsch algorithm [29, 30] to determine the optimal rotation between cameras based on the skeleton points. Given two sets of points, they are first translated so that their centroids are at the origin (Equation 1) and then use Singular Value Decomposition (SVD) is on the Cross-Covariance Matrix of the points to determine the rotation matrix R (Equations 2-3). The translation vector t is calculated by applying the rotation and subtracting the centroid (Equation 4).

$$H = \sum_{j=1}^{N} \left(p_A^i - \text{centroid}_A\right) \left(p_B^i - \text{centroid}_B\right)^T \quad (1)$$

$$[USV] = \text{SVD(H)} \quad (2)$$

$$R = VU^T \quad (3)$$

$$t = -R \times \text{centroid}_A + \text{centroid}_B \quad (4)$$

Although only a small number of joints are computed per skeleton, the calibration is performed over a large number of frames streamed by each camera, since the camera pose does not change from frame to frame. Thus we use RANSAC [31] to optimize the 3D rigid transformation. Smart selection of skeleton joints is performed based on joint confidence measures as well as joint type (body part); At each iteration three joint pairs from corresponding frames of the two cameras are randomly chosen, such that every joint is from a different body region: upper, mid and lower body sections. This

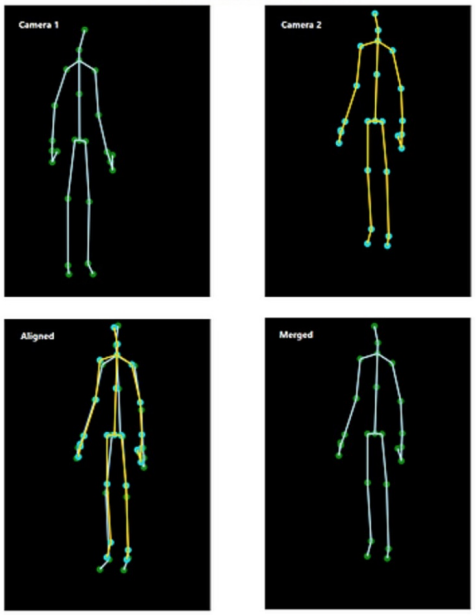

Fig. 6: Aligned and merged body skeletons from 2-Cameras. Top: camera outputs. Bottom: aligned skeleton (left) and merged (right).

improves robustness and viability of the computed 3D transformation. The 3D transformation with the largest number of supporting joints across all frames is returned as output.

3.4 Pose estimation from 3D point clouds

The pose calculated from the skeleton data provides a good estimate for the inter-camera calibration. However due to noise and errors in the skeleton, the resulting 3D transformation can still be improved. Thus we now exploit the more reliable 3D point clouds supplied by the cameras. The inter camera transformation is further refined by aligning the 3D points using the ICP algorithm [32]. This, however, requires a good initial guess, otherwise, ICP may converge to an incorrect local minimum (see examples in Figure 7). Thus, we use the transformations calculated from the skeleton alignment as the initial guess for alignment of the 3D points using ICP. We introduced a further improvement by including, in addition to the 3D body points, the 3D points of the ground plane ("the floor") near the subject's legs. This addition of points increases scene overlap of the 3D points between the two cameras. The resulting 3D transformation shows an improvement of 37% in accuracy over the initial estimate obtained using the skeleton data (see Section 3.8).

3.5 Multiple Camera System

The calibration and pose estimation algorithm described above considered the case of only two cameras. Our proposed system runs on multiple cameras, with typical scenarios ranging between 2-6 cameras. Extending the calibration to more than two cameras, is based on evaluating several pairwise inter-camera calibrations as described in Sections 3.3-3.4. A critical insight is that due to Kinect inaccuracy for non-frontal views (resulting in low confidence or erroneous skeleton joints), it is of significant importance to calibrate between cameras whose skeleton joints are reliable in corresponding frames. This criterion will arise significantly more often when the two cameras have as similar a viewpoint as possible. This will be shown empirically in Section 3.8. Thus, the calibration protocol calibrates pairs of neighboring cameras, resulting in pairwise 3D rigid transformations (see Figure 8). The final transformation from camera to a single coordinate system (e.g. that of Camera 1) is obtained by concatenating the transformations along the shortest path.

3.6 Data Merging

Following calibration and pose estimation, all the acquired data from all cameras are aligned in a common coordinate frame. In this section we describe our methods for merging the data, specifically, the skeleton and later the 3D cloud of points. We propose several merging criteria, each of which targets a weakness of the camera system:

1. The Kinect camera supplies a skeleton per frame with a confidence ranking per joint. Thus joints with low confidence are either disregarded or weighted very low in the merging process.
2. IR sensors are inherently noisy [8]. Additional, distortions are introduced due to scene interference, lighting, scene inter-reflections etc. This noise directly affects the quality of the point clouds and indirectly, the skeletons. Reduction in noise can be introduced by "averaging" data from multiple views.
3. Kinect's skeleton computation algorithm does not perform well on non-frontal views. This is due to self occlusion and probably since the algorithm was mainly trained on near frontal views. This problem suggests a weighting scheme that awards skeleton joints associated with frontal views. Frontality can be deduced by evaluating the depth values of the 2 shoulder joints as well as their confidence.
4. Kinect's skeleton computation algorithm does not distinguish the front vs the back of the subject. Thus, when the subject is back facing the camera, it will yield a skeleton in a frontal pose. However, even if the skeleton is reflected, it's pose still remains very inaccurate. We developed a criterion that recognizes back facing skeletons and assigns its joints, the minimal confidence level.
5. Given the nature of the depth data, distant objects have fewer sampled depth points per area and as such are less reliable. Preference is thus given to joints acquired by the camera closest to the joint.

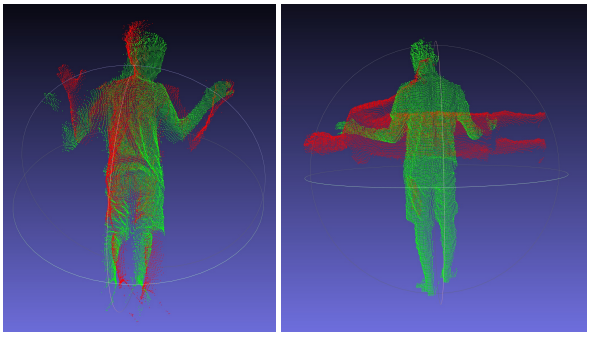

Fig. 7: Aligning cloud points using ICP without good initial guess produces incorrect alignments (compare with successful alignments in Figure 9).

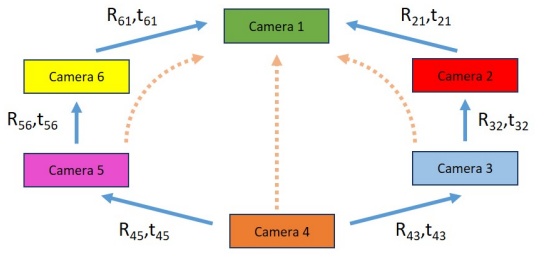

Fig. 8: Calibrating a 6-camera system. Calibration is performed pairwise between neighboring cameras and propagated to the same coordinate system along the shortest path.

Data merging is performed per skeleton joint, per frame. When considering a set of joint measurements P from different cameras in a given time frame, only the measurements with the highest confidence ranking from the set are retained. For each such joint measurement $p \in P$, we measure its frontality angle α_p and its distance from the camera $dist_p$. To obtained the single merged joint p_m, several possible merging techniques are considered:

- Average:

$$p_m = \frac{\sum_{p \in P} p}{|P|} \tag{5}$$

- Frontality:

$$p_m = \arg\min_{p \in P} \alpha_p \tag{6}$$

- Distance:

$$p_m = \arg\min_{p \in P} dist_p \tag{7}$$

- Weighted Average:

$$p_m = \frac{\sum_{p \in P} w_p p}{\sum_{p \in P} w_p} \tag{8}$$

The weights w_p can be equal to the frontality weights w_f, the distance weights w_d, or their product $w_f w_d$, where:

$$w_f = \exp^{-\alpha_p^2/\gamma_a}.$$

$$w_d = \exp^{-(dist_p - d_0)^2/\gamma_d}.$$

The constants were determined empirically (see Section 3.8) and were set to, $\gamma_a = 1.1, \gamma_d = 2.0$, and $d_0 = 0.5m$ (the Kinect minimal sensing distance).

An example of merged skeletons is shown in Figure 6. In Section 3.8 we experiment and analyze these different merging methods.

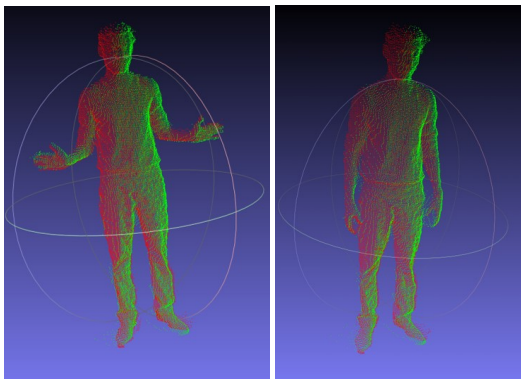

Fig. 9: Aligned and merged cloud points from 2-Cameras

3.7 Cloud Points Registration

An additional outcome of the proposed system is that given the calibration parameters, the 3D cloud points from the different cameras can be aligned and merged. A circular camera setup can transform the partial 3D point clouds into a full body mesh. Thus, we transform the multiple 2.5D data streams into a single and complete 3D representation (Figure 9).

3.8 Camera System Experimentation and Results

Several tests were performed to evaluate the system including evaluating the effects of the number and positioning of multiple cameras, evaluating the precision of the calibration and pose estimations and comparisons between the different data merging methods.

3.8.1 Comparison with Standard Calibration Techniques:

Several techniques are available to perform calibration of multiple Kinect sensors. The well-known checkerboard calibration [33] could, in principle, be used with the Kinect RGB sensor. However, the RGB sensor is positioned away from the IR sensor (Figure 10) which causes the calibrations obtained for the IR and for the RGB, difficult to compare.

Our novel calibration method is based on the 3D Body-Skeleton stream which is calculated based on the 3D depth map. The Kinect depth map has known inaccuracies of up to 18 mm, that arises during the process of mapping the 3D point cloud from the RAW-IR image [34]. To optimally evaluate the quality of our proposed calibration method, we exploit the fact that Kinect cameras provide IR video streams. Thus, the standard checkerboard calibration (Figure 11) [33] can be applied on the Kinect IR images. We compared the checkerboard calibration results with our calibration based on depth sensing and skeleton alignment. Testing was performed by applying both calibration methods under the same 2-camera setup and the comparison between resulting transformations was evaluated using 3 evaluation measurements:

1. Average RMSE - A set of 3D points is transformed by the transformation found by the checkerboard and by the body-skeleton calibrations. The RMSE is calculated between the corresponding 3D points of these two sets and the average RMSE value is calculated.
2. Average Rotation Angle Error - The average difference between corresponding Euler-angles representing the two rotation matrices.

Fig. 10: Kinect RGB and IR sensors positions

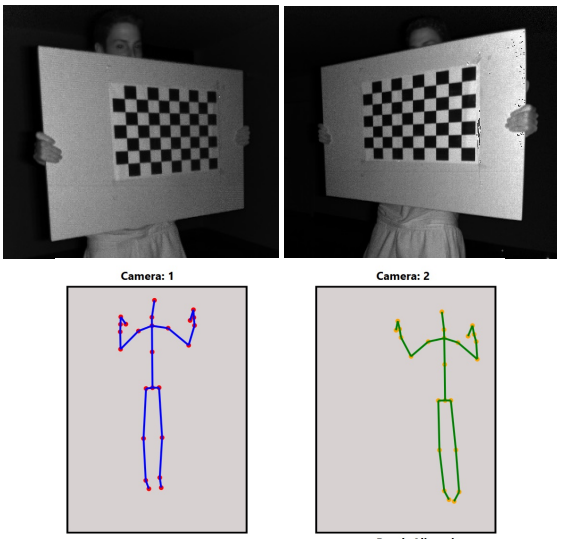

Fig. 11: IR Checker board images (top) and the Skeleton images used in the compared calibrations.

3. Average Translation Distance - The average 3D Euclidean distance between the two translation vectors.

Our calibration method showed accuracy on par with the checkerboard method (Table 1). In order to estimate the accuracy of the checkerboard calibration in itself, we measured the average RMSE between several checkerboard calibration sessions and found it to be 0.92 cm. We also measured the average RMSE between several body-skeleton calibration sessions which was found to be 1.44 cm. The average RMSE between our calibration and the checkerboard calibration was found to be 2.89 cm. The average rotation angle error and the average translations Euclidean distance between the checkerboard and the proposed calibration was found to be 0.79° and 2.5 cm respectively. Considering the known accuracy errors in the Kinect 3D Cloud points which can reach 1.8 cm [34] and the accuracy of each method independently, the results imply that our proposed skeleton based calibration provides results on par with standard calibration techniques (Figure 12).

Table 1 Ground truth calibration vs. Body calibration results

Ground truth error rate (meters)	0.0092
Experimental method error rate (meters)	0.0144
Average RMSE between methods (meters):	0.0289
Average Rotation Error (deg.):	0.7900
Average Translation Error (meters):	0.0250
RAW-IR to 3D internal Kinect Error (meters):	0.0180

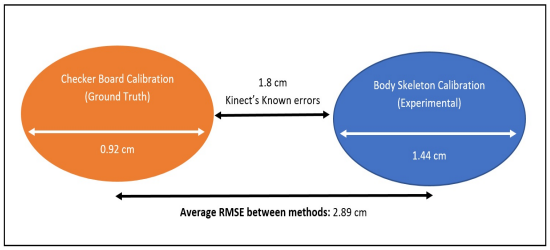

Fig. 12: Average RMSE results.

3.8.2 Calibration Evaluation:

Assuming our calibration method performs well compared to the standard checkerboard calibration in a 2-camera setup, we extend our testing to a multi-camera system with an increased number of cameras. We recorded several sessions in a circular setup of 6-cameras (Figure 8). Per session, these cameras recorded several thousand frames simultaneously, of a single person performing articulated motions in the scene. As mentioned above, calibration of our multi-camera system was performed between pairs of neighboring cameras since larger scene overlap significantly increases accuracy. To obtain the final inter-camera calibration, each camera is mapped to a single coordinate system by concatenating pairwise transformations along the shortest path in the camera circle. To evaluate the calibration we chose a random camera and transformed its 3D points to the farthest camera along the circle in a clockwise path and then in the anticlockwise path (for example, camera 4 data is mapped to camera 1 by concatenating transformations 4 to 3, 3 to 2 and 2 to 1 and then concatenating transformations 4 to 5, 5 to 6 and 6 to 1). A perfectly calibrated system would show perfect alignment of these two mapped sets. In practice we found a the mapping error to be 0.0311 meters following the skeleton pose estimation and 0.0196 meters when 3D point cloud alignment was included in the pose estimation (See Section 3.2). This implies that using the 3D point clouds alignment after the skeleton alignment, improves accuracy by 37%.

3.8.3 *Camera Setup Evaluation:*

In the next test, we evaluated the quality of results as a function of the camera setup, namely, the number of cameras and their positions. Two factors affect the quality of the merged skeleton output: the quality of the skeleton joint measurements acquired by the cameras and the merging algorithm used to combine the the skeleton joints data into a single reliable joint. In this section we evaluate the quality of the measurements as a function of the camera setup.

As described in Section 3.6, for every joint, in every time-frame, a collection P of joint measurements are obtained from a group of cameras. The final merged skeleton joint p_m is calculated from these measurements. We tag the reliability of p_m based on the quality and confidence of its measurements $p \in P$. Specifically, each measured joint p is tagged for confidence by the Kinect camera as high ("well-tracked"), mid ("inferred") or low ("not-tracked"). Furthermore, joint measurements originating from a back facing subjects are detected automatically and are tagged as low confidence in our system. A merged joint p_m is tagged as having high confidence if at least one of the joint measurements is of high confidence. A merged joint is tagged as back pose if all joint measurements are back posed.

For testing, we again consider the circular setup of cameras numbered 1-6 as shown in Figure 8. A camera configuration is a specific subset of these cameras. A configuration depends on the number of cameras as well as their positions within the camera circle. Table 2 shows statistics on the number of high confidence merged joints and the number of back posed joints resulting when merging using a specific camera configuration. Several conclusions may be deduced from the results. First, it is clear that the quality of the measurements improves with the number of cameras in the setup. Both an increase in number of high confident joints as well as a decrease in low confidence back posed joints can be seen. This is due to the fact that more cameras, increase the chances of a frontal view, a closer joint and more reliable joint, implying a higher quality of measurement. The second conclusion involves camera positions within the setup. We distinguish between two types of configurations: Neighboring configuration where all cameras are consecutive in the camera circle, and Circular configurations where cameras are "spread out" along the camera circle. It can be seen that for the same number of cameras, the Circular configuration is preferable, providing higher quality measurements. This can be explained in that Neighboring configurations do not cover all viewpoints of the subject resulting in many frames in which no camera has a frontal view of the subject. In contrast, a Circular configuration increases the chances of at least one of the cameras capturing a reliable frontal view for any pose of the subject.

Table 2 Evaluating Camera Configuration

Camera Configuration	Configuration Type	Back Pose joints	Back Pose joints %	High Confidence joints	High Confidence joints %
[1]	Single camera	37300	51.20	33796	46.39
[1,2]	Neighboring	21900	30.06	49518	67.97
[1,3]	Circular	21125	29.00	49973	68.60
[1,2,3]	Neighboring	20950	28.76	50474	69.28
[1,3,5]	Circular	6325	8.68	66102	90.74
[1,2,3,4]	Neighboring	3775	5.18	68196	93.61
[1,2,3,4,5]	Neighboring	2150	2.95	70641	96.97
[1,2,3,4,5,6]	**All cameras**	**0**	**0**	**72815**	**99.95**

3.8.4 *Skeleton Merging Evaluation :*

In the next test, we evaluated the quality of the merged skeleton based on the merging algorithm used to combine the measured joints into a single reliable joint. We use two measures for evaluating a skeleton's accuracy and robustness. The first evaluates the consistency of the skeleton's bone length across all time frames. For each bone (segment connecting two neighbouring joints) we calculate the standard deviation of the bone length over all confident joints across all video frames. We consider this measure on the 3 dominant and most reliable bones: shoulder to spine, shoulder to elbow and elbow to wrist.

The second measure for skeleton evaluation evaluates the accuracy of the skeleton with respect to new camera measurements. Considering again our configuration of 6 cameras (Figure 8). For each configuration (subset of cameras) we apply calibration and merge the skeletons. Given the merged skeleton, we use the set of remaining cameras for evaluation. For each reliable measured joint from these new cameras, we compute the distance to their corresponding joint in the merged skeleton and count the number of these distances that are smaller than a given threshold d.

Our tests used these two measures on different merging methods for different camera configurations. The merging methods as discussed in Section 3.6 (Equations 5-8) include: simple averaging, distance based, frontality based and weighted averaging (frontality weighted, distance weighted and weighted by both).

Table 3 shows the the resulting measures for different camera configurations and different merging methods. Several conclusions may be deduced from these results. First, frontality seems to take the main role for achieving robust and accurate merged skeletons, in all of its merging variations, whether choosing the "most frontal" camera in each frame or averaging with frontality weight. Another conclusion from the results is that the method of merging using frontality changes with the number of cameras and with the amount of scene overlap that these cameras share. Frontality based merging is advantageous with a small number of cameras (rows marked in bold) and averaging using frontality weights is better for a large number of

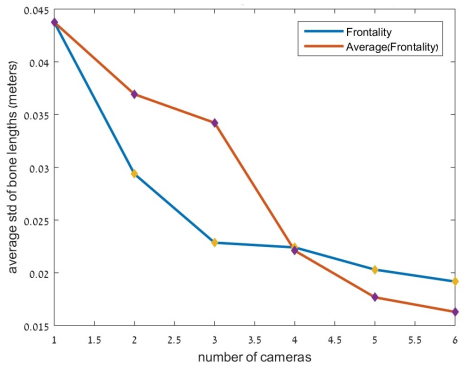

Fig. 13: Skeleton alignment example

Table 3 Evaluating Skeleton Merging

Camera Config	Merging Method	Spine to Shoulder Std (m)	Shoulder to Elbow Std (m)	Elbow to Wrist Std (m)	Supporting Joints % (d=0.05m)	Supporting Joints % (d=0.07m)
[1,2,3]	Average	0.0374	0.0363	0.0445	42.24	55.77
[1,2,3]	Distance	0.0282	0.0332	0.0341	41.11	55.10
[1,2,3]	**Frontality**	**0.0229**	**0.0265**	**0.0454**	**46.06**	**60.23**
[1,2,3]	Avg(Distance)	0.0363	0.0325	0.0410	42.24	55.85
[1,2,3]	Avg(Frontal)	0.0373	0.0364	0.0443	42.36	55.87
[1,2,3]	Avg(Dist+Frontal)	0.0363	0.0325	0.0409	42.37	55.95
[1,3,5]	Average	0.0334	0.0255	0.0405	62.23	76.13
[1,3,5]	Distance	0.0218	0.0349	0.0536	60.95	75.71
[1,3,5]	**Frontality**	**0.0207**	**0.0203**	**0.0270**	**63.70**	**78.19**
[1,3,5]	Avg(Distance)	0.0345	0.0261	0.0407	62.23	76.15
[1,3,5]	Avg(Frontal)	0.0322	0.0227	0.0375	63.39	77.38
[1,3,5]	Avg(Dist+Frontal)	0.0333	0.0233	0.0352	63.41	77.65
[1,2,3,4,5,6]	Average	0.0150	0.0215	0.0288		
[1,2,3,4,5,6]	Distance	0.0167	0.0372	0.0523		
[1,2,3,4,5,6]	Frontality	0.0119	0.0174	0.0282		
[1,2,3,4,5,6]	Avg(Distance)	0.0148	0.0375	0.0338		
[1,2,3,4,5,6]	**Avg(Frontal)**	**0.0089**	**0.0155**	**0.0245**		
[1,2,3,4,5,6]	**Avg(Dist+Frontal)**	**0.0088**	**0.0154**	**0.0282**		

cameras (rows marked in bold). Figure 13 further emphasises this conclusion by plotting the frontality measure vs number of cameras and shows the tradeoff point to be at 4 cameras. This effect can be explained by considering that configurations with large scene overlap tend to produce more than one good (frontal) joint measurement per frame, and thus these should be averaged. However for small scene overlaps (such as [1,3,5]) for most frames, only one camera captures a frontal pose and thus frontality based merging in which the single best measured joint is selected, outperforms averaging.

Finally, comparing the two 3-camera configurations [1,2,3] and [1,3,5] (rows marked in bold) we see support for the conclusion in our previous test, that the circular configurations are more advantageous for skeleton accuracy than neighboring configurations. The circular configuration [1,3,5] improves over the neighboring configuration [1,2,3] under both our measures.

4 Motion Analysis Application: Stroke Rehabilitation

The main purpose of developing our multi camera tracking system is for it to be used for medical and tele-medicine applications. In this section we describe an experiment we performed in using this system for automatically assessing the medical condition of stroke patients. Additional details on the medical experiment can be found in [35].

4.1 Background

Stroke is a serious medical condition which occurs when the blood flow to an area in the brain is cut off. If a stroke is not detected early enough, permanent brain damage or death may occur. Around 800,000 people a year in the US incur a stroke and there are 6.5 million stroke survivors in the US. Stroke accounts for 1 of every 20 deaths in the US and amounts to nearly 133,000 people a year [36]. Rehabilitation of stroke patients is a long and slow process. The level of rehabilitation of a patient is typically evaluated using the Fugl-Meyer Assessment (FMA) [1]. This test involves the patient performing specific motor actions. A physician or skilled medical professional rates the performance on the FMA scale and a score is derived. Thus, this score is subjective and lacks a high degree of objectivity, impartiality and sensitivity.

The Fugl-Meyer Assessment (FMA) is a stroke-specific, performance-based impairment index. It is designed to assess motor functioning, balance, sensation and joint functioning in patients with post-stroke hemiplegia (a weakness of one entire side of the body). FMA is applied clinically and is used to determine disease severity, describe motor recovery, and to plan and assess treatment.

Since stroke is a debilitating disorder, patients often find it difficult to travel to stroke rehabilitation centers for testing and thus do not receive optimal medical care. Recent advances in information and communication technologies connect specialists that are centered in urban areas with population in suburban and rural areas and thus benefit these patients. This technology, which enables treatment of patients in remote areas and in nursing homes is termed *Tele-Medicine* stemming from the use of telecommunication in order to provide health care. Under the idea of Tele-Medicine, we propose an automated system for tracking and evaluating stroke rehabilitation using a cost-effective home-based system. The system must be non-invasive, easy to use, inexpensive and can be activated in a home setting. In order to be consistent with the physician's view of the examinee and also to prevent tracking failures, at least two cameras are necessary for analyzing both sides of the patient independently. Thus, our multi-camera system based on 3D cameras is appropriate for this task.

4.2 Medical experiment

We implemented the multi-camera motion capture system in a medical setting in the context of automating the FMA assessment procedure [1] in a home-setting. Our FMA application uses a two camera setup (Figure 4), ensuring each side of the patient's body is properly viewed, and producing a reliable data to be used in

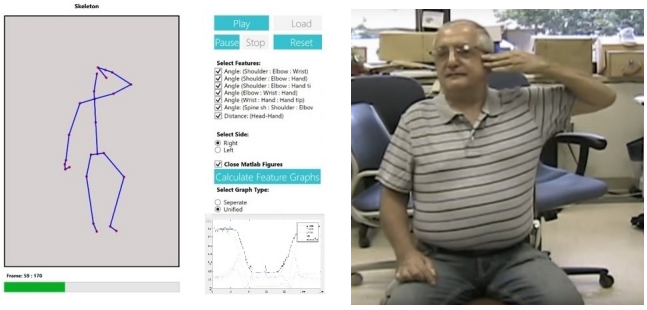

Fig. 14: Analysis of the Fugl-Meyer hand salute test.

Fig. 15: Fugl-Meyer Salute and Hand Lift tests.

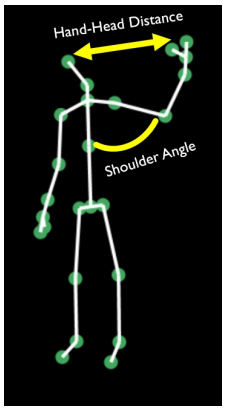

Fig. 16: Skeleton movement features example.

analysis. We conducted an Helsinki-approved study in a major public hospital (ID: 0194-15-NHR, Galilee Medical Center) using our multi-camera tracking system. 22 participants were filmed during their Fugl-Meyer assessments. The participants were twelve stroke patients and ten control healthy subjects. The subjects performed the Fugl-Meyer assessment in the hospital testing room under the guidance of a medical professional, one of the authors, who also provided the FMA rating and FMA score for the patients and the healthy subjects. The subject performed the FMA motions with the hand on the affected side as well as the hand on the unaffected side (termed-healthy hand). Each motion was repeated several times.

Two Kinect cameras were set up in the testing room so that they did not interfere with the testing yet obtained unobstructed views of the subject. The cameras were positioned at a 45 degree angle to the subject's front view, and at a distance of approximately 1.5-2 meters (see Figure 4). The two cameras recorded the body skeleton of the subject performing the motor task. In our experiment we focused on two key motor movements from the Fugl-Meyer scale protocol, the "Salute" and "Hand lift" (90 degrees) tests (Figure 15).

4.3 Fugl-Meyer analysis features

An analysis application was developed for extracting measurements from the tracked body skeleton recordings as shown in Figure 14, and these measurements were then used to detect correlations with the physician's diagnosis. Measurements were extracted from each frame of the acquired skeleton sequence, for each of the two FMA movements. Measurements were derived from the angle defined by three skeleton joints, the distance between a pair of joints or the height moved by a single joint (Figure 16). The list of measurements is given in Table 4. From these measurements , a list of features was calculated per frame for the whole skeleton sequence thus, providing a feature vector for each sequence, per each motion and per each subject. The extracted features per sequence included:

Table 4 List of measurements used in our testing.

Measurements
Angle (Shoulder - Elbow - Wrist)
Distance (Head - Hand)
Angle (Spine Shoulder - Shoulder - Elbow)
Height (Hand)
Height (Elbow)
Height (Wrist)
Angle (Shoulder - Spine Shoulder - Spine Mid)
Angle (Hip - Shoulder - Elbow)
Distance (Hand - Knee)
Distance (Wrist - Hip)
Angle (Knee - Hip - Spine Mid)
Distance (Head - Elbow)
Distance (Camera to Shoulder)

- Sequence time length.
- Minimum, and maximum of each measure within the sequence.
- Average and variance of each measure within the sequence.
- Difference between start and end values of each measure along the sequence.
- Average speed and acceleration of each measure as it changes along the sequence.

The feature vectors associated with each skeleton sequence used in the analysis and the resulting analysis are described in the following section.

4.4 Results

In this section we analyze the data collected during the medical trial. We show that our results concur with the medical guidelines defined by the medical specialists, including the known significant movement features, the differences between patients and healthy subjects behavior and the special case of motor compensation in stroke patients. We used machine learning tools to perform classification and feature ranking. The results are presented in the following three sub sections and include: classifier results, showing success rate at predicting FMA scores, feature selection, presenting the most significant features found by the classifiers, and statistical analysis for detecting motor compensation in the stroke patients movement.

4.4.1 Classifiers results :

Participants in the experiment performed several repetitions of the "Hand Lift" and "Salute" motions using each of the hands. Stroke patients typically suffer from weakness of one entire side of the body (Hemiplegia), but also suffer from a general cognitive deterioration and reduced motor ability that influences all body functioning including the "healthy" side of the body. Thus, in order to distinguish between healthy subjects and patients, we analyze each side of the subject separately as well as analyzing the asymmetry of the subject's motor performance on both sides. Towards this goal we performed several classification tests on the collected data:

1. "Raw data" - each repeated movement on each body side is collected separately as a single sample for classification. Thus, each subject has numerous samples for classification for each movement.
2. "Repetition averaging" - repetitions are averaged on each body side thus, each subject has two samples for classification for each movement, one for each body side.
3. "Asymmetry measure" - repetitions are averaged on each body side and then a measure of asymmetry is applied: $\frac{\|Right-Left\|}{\|Right+Left\|}$. Thus, each subject has one sample per movement for classification.

To test the effect of using our multi camera setup compared to a single camera setup, we analyzed these three data sets under 4 possible camera setups and data merging methods:

(a) Single camera - one of the two cameras in the setup was randomly chosen for each subject as the source camera for data collecting and only data from this camera was used in the analysis.

(b) Two cameras - following the merging of skeletons as described in Section 3.6, we consider 3 possible analyses:

1. Averaging - Where the data from both cameras are averaged.
2. Best choice - Where the best camera data is selected. In this case the data is that which is acquired by the closest camera to the skeleton joint (subject's hand).
3. All samples - Where the data from both cameras are not merged and used as sperate samples in the classification.

In order to classify between patients and healthy subjects, we ran several classifications based on the 3 types of data and the 4 types of camera acquisitions. The features used per sample were those detailed in Table 4, and the class label of each sample was the FMA score provided by the physician (score of 0-1 vs score of 2-3).

To build the classifiers, we used SVM [37], Single Decision Tree (C4.5) [38] and the Random Forest [39] which uses a forest of decision trees and often achieves better results. In preliminary tests, while analysing several setups and data types, SVM did not achieve better classification results than random forests. For example for the "all samples" setup for the "raw" data type for both salute and hand lift movements, SVM achieved 87.63% success rate while Random Forest achieved 90.9%. An additional goal in the classification process is to validate the medical guidelines that were considered during feature extraction, thus the final ranking of the features in the classifier is of interest. Decision trees are advantageous compared to SVM in this aspect due to the built-in feature ranking process. Thus we report our results using only the decision tree classification approach.

The decision tree was built using three folds, with batch size 100, and confidence factor 0.25. The Random Forest classifier used the same parameters and up to 100 trees. A leave-one-out strategy was used for cross-validation, which entailed leaving out one subject at each iteration and training on the rest

Tables 5 and 6 show classification results for the Hand Lift and for the Salute motion respectively. Table 7 shows classification results when data of both motions were used collectively. The tables present the summarized results across all subjects comparing the different data formats and different camera setups.

Analyzing the results shown in the tables, it can be seen that the Salute movement shows better classification results than the Hand Lift movement, with correct percentage of 96-100% (Table 6 rows 1-6). The success rate for the Hand Lift motion (Table 5) is much lower. This may be explained by the difficulty and complexity of the Salute movement for patients. When using all data from both movements (Table 7) good performance is also achieved at 87-93% success rate.

Table 5 Classifiers benchmark summary - Hand lift

Camera setup	Data used	Samples	C4.5	R-F
All samples	**Raw data**	**214**	**78.50**	**82.71**
Best choice	Raw data	134	69.40	82.08
Single camera	Raw data	130	64.61	77.69
Averaging	Repetitions average	44	77.27	79.54
Best choice	Repetitions average	44	88.63	79.54
Single camera	Repetitions average	45	66.66	82.22
Averaging	**Asymmetry measure**	**22**	**36.36**	**77.27**
Best choice	**Asymmetry measure**	**22**	**81.81**	**90.90**

Table 6 Classifiers benchmark summary - Salute

Camera setup	Data used	Samples	C4.5	R-F
All samples	Raw data	149	90.60	97.31
Best choice	Raw data	95	88.42	96.84
Single camera	Raw data	87	82.75	98.85
Averaging	Repetitions average	44	93.18	97.72
Best choice	**Repetitions average**	**44**	**97.72**	**100**
Single camera	**Repetitions average**	**43**	**97.67**	**100**
Averaging	**Asymmetry measure**	**22**	**77.27**	**63.63**
Best choice	**Asymmetry measure**	**22**	**45.45**	**59.09**

Table 7 Classifiers benchmark summary - Both Salute & Hand lift

Camera setup	Data used	Samples	C4.5	R-F
All samples	Raw data	363	77.95	90.90
Best choice	Raw data	229	84.71	89.95
Single camera	Raw data	217	77.88	87.09
Averaging	Repetitions average	88	85.22	90.90
Best choice	**Repetitions average**	**88**	**82.95**	**93.18**
Single camera	Repetitions average	88	90.90	81.81
Averaging	**Asymmetry measure**	**44**	**70.45**	**70.45**
Best choice	**Asymmetry measure**	**44**	**72.72**	**75.00**

The results also show that classifying using the asymmetry measure yields inferior results compared to the other data sets. This, and the fact that the classifications were run on data from both the affected and healthy body sides, imply that both the stroke affected side and the subject's healthy side provide distinguishing features enabling separation of patients from healthy subjects. In the context of the cameras setup, it can be seen that the 2-camera setup outperforms the single camera setup in most cases. The average correct percentage of all cases with a single-camera is 87.5% compared to 90.4% in the two-camera setup (excluding the asymmetry versions). Advantage is seen when using the best choice camera over the averaging method.

4.4.2 Feature selection:

Classification between healthy subjects and stroke patients was shown to performed better, when analyzing each side of the subject independently rather than using the asymmetry measure. Thus, in this section we focus on these high-performance classifiers and explore them further by ranking their features.

We use two feature selection ranking methods:

1. The Information Gain univariate feature ranking (Kullback-Leibler divergence) [40].
2. Random Forest features ranking [39] that summarizes the ranks of all selected features according to their positions in the decision trees.

Tables 8, 9, 10 and 11 show rankings of features from the 6 best classifiers (Tables 5, 6, 7) using the info gain ranking and the random forest ranking respectively. The top 10 features are listed. It can be seen that the 3 most significant features across almost every classifier are (marked by bold rows):

1. Total time - Representing the period of time between start and end of the movement. Patients are typically slower than healthy subjects thus the significance of this feature is clear.
2. Distance (Head - Elbow) Average speed - Representing the speed of the "changing distance" between the head and the elbow while performing the hand lift or salute movement. These two joints are central to performing both movements and the speed in which the movement is performed is reflected in the change in distance between these joints.
3. Angle (Hip - Shoulder - Elbow) Average speed - This feature defines the speed of change in the "shoulder" angle (arm to body angle) which is intuitively the significant angle while estimating the hand lifting or salute movements.

All these top features represent, in some form, the difference of "speed" between healthy and patients. The specific skeleton joints represented in these features are directly related to the medical guidelines of the FMA and the instructions given by the specialist to the patient (e.g. for the Hand lift, the patient is required to lift hand to 90° angle between arm and body).

Table 8 Feature selection results - Information Gain (Part 1)

All Samples	Raw Data	Salute & Hand lift
Top 10 Ranked Features		Rank
Total time		**0.60**
Distance (Head - Elbow) - Average Speed		**0.39**
Distance (Head - Hand) - Variance value		0.32
Distance (Head - Hand) - Start-Stop Difference Value		0.28
Distance (Head - Elbow) - Variance Value		0.27
Distance (Head - Elbow) - Start-Stop Difference Value		0.26
Angle (Hip - Shoulder - Elbow) - Average Speed		**0.26**
Distance (Hand - Knee) - Max Acceleration		0.23
Height (Hand) - Min Speed		0.21
Distance (Hand - Knee) - Average Speed		0.21

Best choice	Repetitions avg	Salute & Hand lift
Top 10 Ranked Features		Rank
Total time		**0.68**
Distance (Head - Elbow) - Average Speed		**0.48**
Angle (Hip - Shoulder - Elbow) - Average Speed		**0.46**
Distance (Hand - Knee) - Average Speed		0.45
Angle (Hip - Shoulder - Elbow) - Min Speed		0.40
Distance (Head - Elbow) - Max Speed		0.40
Distance (Hand - Knee) - Max Acceleration		0.33
Distance (Head - Elbow) - Variance value		0.32
Angle (Shoulder-Spine Shoulder-Spine Mid) - Avg Speed		0.31
Distance (Head - Elbow) - Start-Stop Difference value		0.31

All samples	Raw data	Hand lift
Top 10 Ranked Features		Rank
Total time		**0.42**
Distance (Head - Elbow) - Average Speed		**0.29**
Angle (Hip - Shoulder - Elbow) - Variance value		0.27
Distance (Head - Hand) - Variance value		0.24
Height (Hand) - Start-Stop Difference value		0.22
Height (Wrist) - Start-Stop Difference value		0.22
Distance (Head - Hand) - Average Speed		0.21
Distance (Head - Elbow) - Start-Stop Difference value		0.20
Distance (Head - Hand) - Start-Stop Difference value		0.20
Height (Hand) - Variance value		0.18

Table 9 Feature selection results - Information Gain (Part 2)

Best choice	Repetitions Avg	Hand lift
Top 10 Ranked Features		Rank
Total time		**0.64**
Angle (Spine Shoulder - Shoulder - Elbow) - Max Accel.		0.38
Angle (Hip - Shoulder - Elbow) - Start-Stop Difference value		0.34
Angle (Hip - Shoulder - Elbow) - Variance value		0.33
Angle (Hip - Shoulder - Elbow) - Average Speed		**0.33**
Angle (Shoulder - Elbow - Wrist) - Max Acceleration		0.33
Distance (Head - Elbow) - Average Speed		**0.32**
Height (Hand) - Variance Acceleration		0.31
Height (Hand) - Max Acceleration		0.31
Distance (Hand - Knee) - Max Acceleration		0.31

All samples	Raw data	Salute
Top 10 Ranked Features		Rank
Total time		**0.84**
Distance (Head - Elbow) - Average Speed		**0.73**
Distance (Head - Hand) - Average Speed		0.62
Distance (Hand - Knee) - Average Speed		0.61
Angle (Hip - Shoulder - Elbow) - Average Speed		**0.60**
Distance (Head - Elbow) - Start-Stop Difference value		0.55
Distance (Wrist - Hip) - Average Speed		0.51
Angle (Shoulder - Spine Shoulder - Spine Mid) - Max Accel.		0.44
Distance (Head - Elbow) - Variance value		0.44
Distance (Head - Hand) - Start-Stop Difference value		0.41

All Samples	Raw Data	Salute
Top 10 Ranked Features		Rank
Total time		**0.99**
Distance (Head - Elbow) - Average Speed		**0.86**
Angle (Hip - Shoulder - Elbow) - Average Speed		**0.77**
Angle (Shoulder - Spine Shoulder - Spine Mid) - Average Speed		0.70
Distance (Head - Elbow) - Start-Stop Difference value		0.68
Angle (Hip - Shoulder - Elbow) - Min Speed		0.64
Distance (Head - Hand) - Average Speed		0.64
Distance (Head - Elbow) - Max Speed		0.58
Distance (Hand - Knee) - Average Speed		0.56
Distance (Wrist - Hip) - Average Speed		0.50

We summarize the experimentation in concluding that FMA can be automatically scored using the features above, to distinguish between patients with high severity and low severity (and healthy) FMA scores.

4.4.3 *Compensation Statistical analysis of Compensation:*

Motor Compensation refers to the alternative strategies developed by stroke patients in order to compensate for their difficulty or inability to perform a motor task [41]. In the context of FMA, this is expressed as increased movement in body parts that are unrelated to the motor task, such as the movement of the spine or the shoulders during hand lifting [41]. In the current study, we analyzed the measured motion and position of stroke patient body parts to uncover motor compensation in patients during FMA.

Following the medical guidelines [41], we analyzed the following motion features in order to detect well known compensation strategies used by stroke patients when performing Hand Lift and Salute motion:

1. Elbow Angle - (Shoulder - Elbow - Wrist) -
   Tests if the hand is straight during movement.
2. Spine Angle - (Knee - Hip - Spine Mid) -
   Tests if the back is straight during movement.
3. Shoulder Distance - (Camera to Shoulder) -
   Tests if the shoulders are stable during movement.

Patients should present increased motion signals for these features when they perform motor compensations, thus, we focus on the STD values of these features along the time course of the analyzed movement. We used the T-Test [42], to evaluate each compensation feature independently. The results, given in Table 12, show that for the Hand Lift motion, patients showed increased motion in all cases, with significant results, at threshold of 0.05, in the spine angle $(t(-2.437) = 42, p = 0.019)$ and the shoulder distance $(t(-2.105) = 35.5, p = 0.042)$. For the Salute motion (Table 13), we exclude the Elbow angle, since saluting does not require maintaining a straight arm. In this case, only the spine angle is significant $(t(-2.851) = 35.43, p = 0.007)$, whereas shoulder distance shows an insignificant opposite trend.

4.5 *Discussion*

A novel multi-camera tracking system was applied to evaluating motor movement of stroke patients as part of our stroke rehabilitation project and with the goal of allowing home assessment for patients. We showed very high classification rates between stroke patients and healthy subjects using our Fugl-Meyer tracking and analysis system. In addition, the top-ranking features were found to strongly relate to the Fugl-Meyer instructions and indicate the significance of speed of motion in determining the FMA score. Two movements were tested, both in the category of upper-limb function ability. The results show that a complex movement such as the Salute is a much better indicator than a simple movement such as the Hand Lifting. Our experiment also showed that the asymmetry between movements in patients' two hands is not a distinguishing factor and it is advantageous to analyze each hand independently.

We also found that patients show higher levels of compensation than healthy individuals. These results show, for the first time, that compensation can be detected and tracked using a consumer camera and suggests that in the future, such systems will be able to track and quantify in-depth rehabilitation processes.

Our system showed results that are consistent those obtained using expensive and invasive high-end motion capture systems: our analysis showed that stroke patients move slower and take longer to

Table 10 Feature selection results - Random Forest (Part 1)

All Samples	Raw Data	Salute & Hand lift
Top 10 Ranked Features		Rank
Total time		**25.97**
Distance (Head - Elbow) - Variance value		23.37
Distance (Head - Elbow) - Start-Stop Difference value		21.00
Angle (Hip - Shoulder - Elbow) - Average Speed		**20.71**
Height (Hand) - Start-Stop Difference value		20.63
Distance (Wrist - Hip) - Average Speed		19.43
Angle (Hip - Shoulder - Elbow) - Min Speed		18.90
Distance (Head - Elbow) - Average Speed		**18.51**
Height (Wrist) - Start-Stop Difference value		17.63
Distance (Wrist - Hip) - Max Acceleration		16.89

Best choice	Repetitions avg	Salute & Hand lift
Top 10 Ranked Features		Rank
Total time		**22.50**
Angle (Hip - Shoulder - Elbow) - Min Speed		17.75
Distance (Wrist - Hip) - Max Acceleration		16.00
Distance (Head - Elbow) - Variance value		15.62
Distance (Head - Elbow) - Max Speed		14.00
Height (Hand) - Variance Speed		13.12
Distance (Head - Elbow) - Average Speed		**12.00**
Angle (Shoulder - Elbow - Wrist) - Max Acceleration		11.75
Angle (Shoulder - Spine Shoulder - Spine Mid) - Avg Speed		11.18
Height (Wrist) - Max Acceleration		11.00

All Samples	Raw Data	Hand lift
Top 10 Ranked Features		Rank
Distance (Head - Elbow) - Average Speed		**26.14**
Total time		**24.61**
Angle (Hip - Shoulder - Elbow) - Variance value		18.06
Distance (Head - Hand) - Variance value		17.03
Height (Wrist) - Start-Stop Difference value		16.23
Height (Hand) - Start-Stop Difference value		16.04
Height (Hand) - Max Acceleration		16.00
Distance (Head - Elbow) - Variance value		15.42
Height (Hand) - Variance value		14.38
Angle (Hip - Shoulder - Elbow) - Average Speed		**13.93**

Table 11 Feature selection results - Random Forest (Part 2)

Best choice	Repetitions avg	Hand lift
Top 10 Ranked Features		Rank
Total time		**17.50**
Angle (Hip - Shoulder - Elbow) - Max Acceleration		16.50
Angle (Hip - Shoulder - Elbow) - Start-Stop Difference value		14.25
Distance (Head - Elbow) - Max Acceleration		13.50
Angle (Hip - Shoulder - Elbow) - Average Speed		**10.50**
Angle (Shoulder - Elbow - Wrist) - Max Acceleration		9.75
Distance (Hand - Knee) - Max Acceleration		9.50
Distance (Wrist - Hip) - Max Acceleration		8.25
Height (Wrist) - Max Acceleration		7.25
Distance (Wrist - Hip) - Average Speed		7.00

All Samples	Raw Data	Salute
Top 10 Ranked Features		Rank
Total time		**27.28**
Angle (Shoulder - Elbow - Wrist) - Average Speed		17.25
Distance (Head - Hand) - Average Speed		16.00
Distance (Head - Elbow) - Average Speed		**15.75**
Height (Elbow) - Average Speed		15.625
Distance (Head - Elbow) - Variance value		14.50
Distance (Head - Elbow) - Start-Stop Difference value		14.00
Height (Hand) - Variance Speed		12.00
Angle (Hip - Shoulder - Elbow) - Average Speed		**11.87**
Angle (Shoulder - Spine Shoulder - Spine Mid) - Min Speed		11.75

Best choice	Repetitions avg	Salute
Top 10 Ranked Features		Rank
Total time		**20.00**
Angle (Hip - Shoulder - Elbow) - Average Speed		**16.00**
Distance (Head - Hand) - Average Speed		15.00
Distance (Head - Elbow) - Average Speed		**15.00**
Distance (Head - Elbow) - Max Speed		13.00
Angle (Hip - Shoulder - Elbow) - Min Speed		12.00
Distance (Head - Elbow) - Variance value		9.75
Distance (Head - Elbow) - Min value		9.50
Distance (Head - Elbow) - Max Acceleration		8.00
Angle (Shoulder - Spine Shoulder - Spine Mid) - Avg Speed		7.75

perform a motor task compared to healthy subjects. This corresponds with [43] where infrared light emitting diodes (IREDS) were used invasively to show this effect. Our findings (Tables 12 and 13) also show increase in trunk flexion (spine motion) in patients attempting to move their hand to the target position compared to healthy subjects. This was found in [44] by using an optical motion analysis system, where eight infrared emitting diodes (IREDs) were placed on body landmarks of the hand, arm and trunk.

The high classification rate between stroke patients and healthy subject and the consistency with high-end systems show that, with additional effort, our system is suitable for stroke rehabilitation quantification from the patient's home. Additional effort is needed in developing a dedicated user interface for a system operated by unprofessional end users.

Table 12 Compensation levels in Stroke Patients and Healthy subjects for the hand lift movement

	Elbow angle	Spine angle	Shoulder angle
Patient	8.80±.88	3.37±.31	0.03±.004
Healthy	7.96±.68	2.44±.20	0.02±.002
Sig (2-tailed)	0.467	0.019	0.042

Table 13 Compensation levels in Stroke Patients and Healthy Subjects for the hand salute movement

	Spine angle	Shoulder distance
Patient	2.47±.29	0.071±.01
Healthy	1.53±.16	0.074±.01
Sig (2-tailed)	0.007	0.879

5 Conclusions and Future Work

In this research we introduced a novel multi-camera human tracking system. The system is inexpensive, portable and marker-less. System calibration is adaptive and performed on the fly based on the human body being tracked, and so requires no calibration session nor special calibration equipment. Thus the system is well suited for home use and for tele-medicine applications. The system performs run time merging of the skeleton data, resulting in a more reliable and stable skeleton representation. 3D point cloud alignment and merging can be performed as well to form a full 3D body representation. We show excellent performance of the calibration algorithm (less than 2cm) and of the skeleton merging (less than 1.7cm std in our measure of "Skeleton" bone length).

Finally, through our testing, we reached several conclusions and practical recommendation:
More cameras are better, Circular configurations (360° coverage of scene) is preferred over neighboring configurations, the best skeleton merging method involves frontality with simple merging (choose the single best frontal view) for configurations with few cameras, and frontality weighted averaging for configurations with a larger number of cameras. Thus camera setup should be designed to maximize frontality while maintaining a circular configuration.

In this study, the developed multi-camera system was applied to evaluating motor movement of stroke patients as part of our stroke rehabilitation project and with the goal of allowing home assessment for patients. We showed very high classification rates between stroke patients and healthy subjects using our Fugl-Meyer analysis application. In addition, the top ranking features were found to strongly relate to the Fugl-Meyer instructions and indicate the signifcance of speed of motion in determining the FMA score.

Two movements were tested, both in the category of upper-limb function ability. The results show that a complex movement such as the salute is a much better indicator than a simple movement such as the hand lifting. Our experiment also showed that the asymmetry between movements in patients two hands is not a distinguishing factor and it is better to analyse each hand independently.

Future medical studies will extend the classification capabilities to distinguish between the three known stroke severities. This will, eventually, demonstrate that the Fugl-Meyer Assessment can be performed automatically without a need for a physician to be present and in a home setting using our flexible multi-camera motion capture system.

6 References

1 A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. a method for evaluation of physical performance," *Scandinavian journal of rehabilitation medicine*, vol. 7, no. 1, pp. 13–31, 1975.

2 VICON, "Motion capture system homepage." http://www.vicon.com, accessed May 2018.

3 OPTOTRAK, "Motion capture system homepage." https://www.ndigital.com/msci/products/optotrak-certus/, accessed May 2018.

4 IPIsoft, "Motion capture system homepage." http://www.ipisoft.com/, accessed May 2018.

5 Microsoft, "Kinect for xbox one." http://www.xbox.com/en-US/xbox-one/accessories/kinect, accessed May 2018.

6 J. Shotton, A. Fitzgibbon, M. Cook, *et al.*, "Real-time human pose recognition in parts from single depth images," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, IEEE, 2011.

7 J. Shotton, T. Sharp, A. Kipman, *et al.*, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

8 T. Mallick, P. P. Das, and A. K. Majumdar, "Characterizations of noise in kinect depth images: A review," *IEEE Sensors journal*, vol. 14, no. 6, pp. 1731–1740, 2014.

9 B. Choo, M. Landau, M. DeVore, and P. A. Beling, "Statistical analysis-based error models for the microsoft kinect depth sensor," *Sensors*, vol. 14, no. 9, pp. 17430–17450, 2014.

10 D. Falie and V. Buzuloiu, "Noise characteristics of 3d time-of-flight cameras," in *International Symposium on Signals, Circuits and Systems (ISSCS).*, vol. 1, pp. 1–4, IEEE, 2007.

11 A. Fetić, D. Jurić, and D. Osmanković, "The procedure of a camera calibration using camera calibration toolbox for matlab," in *MIPRO, 2012 Proceedings of the 35th International Convention*, pp. 1752–1757, IEEE, 2012.

12 B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease," *Gait & posture*, vol. 39, no. 4, pp. 1062–1068, 2014.

13 Š. Obdržálek, G. Kurillo, F. Ofli, *et al.*, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *Engineering in medicine and biology society (EMBC), 2012 annual international conference of the IEEE*, pp. 1188–1193, IEEE, 2012.

14 R. A. Clark, Y.-H. Pua, A. L. Bryant, and M. A. Hunt, "Validity of the microsoft kinect for providing lateral trunk lean feedback during gait retraining," *Gait & posture*, vol. 38, no. 4, pp. 1064–1066, 2013.

15 Q. Wang, G. Kurillo, F. Ofli, and R. Bajcsy, "Evaluation of pose tracking accuracy in the first and second generations of microsoft kinect," in *Healthcare Informatics (ICHI), 2015 International Conference on*, pp. 380–389, IEEE, 2015.

16 B. Bonnechere, B. Jansen, P. Salvia, *et al.*, "Validity and reliability of the kinect within functional assessment activities: comparison with standard stereophotogrammetry," *Gait & posture*, vol. 39, no. 1, pp. 593–598, 2014.

17 M. Plotnik, S. Shema, M. Dorfman, , *et al.*, "A motor learning-based intervention to ameliorate freezing of gait in subjects with parkinsons disease," *Journal of neurology*, vol. 261, no. 7, pp. 1329–1339, 2014.

18 Funaya, Shibata, Wada, and Yamanaka, "Accuracy assessment of kinect body tracker in instant posturography for balance disorders," in *2013 7th International Symposium on Medical Information and Communication Technology (ISMICT)*, pp. 213–217, 2013.

19 T. T. Liu, C. T. Hsieh, R. C. Chung, and Y. S. Wang, "Physical rehabilitation assistant system based on kinect," in *Applied Mechanics and Materials*, vol. 284, pp. 1686–1690, Trans Tech Publ, 2013.

20 D. S. Alexiadis, P. Kelly, P. Daras, N. E. O'Connor, T. Boubekeur, and M. B. Moussa, "Evaluating a dancer's performance using kinect-based skeleton tracking," in *Proceedings of the 19th ACM international conference on Multimedia*, pp. 659–662, ACM, 2011.

21 A. Kitsikidis, K. Dimitropoulos, S. Douka, and N. Grammalidis, "Dance analysis using multiple kinect sensors," in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, vol. 2, pp. 789–795, IEEE, 2014.

22 J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

23 A. Maimone, J. Bidwell, K. Peng, and H. Fuchs, "Enhanced personal autostereoscopic telepresence system using commodity depth cameras," *Computers & Graphics*, vol. 36, no. 7, pp. 791–807, 2012.

24 S. Asteriadis, A. Chatzitofis, D. Zarpalas, D. S. Alexiadis, and P. Daras, "Estimating human motion from multiple kinect sensors," in *Proceedings of the 6th international conference on computer vision/computer graphics collaboration techniques and applications*, p. 3, ACM, 2013.

25 P. Otten, J. Kim, and S. H. Son, "A framework to automate assessment of upper-limb motor function impairment: A feasibility study," *Sensors*, vol. 15, no. 8, pp. 20097–20114, 2015.

26 J. Wang, L. Yu, J. Wang, L. Guo, X. Gu, and Q. Fang, "Automated fugl-meyer assessment using svr model," in *Bioelectronics and Bioinformatics (ISBB)*, pp. 1–4, IEEE, 2014.

27 W. S. Kim, S. Cho, D. Baek, H. Bang, and N. J. Paik, "Upper extremity functional evaluation by fugl-meyer assessment scoring using depth-sensing camera in hemiplegic stroke patients," *PloS one*, vol. 11, no. 7, p. e0158640, 2016.

28 OpenKinect, "Project homepage." https://openkinect.org, accessed May 2018.

29 W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica*, vol. 32, no. 5, pp. 922–923, 1976.

30 W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica*, vol. 34, no. 5, pp. 827–828, 1978.

31 M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

32 P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611, pp. 586–607, International Society for Optics and Photonics, 1992.

33 Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 1, pp. 666–673, IEEE, 1999.

34 O. Wasenmüller and D. Stricker, "Comparison of kinect v1 and v2 depth images in terms of accuracy and precision," in *Asian Conference on Computer Vision*, pp. 34–45, Springer, 2016.

35 N. Eichler, H. Hel-Or, I. Shimshoni, D. Itah, B. Gross, and S. Raz, "Non-invasive motion analysis for stroke rehabilitation using off the shelf 3d sensors," in *Neural Networks (IJCNN), 2018 International Joint Conference on Neural Networks*, IEEE, 2018.

36 E. J. Benjamin, M. J. Blaha, S. E. Chiuve, *et al.*, "Heart disease and stroke statistics-2017 update: a report from the american heart association," *Circulation*, vol. 135, no. 10, pp. e146–e603, 2017.

37 N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

38 J. R. Quinlan, "C4.5: Programming for machine learning," *Morgan Kauffmann*, vol. 38, p. 48, 1993.

39 T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

40 S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

41 M. F. Levin, J. A. Kleim, and S. L. Wolf, "What do motor ŞrecoveryŤ and ŞcompensationŤ mean in patients following stroke," *Neurorehabilitation and neural repair*, vol. 23, no. 4, pp. 313–319, 2009.

42 J. F. Box, "Guinness, gosset, fisher, and small samples," *Statistical science*, vol. 2, no. 1, pp. 45–52, 1987.

43 M. Cirstea and M. F. Levin, "Compensatory strategies for reaching in stroke," *Brain*, vol. 123, no. 5, pp. 940–953, 2000.

44 S. M. Michaelsen, S. Jacobs, A. Roby-Brami, and M. F. Levin, "Compensation for distal impairments of grasping in adults with hemiparesis," *Experimental Brain Research*, vol. 157, no. 2, pp. 162–173, 2004.