Robust Tracking of the Upper Limb for Functional Stroke Assessment

Sonya Allin, Nancy Baker, Emily Eckel, and Deva Ramanan

***Abstract*—We present a robust three-dimensional parts-based (PB) tracking system designed to follow the upper limb of stroke survivors during desktop activities. This system fits a probabilistic model of the arm to sequences of images taken from multiple angles. The arm model defines shapes and colors of limbs and limb configurations that are more or less likely. We demonstrate that the system is 1) robust to cluttered scenes and temporary occlusions; 2) accurate relative to a commercial motion capture device; and 3) capable of capturing kinematics that correlate with concurrent measures of post stroke limb function. To evaluate the PB system, the functional motion of 7 stroke survivors was measured concurrently with the PB system and a commercial motion capture system. In addition, functional motion was assessed by an expert using the Fugl-Meyer Assessment (FMA) and related to recorded kinematics. Standard deviation of differences in measured elbow angles between systems was 5.7 degrees; deviation in hand velocity estimates was 2.6 cm/s. Several statistics, moreover, correlated strongly with FMA scores. Standard deviation in shoulder velocity had a significant correlation coefficient with FMA score below -0.75 when measured with all systems.**

***Index Terms*—Computer Vision, Functional Assessment, Stroke Rehabilitation, Human Tracking**

I. Introduction

ONE of the great benefits of technology for stroke rehabilitation is its ability to objectively and repeatedly measure movement in a way that documents motor recovery. Studies with robots, for example, have quantitatively demonstrated that changes in the speed and smoothness of point to point motions correlate strongly with corresponding changes in the functional status of the upper limb [1]. Studies with commercial motion tracking systems have identified related three-dimensional functional kinematics, such as the extent to which subjects depend on the motion of their torso during a reach [2] and the amount of synergistic motion about joints [3], to discriminate post-stroke functional status. Many kinematics, such as movement speed or presence of joint synergies, have, in fact, long been encoded into evaluation criteria of common clinical observation-based limb assessments, such as the Fugl-Meyer Assessment (FMA) [4]. Technology, however, has enabled these observations to be made repeatedly, continuously and over extended periods of time.

S. Allin is with the Department of Occupational Therapy, University of Toronto, Toronto, ON M5G1V7 CANADA (phone: 416-946-8573; fax: 416-946-8570; email:s.allin@utoronto.ca).

N. Baker is with the Department of Occupational Therapy, University of Pittsburgh, Pittsburgh, PA 15261 USA (email:nab36@pitt.edu).

E. Eckel is with the Department of Occupational Therapy, Chatham University, Pittsburgh, PA 15232 USA (email:eeckel@chatham.edu).

D. Ramanan is with the Department of Computer Science, University of California at Irvine, Irvine, CA 92697 USA (email:d.ramanan@uci.edu).

Continuous quantitative measures can be useful not only to assess disability status and rehabilitation outcomes, but also to motivate and guide therapy [5]. Unfortunately, tools to make quantitative kinematic measures of gross limb function tend to be high in cost and impractical for use outside of clinics. Robots large enough to train gross reaching motions are costly, as are commercial devices that track the motion of infra-red reflective markers, such as the VICON^{TM}. Both require in-laboratory use due to the size or requirements of the equipment.

Some cost-effective, home appropriate kinematic tracking alternatives include accelerometers [6], force feedback joysticks, and computer mice [7]. While accelerometers are well designed to quantify overall levels of activity in the arm [6], reconstructed kinematics from these sensors tend to be noisy and sensitive to measurement drift [8]. Accuracy can be increased by introducing gyroscopes or ultrasound [8] at the expense of requiring more, potentially movement inhibiting, instrumentation. Joysticks and mice have comparable limitations in that they operate in confined spaces and capture movement only in one plane. Although many in-plane movement features, like smoothness of hand motion during a reach, correlate positively with motor improvement after stroke [1], features that correlate negatively, like torso compensation [2], cannot be captured directly in this way.

Such cost and configuration limitations have led several rehabilitation researchers to explore the kinematic tracking alternatives offered by computer vision. In the context of post-stroke rehabilitation of functional arm motion in particular, research has applied stereo vision to hand tracking [9] and monocular vision to tracking the gross position of the arm [10]. The potential advantages of video based kinematic tracking alternatives are many, as the measurement tools are very low in cost and relatively easily integrated into locations, like kitchens and offices, where functional arm motions routinely occur.

In the research presented here, we explore the application of parts-based (PB) kinematic tracking [11]–[15] with no dynamic assumptions to functional limb assessment and monitoring. Parts-based kinematic trackers generally represent the human body as an assemblage of parts (i.e. a head, a torso, arms, etc.), each with static properties like shape and appearance [14], [15] and/or dynamic properties like velocity and inertia [11]. Models with dynamic properties may predict a configuration of body parts at one frame given the prior configuration [15], while models with configuration and appearance properties alone may track by locating body parts on a frame by frame basis [14]. Trackers without dynamic assumptions have proven to be robust over long periods of

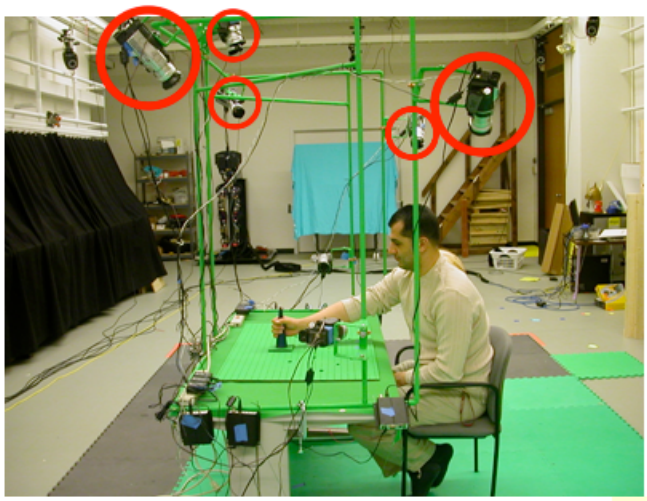

Fig. 1. System overview. The complete system is composed of 6 JVC cameras, located around a desktop. Each camera is circled in red in the image. On the tabletop are markings which indicate start and stop positions for the hands as well as objects, as required by the AMAT instruction manual.

time and in tracking situations where humans are moving in relatively unusual ways (as occurs during a baseball game, for example) [14]. As stroke survivors may make atypical motions as well, we focus on similar, dynamics-free models.

We additionally limit ourselves to development of tools that respect the following constraints. They:

1. Do not depend on a background model, i.e., they do not assume the background is static and that a subject can be detected by foreground segmentation. This is because we would like to be able to track individuals in real environments where the background may periodically and unexpectedly change (caregivers may move in the background, for example).

2. Do not depend of motion, as we would like to track individuals when they remain still. Optic flow based trackers [16] specifically model motion and may therefore experience difficulty tracking people who do not move. Similarly, trackers that depend on adaptive background models [17] risk associating an individual who does not move with the background as time goes by.

3. Are tolerant to occlusions, as we expect individuals periodically to be blocked from view by family members or caregivers in real environments.

4. Recover robustly from error, as, even in controlled situations, tracking errors will likely periodically take place.

Our parts-based (PB) tracking system avoids background models by creating strong models of the human in the foreground. These models make no dynamic assumptions and are capable of tracking individuals who do not move. Tolerance to occlusions is facilitated by the use of multiple cameras. In addition, our system not only detects humans but also assigns each detected configuration of limbs a probability. The assignment of a probability allows us to eliminate views that are poor (i.e. low in probability, given the model) before limb reconstruction takes place. Finally, our PB tracking system has been designed to detect the limbs independently at each frame, which facilitates recovery from error. The system as it currently stands consists of 6 synchronized cameras stationed about a desktop, and is illustrated in Figure 1.

In this paper we describe our system and an experiment in which the kinematics measured with a PB tracker are 1) compared to a standard commercial motion tracking system; and 2) associated with concurrent measures of functional limb disability, as measured by the FMA, in stroke survivors.

II. Parts Based Tracking System Algorithm

A. Overview

Like the authors of [14], [15], we model the upper body in two dimensional images as a configuration of segments, each with distinct shape and appearance properties. Our current model consists of three segments per arm, with one segment at the upper arm, one at the forearm and one at the hand. Shapes of segments are defined as ellipses, colors by red, green and blue (RGB) values, and configurations by relative positions and orientations of adjacent body parts. Training the various models requires a human to segment bodies in example images before tracking begins. When tracking takes place, shape, color and configuration information is combined to determine likely arm locations in images. Each stage in the training and tracking process is covered below.

B. Training limb appearances and shapes

Prior to use, the PB tracker system must be trained once for each subject to identify the appearance of limb segments. To ensure robust appearance models, in initial experiments we asked subjects to wear a colorful shirt while they were tracked. Our experience indicates individuals may be willing to wear such an outfit during a therapy session should it demonstrably enhance rehabilitation. Alternative and markerless appearance models, however, exist in the literature; these may characterize appearances based not only on limb color but on texture, size or edge information [13]–[15]. Additional work explores automatic acquisition of appearance models [14], thereby removing the need to train for appearances altogether. Our initial experiments can be seen as a best case scenario for such appearance modeling techniques.

To train limb colors, a human operator provides examples of image regions that correspond to segments of the shirt (i.e., upper arm and forearm) and hands. In experiments, examples were images of each subject as he or she was seated with arms resting on a tabletop. One image was selected from each camera angle.

With image regions selected, a quadratic logistic regression is tuned to the color of each body segment. More specifically, the likelihood a pixel belongs to a particular arm segment is defined as a function of its RGB colors. Let y be the label of a pixel (that indicates it belongs to the hand, forearm or upper arm). Then:

$$P(y|R,G,B) = 1/(1 + exp(-w_y * \phi(R,G,B)) \quad (1)$$

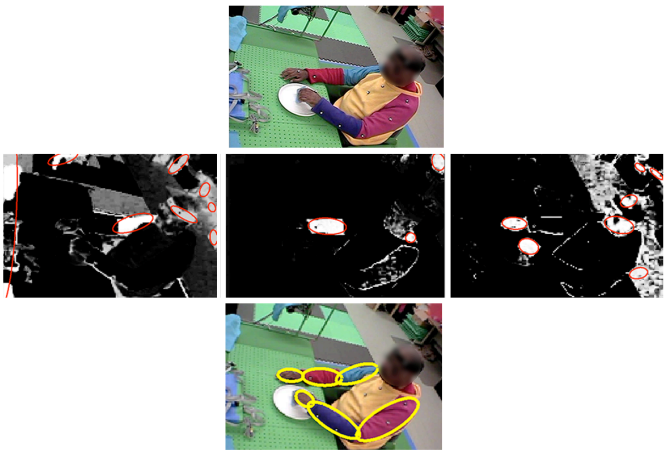

Fig. 2. Detecting candidate parts. At the top is the original image. In the next row, the three images from left to right are images of pixels detected, via logistic regression, as likely to belong to 1) the upper arm, 2) lower arm and to 3) skin (the hand). Candidate parts have been circled in red. Finally, at the bottom of the image is the combination of hands, lower and upper arms determined to be most likely according to the appearance and configuration model.

Here, ϕ is a function which returns a vector that consists of the original RGB values for a given pixel in addition to their squares, cross terms (i.e., $R*G$, $G*B$ and $B*R$) and an offset term (i.e. 1). The vector w_y is made of coefficients for each of these terms. During training, values for each $P(y)$ are set to 1 if the pixels belong to a segment of interest and to 0 otherwise; the best fitting w_y is then learned. During tracking, trained coefficients for each segment (w_y) are used to map incoming RGB pixels onto the [0,1] range.

Training for likely shapes makes use of the same hand segmentations used for color modeling. More specifically, segmented parts are parameterized as ellipses, and the center, minor and major axis lengths are stored. During tracking, this shape information is compared with the shapes of candidate parts, as is described in Section D.

C. Tracking step 1: Detecting candidate parts

The first step in tracking involves detection of candidate body parts. To do this, we apply the trained color sensitive logistic curves to each image. Values are then thresholded at 0.75; pixels with values lower than this are discarded. Remaining pixels are collected into connected groups and groups that are less than 15 pixels in size are removed. Groups of more than 15 pixels are then parameterized as ellipses according to their location, size and orientation. More specifically, for a body segment like the hand, we detect i groupings of pixels and parameterize each of these groupings as $eh_i = \{x_i, y_i, major_i, minor_i, orientation_i\}$. Here, $major_i$ and $minor_i$ are the lengths of axes of the best fit ellipse surrounding the group of pixels and x_i, y_i are its center. We locate and parameterize additional sets of ellipses for the other body parts (i.e. sets of eu_i for the upper arm and ef_i for the forearm) in the same way. Examples of candidate parts for various segment types are illustrated in Figure 2.

D. Tracking step 2: Using the model to locate likely arms.

Certain configurations of ellipsoids are more likely than others to correspond to an arm. We know, for example, that one's forearm is usually close to his or her upper arm and that the elbow is unlikely to hyperextend. To encode such assumptions requires us to model, in addition to shape and color, likely configurations of arm segments. This configuration model is based not only on limb segment dimensions that are specific to each individual, but on general anthropometric information about the human arm.

More formally, we say that the probability a collection of ellipsoids corresponds to an arm depends upon the appearance of each ellipsoid (i.e., if it is the right color and shape) and their relative orientation (i.e., if they are close together, far apart, etc.). In addition, we claim conditional independence of parts that are not immediately adjacent to one another, assuming that we are given the intervening parts. The location of the right hand, then, is considered conditionally independent of the location of the right upper arm, assuming we know the location of the right forearm.

Let eh, ef and eu be random variables associated with ellipses for the true hand, forearm and upper arm in an image, respectively. We score the arm-ness of any set of three ellipses with the following probability model:

$$\begin{aligned} P(eh = eh_i, ef = ef_j, eu = eu_k | im) \propto \\ P(eh = eh_i, ef = ef_j, eu = eu_k) * \\ P(im | eh = eh_i, ef = ef_j, eu = eu_k) \end{aligned} \tag{2}$$

Here, im represents the particular image data being analyzed. The proportionality sign in the equation ensures the right hand side is a proper probability distribution over ellipse assignments (i.e., that it integrates to one).

The first term on the right hand side of the equation scores the spatial geometry of ellipses, with a higher score being given to "valid" hand, forearm, and upper-arm configurations that consist of appropriately shaped ellipses. More specifically:

$$\begin{aligned} P(eh = eh_i, ef = ef_j, eu = eu_k) = \\ P(eh = eh_i | ef = ef_j) * \\ P(ef = ef_j | eu = eu_k) * P(eu = eu_k) \end{aligned} \tag{3}$$

Where:

$$\begin{aligned} P(eh = eh_i | ef = ef_j) = 0 \\ P(ef = ef_j | eu = eu_k) = 0 \end{aligned} \tag{4}$$

if the pairing of adjacent ellipses is "invalid", else:

$$\begin{aligned} P(eh = eh_i | ef = ef_j) = N_{hf}(d(eh_i, ef_i) | \mu_{hf}, \sigma_{hf}) \\ P(ef = ef_j | eu = eu_k) = N_{fu}(d(ef_j, eu_k) | \mu_{fu}, \sigma_{fu}) \end{aligned} \tag{5}$$

We define "valid" parings of adjacent ellipses as those whose centers: 1) do not overlap; 2) are less than an arm segments length away and, 3) for the forearm and upper arm (i.e., eu and ef), claim the lower arm to be below the upper arm (in the front camera views) or to the one side (in the lateral views). For "valid" pairings, $P(eh = eh_i | ef = ef_i)$ and $P(ef = ef_j | eu = eu_k)$ represent the likelihood a particular ellipse (i.e. eh_i, or ef_j) corresponds to a true part (i.e. eh or

ef) given the parts adjacent to it in the model. The likelihood functions, N_{hf} and N_{fu}, are functions of the x and y distances between the centers of adjacent ellipses (i.e. d). Distances are computed along the major and minor axes of parent ellipses in the model, i.e. along the axes of ef_j or eu_i. More specifically, N_{hf} and N_{fu} are normal distributions centered at μ_{hf} and μ_{fu}, the distances between adjacent parts corresponding to eh and ef or ef and eu in the training images. In our experiments, the variance on distributions, σ_{hf} and σ_{fu}, was set to 25; this empirically was found to suppress part pairings that were too close or far apart.

The second term favors ellipses whose local image evidence is consistent with color and shape models learned during training. We assume this portion of the score factors into a product of image evidence for each body part, i.e.:

$$P(im|eh = eh_i, ef = ef_j, eu = eu_k) = \prod_i P(im|eh_i) \prod_j P(im|ef_j) \prod_k P(im|eu_k) \quad (6)$$

We further assume image evidence for each part factors into shape and appearance terms. This same factorization applies to $P(im|eh = eh_i)$, $P(im|ef = ef_j)$ and $P(im|eu = eu_k)$. We write the factorization for $P(im|eh = eh_i)$ below, and note that the equations for $P(im|ef = ef_j)$ and $P(im|eu = eu_k)$ are identical, but with alternate mean and variance terms:

$$P(im|eh = eh_i) = Nshape_h(eccen(eh_i)|\mu s_h, \sigma s_h) * Ncolor_h(color(eh_i)|\mu c_h, \sigma c_h) \quad (7)$$

Here, $eccen(eh_j)$ is the eccentricity of the ellipse (i.e. the ratio of its two axes length), and $color(eh_i)$ is the average value of part-specific logistic color response within the ellipse. $Nshape_h$ is a normal distribution whose mean (μs_h) lies at the ratio of axis lengths of for the hand segment in the training images. $Ncolor_h$ is also a normal distribution, whose mean (μc_h) lies at the maximum color response (i.e. at 1). $Nshape_f$ and $Nshape_u$ are centered at the corresponding ratios in the training images. For all segments, σs_h, σs_f and σs_u were set to 1 while σc_h, σc_f and σc_u were set to 0.1, as these led to empirically good results. In practice, this function favors upper and lower limb ellipses that are highly eccentric, hand ellipses that are round and parts that have a high average color response.

Using this model allows us to assign a probability value, ranging from 0 to 1, to all combinations of detected potential hands, forearms and upper arms. Finding the most "arm like" combination of a hand, forearm and upper arm then involves finding the single combination of ellipses that have the highest probability value based on their shape, configuration and color. To do this, a combinatorial search is executed across all detected ellipses; this search made slightly more efficient via dynamic programming. Figure 2 illustrates not only candidate limb segments, but the highest probability configuration of parts, selected using the shape and configuration models.

The part detection method, as it stands, is robust to moderate changes in the scale of the arms in an image. A lower bound on the size of a detected part is determined by color segmentation operations, which discard groups of pixels that number less than 15. An upper bound on the size of detected arms is determined by the configuration constraints, which discard configurations in which parts are significantly farther than a trained "arm's length" from one another. "Significantly farther" is defined by the variance terms σh_f and σf_u, which are currently set at 25 pixels.

E. Tracking step 3: Three Dimensional Triangulation

The last step in the reconstruction process involves triangulating ellipses across each two dimensional view to make a complete three dimensional reconstruction of the upper arm. Before triangulation, however, we use the probability values from the prior step to discard spurious arm detections. Examples of probability values assigned to arm configurations from one view are illustrated in Figure 3. In the figure, probabilities associated with the left arm configuration are indicated in green, while those for the right are indicated in red. When the right arm is partially occluded, its net probability drops. When the left arm is self occluded it is not correctly detected and its net probability drops. We have found good reconstructions result when the three highest scoring arms in available views are selected for triangulation. The addition of penalties to part data from views unlikely to image particular arm parts well also improves results slightly (views lateral to the body, for example, are not likely to capture the opposing upper arm).

Finally, across high scoring views we identify point correspondences. To do this, four points are selected to represent each two dimensional ellipse; these lie on ellipse contours at the extremes of major and minor axes. Across initial images, point correspondences are determined by selecting the ordering that minimizes re-projection error. In subsequent images, point orderings are selected that are closest in proximity to point projections in the previous frame.

We use the following equation, which can be solved using standard least squares techniques, to triangulate points [20]:

$$\sum_{c \in cameras} (Id - u_c u_c')P = \sum_{c \in cameras} (Id - u_c u_c')center_c \quad (8)$$

Here, c is an index that refers to a particular camera view, P is a reconstructed 3D point, $center_c$ is the cth cameras center and u_c is the direction of the ray extending from the camera's center to the point. Id is the identity matrix.

Examples of reconstructed hand velocity and elbow flexion measurements are found in Figure 4. Raw wrist positions from both the PB system and a commercial motion capture system are also illustrated in this figure. From such reconstructions, several movement kinematics can be estimated in three dimensions, including movement smoothness, path lengths, average speeds, etc.

F. Summary

The tracking method we have described has a number of features that are advantageous for use in environments outside of laboratories and other controlled environments. For one, we do not depend on a constant or static background for images. As illustrated in Figure 2, although backgrounds are controlled

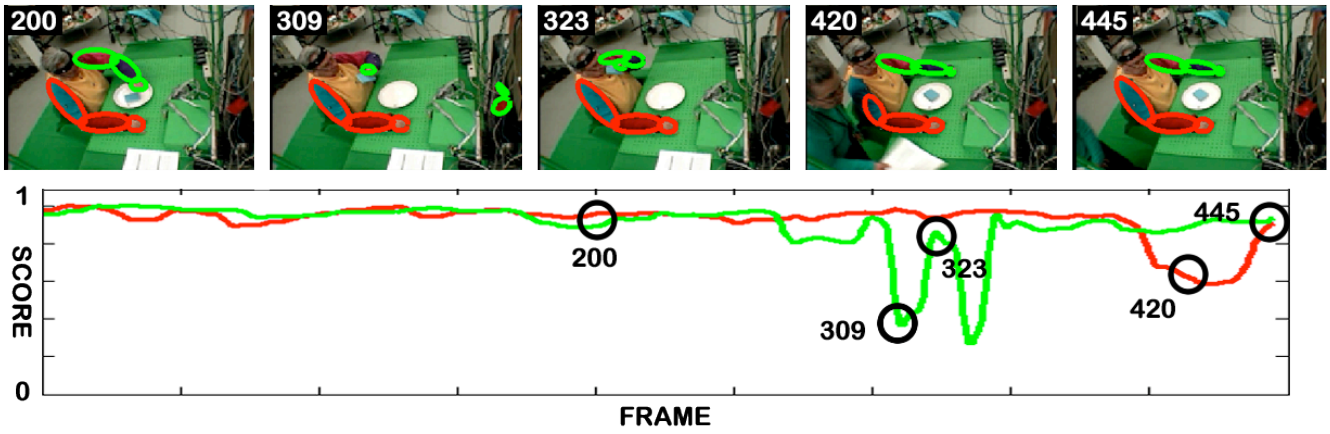

Fig. 3. Probability values assigned to arm configurations from one view. Arm configurations are at the top of the figure and probability scores for the configurations are at the bottom. Probabilities associated with the left arm are indicated in green, while those for the right are in red. When the right arm is partially occluded, its net probability drops. When the left arm is self occluded, its net probability drops. Drops can be used to detect and discard spurious configurations before reconstruction.

in our experimental environment they still feature some clutter, and yet detection of high quality (or highly probable) arm configurations is still possible. In the great majority of homes and offices, we can expect backgrounds with clutter on a day to day basis. A tracking paradigm that is robust to these conditions, then, is essential.

In addition, because we make use of multiple cameras and use probability to weight detected arms in various views, we are able to generate reconstructions even in the face of temporary occlusions. As illustrated in Figure 3, an occlusion results in a situation where detected arms in some images have low probability values assigned to them; these bad views can be detected and discarded before reconstruction takes place. In homes and offices, we can expect similar occlusions to take place due to the motion of family members or caretakers, for example. In our own experiments, an attending therapist often obstructed cameras in order to assist subjects, and yet tracking continued. We have found a system with at least three high quality views of arm motion yields good reconstructions in situations where such occlusions take place.

Third, because we are detecting and assembling parts in each image independently and make use of no dynamic priors, we are capable of quickly recovering from tracking errors, like the occlusion illustrated in Figure 3. We are also insensitive to moderate changes in the scale of arms in images and, because we do not track features that depend on motion like optic flow [16], we can track the body when it remains still.

III. Evaluation

A. Evaluation metrics

Two evaluation metrics were employed. The first was designed to validate the PB system against measures of kinematics from a commercial motion capture system. The second was designed to explore the potential for the PB system to capture clinically meaningful information. This second evaluation involved comparing kinematic measures from each subject to "ground truth" measures of functional health.

Metric 1: *accuracy relative to commercial motion capture*

To evaluate raw accuracy, absolute and relative measurements from the PB system and a 12 camera infra-red tracking (VICON^{TM}) system were compared. Relative measures included elbow angle reconstructions as well as hand and shoulder velocity estimates. Absolute measures related position estimates of the wrist, elbow and shoulder from either system. All motion used to make evaluations related explicitly to the functional activity of stroke survivors.

In the PB system, position and angle estimates were based on motion of the most proximal point on the reconstructed upper arm on the hemiparetic side (lying close to one acromion), the most distal point on the reconstructed forearm, and the point closest to axes of upper and lower arm (lying roughly at the elbow). Elbow flexion was defined as the angle formed by these points. When using the VICON^{TM}, corresponding position and angle estimates were determined by the motion of IR markers located at the corresponding acromion process, the medial epicondyle of the humerus and the lateral epicondyle of the ulna. Hand velocities, in the PB system, were based on the motion of the center of reconstructed hemiparetic hands. In the VICON^{TM}, hand velocities were based on the motion of an IR reflective marker on the metacarpo-phalangeal joint of the index finger. Finally, VICON^{TM} measures of shoulder velocity were based on the movement of the marker on the acromion while the corresponding point in the PB system was the most proximal point on the reconstructed upper arm.

To compare the relative measures, we related average deviations between the commercial and PB tracking devices as well as standard deviations in measurement discrepancies. Before standard deviations were computed, however, average errors were subtracted to remove measurement bias between the two systems. To compare the absolute measures, each reconstructed PB arm was first aligned with the same arm as measured by the motion capture system. This was done by estimating, using least squares, an affine transformation

TABLE I
SUBJECT CLINICAL DETAILS

Subject	Age	YPS	Side	Dom	FMA
1	75	2	L	R	64
2	60	1.5	R	R	44
3	47	22	L	R	46
4	82	12	R	R	53
5	58	13	L	R	64
6	78	7	R	R	39
7	63	35	R	R	64

'YPS' is 'years post-stroke'. 'Dom' is dominant side, 'Side' is lesion side and 'FMA' is Fugl -Meyer Assessment score (upper extremity only).

that minimized position differences between the two systems across shoulder, elbow and wrist estimates. Applying this transformation adjusted for bias introduced 1) during the calibration process (i.e., due to discrepancies between the global coordinate frames of either imaging system) and 2) as a result of discrepancies between the IR marker locations (which were on joint protrusions) and PB positions (which were located along axes of reconstructed arm cylinders, not arm surfaces). After alignment, mean absolute differences between corresponding positions were calculated, as well as standard deviations and root mean squared errors (RMSEs).

Metric 2: *validity*

Evaluation of clinical measurement validity was done by associating measures with a concurrent "ground truth" assessment of arm function after stroke. The functional measure chosen to ground assessments was the upper extremity (UE) portion of the Fugl-Meyer Assessment (FMA); this is a common, comprehensive, reliable test of individuals' capacity to move within and without synergetic muscle couplings that are common after stroke [4]. Scores on the UE portion of the FMA range from 0 (indicating no ability to move) to 66 (indicating an ability to move without synergistic muscle activity interference).

Univariate Pearson's correlations were used to relate kinematic summary statistics, including means and standard deviations, to FMA scores. T-tests were used to test for correlations that significantly deviated from zero; thresholds for significant deviation in these correlations were set at 0.05.

B. Evaluation methods

Arm motions of seven chronic stroke survivors were recorded with the 12 camera VICON^{TM} system and the PB tracker of our design. Four of the stroke survivors were hemiparetic on the left side and three were hemiparetic on the right. Participants were on average 66 years old (SD = 13) and 13 years post stroke (SD = 12). All subjects were assessed on the UE FMA by a licensed occupational therapist just prior to the kinematic recording sessions. Time to rest was provided between FMA assessments and the time of the recordings. Basic clinical details are in Table I.

The PB tracker was constructed with six JVC GRDV camcorders. Calibration of intrinsic parameters for the video capture system was done using a checker-board calibration grid and Matlab's Camera Calibration Toolbox [21]. Approximately 40 grid images were used to calibrate each camera; square image pixels and an absence of skew were assumed. The process estimated a focal length, principal point, and radial distortion coefficients for each of the six cameras. Extrinsic parameters were determined by placing the checkerboard directly on the desktop and capturing a single image of it from all cameras simultaneously. A rotation and translation of each camera with respect to this grid was then solved using Toolbox routines. After calibration, re-projection error across all cameras was under 2 pixels in both x and y directions. To calibrate the VICON^{TM}, VICON^{TM}'s proprietary software was used. The world origin for the VICON^{TM} was established at a corner of the desktop. Each camera had one CCD camera, recorded video at 30 Hz and at a resolution, after down-sampling, of 180 by 120 pixels. Video was encoded in real time in MPEG4 format and stored to disk.

Before recordings, subjects were asked to put on a shirt with a distinct color on each arm segment. IR reflective markers used by the VICON^{TM} device were then attached to each acromion process, lateral epicondyle of the humerus, ulnar epicondyle and second IP joint. The VICON^{TM} recorded motion at a rate of 60 Hz; these data were later down-sampled to match the temporal resolution of the video cameras.

The task subjects were asked to perform was drawn from an explicitly functional test of arm movement called the Arm Motor Ability Test (AMAT) [18]; this task involved lifting a sandwich, made of foam, from a plate on a table to the lips. The task was performed with the arm most affected by hemiparesis. The sandwich was served on a plate that was in line with the midline of subjects bodies, as specified by AMAT instructions. The center of the plate was 21 cm from the front edge of the table. Subjects started with their torsos touching the back of the chair, elbows at roughly 90 degrees, and hands laying flat on the desktop. The task was performed at the command of an attending therapist and at a comfortable movement speed. Once the task was completed, subjects were asked to return their hands to their starting locations. After demonstration of the motion by a therapist, a single lift was recorded from each subject.

To synchronize data acquired with the two tracking systems, a 5V pulse from a Data Translation board was sent simultaneously to the VICON^{TM} and to red LEDs in the field of view of each video camera. Light blinks were later detected in each view and corresponding video segments extracted for comparison with the motion capture data. Electronic pulses were generated at the push of a button.

After kinematic reconstruction and down-sampling of VICON^{TM} data, movement data were smoothed with a second-order Butterworth filter (with a cutoff frequency of 2 Hz). Then, movement data were segmented based on the computed velocity of the hemiparetic hand. The beginning of motion was said to correspond to the instant in time when hand velocity exceeded 5% of its maximum, and the end of motion was said to correspond to the last point in time when velocity exceeded this threshold. Means and standard deviations in elbow angle, hand and shoulder velocity estimates were computed from these motion segments and related to

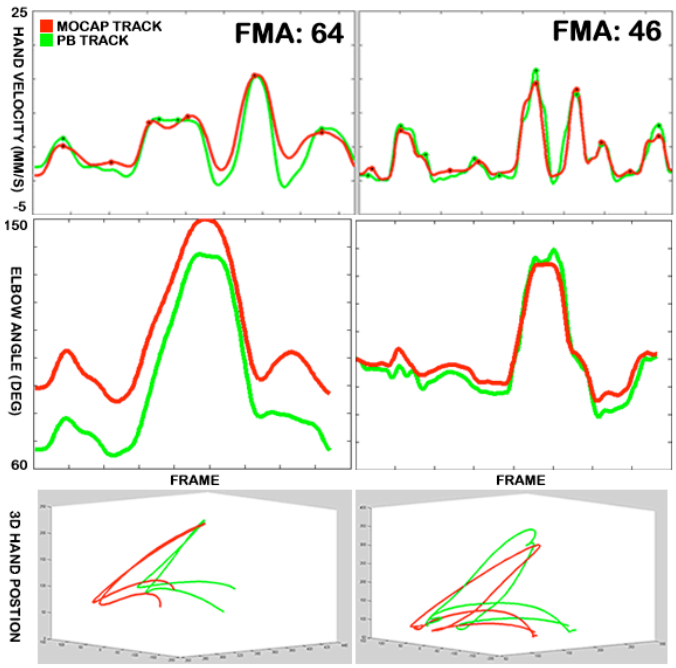

Fig. 4. From the top row to the bottom: reconstructed hand velocities, elbow angles and raw wrist positions for two subjects. Detected peaks in the hand velocity profiles have been indicated with black dots. Motion capture reconstructions are red and PB tracker reconstructions are green. Statistics for an individual with a relatively high FMA score are at the left; those for a relatively low scoring individual are at right.

TABLE II
COMPARISON BETWEEN MOTION CAPTURE & VIDEO MEASURES

Statistic	mean error (mocap-video)	standard dev.
shoulder velocity	0.30 cm/sec	0.50 cm/sec
hand velocity	1.80 cm/sec	2.57 cm/sec
elbow flexion	10.9 degrees	5.70 degrees
shoulder position	25.78 mm	33.20 mm
elbow position	19.37 mm	17.02 mm
hand position	23.97 mm	24.65 mm

FMA scores. Standard deviations were used as slightly more robust measures of range of motion than differencing extreme joint positions. Standard deviations reflect more individual measurements and therefore may be relatively tolerant to noise. In addition, the number of peaks in hand velocity profiles was also computed for both systems and related to FMA scores. Velocity peaks have been computed in prior studies as a robust measure of movement smoothness for both healthy subjects [24] as well as stroke survivors [25]. Fewer peaks represent a motion that has fewer periods of acceleration and deceleration, and is therefore relatively smooth. As in [24], peaks were defined at time points where a given velocity was high relative to neighboring time points. Hand velocities as well as detected peaks for two recoded subjects are illustrated in Figure 4.

IV. RESULTS

Metric 1: *accuracy relative to commercial motion capture*

In Table II we report means and standard deviations in differences between velocity, angle and position estimates from both systems. These deviations were computed across all subject data. In the same table, we report means and standard deviations in the absolute differences between positions of the shoulder, elbow and wrist. These statistics were computed after estimating and applying an affine transformation to align points from the PB system with those from the VICON^{TM}, as explained in Section III.

TABLE III
CORRELATION BETWEEN SUMMARY STATISTICS & FMA SCORES (UE).

Statistic	Video (p-val)	VICON^{TM} (p-val)
mean shoulder velocity	-0.62 (0.13)	-0.71 (0.07)
std shoulder velocity	-0.76 (0.05)*	-0.81 (0.03)*
mean hand velocity	0.33 (0.46)	0.57(0.18)
std hand velocity	0.20 (0.66)	0.81(0.03)*
mean elbow flexion	-0.55 (0.25)	-0.20 (0.66)
std elbow flexion	0.67 (0.09)	0.83 (0.02)*
peaks in hand velocity	-0.73 (0.06)	-0.76 (0.05)*

P-values are in parentheses; those ≤ 0.05 have an asterisk.

The custom tracker represented elbows as 11 degrees less flexed than the commercial motion capture system, on average. A bias towards less measured flexion on the part of the PB system was consistent across subjects, although the degree of this bias varied somewhat from subject to subject (see Figure 4). Standard deviation in the discrepancy between elbow flexion estimates was 5.70 degrees, making a 95% confidence interval span roughly 12 degrees. Hand velocity estimates similarly tended be under estimated by the parts based system, albeit relatively slightly (by 1.80 cm/s, on average); the standard deviation for these discrepancies was 2.57 cm/s. Less average discrepancy was exhibited between shoulder velocity measures. Mean discrepancies in shoulder velocities were under 0.3 cm/s, on average, with a standard deviation under 0.5 cm/s.

Deviations in absolute position estimates, after coordinate frame alignment, were fairly consistent across the wrist, elbow and shoulder positions. The mean absolute error in estimates of shoulder and wrist position was 23 mm and 25 mm respectively. Elbow positions were measured slightly more accurately than either the shoulder or wrist; mean error here was 19 mm. Standard deviations in errors, after alignment, were on the order of 3 cm across positions, leading to RMSE values of 33 mm, 25 mm and 34 mm for the shoulder, elbow and wrist respectively.

Metric 2: *validity*

Research indicates that, when reaching forward across the desktop, individuals with more extensive arm impairment may recruit more proximal parts than controls [2]. This means that impaired individuals may move their shoulders towards an object comparatively more than their hands. Our results are consistent with this prior research. In our cohort, velocity of the shoulder was negatively and significantly associated with FMA scores when measured with both the PB system as well as the VICON^{TM} device (correlation

with FMA score was -0.81 for the VICON^{TM} and -0.76 for the PB tracker; see Table III). In addition, a higher average movement speed at the hand was positively correlated with FMA score, as was a greater standard deviation in velocity. Significance here, however, was only achieved when relating FMA scores to standard deviations in velocity as measured by the VICON^{TM}. Counts of peaks in measured hand velocity, however, did correlate strongly with FMA scores across devices (correlation with FMA score was -0.76 for the VICON^{TM} and -0.73 for the PB tracker); the association was close to significant for the PB measurements (p-value = 0.06) and significant for those from the VICON^{TM} (p-value = 0.05). The more functionally disabled individuals, then, tended to move more slowly, and their velocity profiles were less smooth. Finally, the deviation in elbow flexion was positively and significantly correlated with FMA score at the 0.05 level for the VICON^{TM} measurements; the p-value relating the same statistic to the PB tracker measurement was 0.09.

V. Discussion

Discrepancies between individual points on the body as measured by a commercial system and our system can be compared to results previously reported in the literature [15], [22]. The authors of [15], for example, report average absolute discrepancies between 15 IR markers measured by a VICON^{TM} and 15 corresponding virtual markers measured with a parts-based tracker to be 80 mm, with a standard deviation of 5 mm. This average error is likely the result of consistent measurement bias of the video system, at least in part, given the relatively tight standard deviations. In addition, these same authors found more distal landmarks on the body to be measured less accurately relative to motion capture than proximal landmarks, especially during the performance of activities like running. The authors of [22] report much smaller average deviations between measurements of knee, hip and ankle joint centers made using a markerless tracking system and manually identified joint centers from laser scans. RMSEs here were less than 20 mm for all joints and errors in knee angle reconstructions during gait were bound by 2 degrees.

Both standard deviations in absolute errors and RMSE values are higher in our work than in [15] or [22]. Both [15] and [22], however, make use of substantially higher fidelity models trained using motion capture data or data from a laser scanner. More specifically, in [22], video derived visual hulls [26] of an individuals body are registered against scanner-derived models of the body to obtain join estimates. In [15], video-derived silhouettes and edges are compared to silhouettes and edges that are hypothesized using a motion capture trained model of motion. In addition, both [15] and [22] employ a layer of optimization to refine shape and pose estimates from completely markerless data at each frame. The authors of [22] makes use of an iterative optimization technique while [15] uses the condensation algorithm [27]. Our tools require comparatively limited training using video data alone, no dynamic priors, and employ no optimization to refine estimates; this reduces computational complexity at the cost of increased error. In terms of the speed of processing, our current Matlab implementation operates at roughly 2 seconds a frame. We expect an optimized C implementation should be able to operate in real time.

In our work, measurement error resulted, at least in part, from the fact that colors on a shirt were measured as opposed to anatomical landmarks. When individuals leaned over the table during experiments, different parts of the shirt were exposed to the cameras and, as a result, different arm segmentations and reconstructions resulted. Errors were relatively subject-specific, as can be seen in Figure 4. We hope to reduce such errors in the future by more rigorously standardizing the fit of clothes and by reorienting cameras to image not only the front of each subject but his or her back as well.

Even with such a relatively rudimentary kinematic tracker that makes no dynamic assumptions, however, errors in position estimates are still relatively small (under 3 cm, on average). More importantly, tracking results have proven to be accurate enough to carry information about subjects functional scores on the FMA. Standard deviation in shoulder velocity consistently correlates negatively (under -0.75) and significantly ($p \leq 0.05$) with FMA scores, while measures of elbow flexion consistently, strongly and positively correlate (over 0.65) with FMA scores. In addition, measurements of the peaks in hand velocity profiles consistently and negatively (under -0.70) correlate with functional scores. Such associations are consistent with prior clinical research, which has demonstrated increased smoothness of hand motion to reflect functional recovery after stroke [1] and increased range of elbow extension to result from physical rehabilitation [19]. Similarly, increased movement of the torso has been linked to increased functional disability after stroke [2]; measured shoulder velocity in this study can be considered a proxy for this measure. What is significant about results we present here, however, is that they demonstrate that features can be measured relatively cheaply, robustly and in a way that retains clinical salience with off the shelf video cameras and a parts-based (PB) tracking technique.

Although absolute errors between various position estimates were comparable in this research, errors that related to more distal parts of the body (i.e. the elbow and hand) more seriously impacted correlations between motor statistics and FMA scores. At the elbow, significance in associations was retained across devices only with a relaxed threshold on p-values of 0.1. At the hand, significance in velocity associations was lost entirely for the PB device. At the shoulder, by contrast, significance in associations between kinematics and FMA scores were retained across measurement devices, even in the presence of comparable measurement noise. The strength of association between proximal measures and functional performance after stroke is encouraging in that relatively proximal parts of the body tend to be more accurately measured by parts-based trackers [14], [15]. This is partly because the torso produces more image information than a more distal part, like the hand; the problem of associating image data with the torso is therefore simplified.

It is debatable as to how willing stroke survivors may be to wear a colorful garment, like the one wew use here, in

real functional situations. Our experience, however, indicates stroke survivors may be willing to wear such an outfit, assuming the functional tracking system yields real functional benefit. Our tracking methodology, moreover, is sufficiently general to be applied to situations where limb segments have unique, well defined appearance. A bright t-shirt, for example, creates a situation in which the upper and lower limbs are distinct in color. Alternate appearance models attempt to capture features like edges [15] or texture [13] instead of or in addition to color, and have produced good tracking results in diverse imaging conditions like the outdoors [13], [14]. Alternative models, then, may ultimately allow for more naturalistic tracking situations, where no markers of any kind are required. Research presented here can be seen as a baseline or best case scenario for such alternative techniques.

In summary, this research provides a foundation for the future. We have demonstrated that diagnostic kinematic statistics can be measured in stroke survivors using a kinematic parts based (PB) tracker that makes no dynamic assumptions and is appropriate for functional environments. The potential advantages of the approach are that it 1) does not depend on motion; 2) does not depend on a static background; 3) is robust to occlusions and 4) is capable of recovering from error. In future work, we expect to explore reconstruction errors as they relate to system specific features like image compression and frame rate, and to incorporate appearance and shape information about the torso directly, so as to avoid the use of the shoulder as a proxy measure for this part. Finally, we expect to use kinematics recovered from video to facilitate feedback about motor performance during rehabilitation.

REFERENCES

[1] B. Rohrer, S. Fasoli, H. Krebs, R. Hughes, B. Volpe, W. Frontera, J. Stein, N. Hogan N. Movement smoothness changes during stroke recovery. J Neurosci., vol. 22, no. 18, pp. 8297-304, 2002.

[2] S. Michaelsen, S. Jacobs, A. Roby-Brami, M. Levin. Compensation for distal impairments of grasping in adults with hemiparesis. Exp Brain Res., vol 157, no. 2, pp. 162-73, 2004.

[3] M. Cirstea, A. Mitnitski, A. Feldman, M. Levin. Interjoint coordination dynamics during reaching in stroke. Exp Brain Res., vol. 151, no. 3, pp. 289-300, 2003.

[4] A. Fugl-Meyer, L. Jaasko, I. Leyman, S. Olsson,. S. Steglind. The post-stroke hemiplegic patient. Scand J Rehabil Med., vol. 7, pp. 13-31, 1975.

[5] B. Volpe, M. Ferraro, H. Krebs, N. Hogan. Robotics in the rehabilitation treatment of patients with stroke. Curr Atheroscler Rep., vol. 4, no. 4, pp. 270-6, 2002.

[6] G. Uswatte, C. Giuliani, C. Winstein, A. Zeringue, L Hobbs, S Wolf. Validity of accelerometry for monitoring real-world arm activity in patients with subacute stroke. Arch Phys Med Rehabil. vol. 87, no. 10, pp. 1340-5, 2006.

[7] D. Pang, C. Nessler, J. Painter, C. Reinkensmeyer. Web-based telerehabilitation for the upper extremity after stroke. IEEE Trans on Neur. Sys and Rehab Eng., vol. 10, pp. 2102-108, 2002.

[8] D. Vlasic, R. Adelsberger, G. Vannucci, J. Barnwell, M. Gross, W. Matusik, and J. Popovic. Practical Motion Capture in Everyday Surroundings. Proceedings of ACM SIGGRAPH, vol. 26, no. 3, 2007.

[9] L. Sucar, R. Leder, D. Reinkensmeyer, J. Hernndez, G. Azcrate, N. Casteeda, P. Saucedo. Gesture Therapy - A Low-Cost Vision-Based System for Rehabilitation after Stroke. Biomedical Engineering Systems and Technologies, vol. 25, 2008.

[10] M. Goffredo, I Bernabucci, M Schmid, S Conforto. A neural tracking and motor control approach to improve rehabilitation of upper limb movement. J NeuroEng and Rehab, vol. 5, no. 5, 2008.

[11] A. Brubaker, D. Fleet: The Kneed Walker for human pose tracking. Computer Vision and Pattern Recognition (CVPR), 2008.

[12] G.Mori, J.Malik. Estimating human body congurations using shape context matching. European Conference on Computer Vision (ECCV), 2002.

[13] H. Sidenbladh, M Black. Learning the statistics of people in images and video. International Journal of Computer Vision, vol. 54, nos. 1-3, pp. 183-209, 2003.

[14] D. Ramanan, D. Forsyth., A. Zisserman. Tracking People by Learning their Appearance. IEEE Pattern Analysis and Machine Intelligence, vol. 29 , no. 1, 2007.

[15] L. Sigal, A. Balan, M. Black. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. International Journal of Computer Vision, vol. 87, no. 1, pp. 4-27, 2010.

[16] S. X. Ju, M. J. Black, and Y. Yacoob, Cardboard people: A parameterized model of articulated image motion, Proc. Int. Conference on Face and Gesture, pp. 561567, 1996.

[17] A. Senior, Tracking people with probabilistic appearance models, IEEE Workshop on Performance Evaluation Tracking Surveillance, pp. 4855, 2002.

[18] B Kopp, A Kunkel, H Flor, T Platz, U Rose, KH Mauritz, K Gresser, KL McCulloch, and E Taub. The Arm Motor Ability Test: reliability, validity, and sensitivity to change of an instrument for assessing disabilities in activities of daily living. Archives of physical medicine and rehabilitation, vol. 78, no. 6, pp. 615-620, 1997.

[19] M. Woodbury, D. Howland, T. McGuirk, S. Davis, C, Senesac, S. Kautz , L. Richards. Effects of trunk restraint combined with intensive task practice on post-stroke upper extremity reach and function. Neurorehabil Neural Repair; vol. 23, no.1, pp. 78-91, 2009.

[20] R. Hartley, A. Zisserman. Multiple View Geometry. Cambridge University Press, 2004.

[21] J.Y.Bouguet. Camera calibration toolbox for Matlab. http:www.vision.caltech.edu//bouguetj//calib_doc.

[22] S. Corazza, E. Gambaretto, L. Mundermann, T. Andriacchi. Automatic Generation of a Subject Specific Model for Accurate Markerless Motion Capture and Biomechanical Applications. IEEE Trans Biomed Eng, vol. 57, no. 4, pp. 806-812, 2008.

[23] D. Forsyth, O. Arikan, L. Ikemoto, J OBrien, D. Ramanan, Computational Studies of Human Motion: Part 1, Tracking and Motion Synthesis, Foundations and Trends, Computer Graphics and Vision, vol. 1, nos. 2/3, pp. 255, 2006.

[24] VB Brooks, JD Cooke, and JS Thomas. Control of posture and locomotion. 1973.

[25] LE Kahn, ML Zygman, WZ Rymer, and DJ Reinkensmeyer. Robot-assisted reaching exercise promotes arm movement recovery in chronic hemiparetic stroke: a randomized controlled pilot study. Journal of NeuroEngineering and Rehabilitation, vol. 3, no. 12, 2006.

[26] A. Laurentini, The Visual Hull Concept for Silhouette-Based Image Understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, no. 2, pp.150-162. 1994.

[27] M. Isard and A. Blake. Condensation - Conditional density propagation for visual tracking. International Journal of Computer Vision (IJCV), vol. 29, no. 1, pp. 5-28, 1998.