Trial Map : A Visualization Approach for Verification of Stroke Impairment Assessment Database

Jae-Yoon Jung, Janice I. Glasgow, and Stephen H. Scott

***Abstract*— Robotic/mechanic devices have become widely used for various medical assessments recently. While using these devices are beneficial in terms of accuracy and objectiveness, validation and consistency problem may occur when combining these data with traditional clinical information. Here we propose a visualization tool that can summarize the experimental data and compare them with the clinical data, in the stroke impairment assessment domain. This visual tool is based on a neural network ensemble that is trained to match the experimental data with Chedoke-McMaster scale, one of the major outcome measure for stroke impairment and recovery assessment. We compare our ensemble model with ten combinations of different classifiers and ensemble schemes, showing that it outperforms competitors. We also demonstrate that our visualization approach is consistent with clinical information, and reliable in a sense that output of our ensemble can be an estimator for the corresponding clinical data when Chedoke-McMaster scores are missing.**

I. Introduction

In recent years, robotic/mechanic devices have become utilized extensively as assessment tools in medical domains [1]. Compared to clinical examinations performed by trained physicians, this approach enables us to obtain quantitative, precise and objective (i.e., rater-independent) information [2], and to build a large-scale database in an efficient manner. However, it may bring some practical issues while integrating experimental data from robotic devices and clinical information from physicians, as follows.

First, a quick verification tool for the assessment data from the devices is required in order to ensure that the data are correct and no critical errors occurred while being transferred to the database, before they are used for further analysis. As experiments with mechanical devices typically have large numbers of repeated trials and tend to produce real-time observation data, validating these data can hardly be done manually because of their volume. Second, consistency between experimental data and clinical information should be examined. Even when the experimental data are checked to be correct in the first step, conflicts still may occur due to various reasons such as experimental configuration changes and errors in clinical assessments. Third, when these two conditions are met, missing features in the clinical information can be interpolated based on the experimental data. Typical omitted data in medical domains are missing not at random. So the only way to obtain an unbiased estimator for such data without additional information is to model missing data itself [3]. This condition can be mitigated with verified, experimental data if they are known to be related to the missing data.

We use the KINARM (Kinensiological Instrument for Normal and Altered Reaching Movements, BKIN Technologies Inc.) platform in order to measure sensorimotor performance of upper limb for stroke (cerebrovascular accident) and control subjects. Stroke is defined as the sudden loss of neurological function in the brain [4], but it can affect any part of body. We focus on upper limb functionality as paresis of upper limb is a common consequence of stroke that limits a survivor's quality of life [5].

Objective and quantitative assessment of stroke intervention is crucial in order to guide patient rehabilitation in an appropriate manner [6]. Several outcome measures have been proposed to assess various levels of stroke interventions including body function/structure (impairment), activities (disability), and participation (handicap). Most current assessment measures require trained physicians who have specialized knowledge on the field, but the assessment results may suffer from reliability problems such as inter-rater discrepancy and from poor responsiveness [7]. In a companion paper [8], we demonstrate that we can build a reliable and responsive outcome measure based on the experimental data, using neural network ensembles.

In this work, we present a visualization approach to verify the experimental data and to compare them with

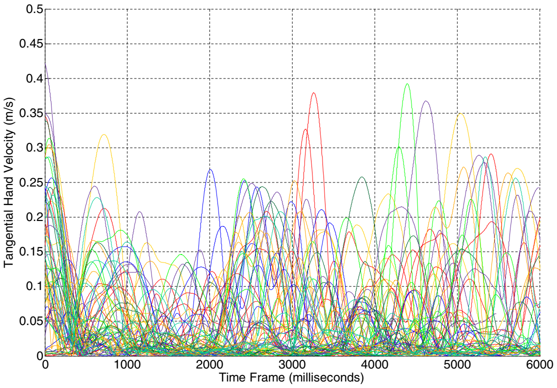


Fig. 1. An example of the raw tangential hand velocity, taken from a typical stroke subject's data. A color coding is used to group trials of the same direction together.

Jae-Yoon Jung and Janice I. Glasgow are with the School of Computing, Queen's University, Kingston, Ontario, Canada (email: {jung, janice}@cs.queensu.ca).

Stephen H. Scott is with the Department of Anatomy and Cell Biology, Canadian Institute of Health Research Group in Sensory-Motor Systems, Centre for Neuroscience Studies, Queen's University, Kingston, Ontario, Canada (email: steve@biomed.queensu.ca)

978-1-4244-1821-3/08/$25.00©2008 IEEE

clinical information, addressing the above mentioned issues in combining data from different sources. First, we establish a classification system to determine if the current subject has an upper limb impairment, based on the experimental data. A neural network ensemble is used for the classifier and the performance is compared with other classifiers and ensemble schemes. For the classification label, we adopt the Chedoke-McMaster score [6], which measures the physical impairment and disabilities associated with stroke. Next, we build a visual tool that compactly summarizes all experimental results as well as clinical information. Finally, missing Chedoke-McMaster values in the clinical information are calculated back based on the result of the trained ensemble networks.

II. Methods

A. Experimental Task

We use the unloaded reaching task [9], [10] to monitor upper limb functionality. This task is designed to measure generic motor performance of hand, arm, and shoulder when no mechanical perturbations are applied. Subjects are instructed to move their testing arm (left or right, one at a time) to a predefined center point and to wait for a target light illumination. Subjects cannot see their hand movement directly, and only the center point, targets, and the index finger position are illustrated in a screen. Targets are one of the eight fixed peripheral angles $(0°, 45°, ..., \ 315°)$ and are turned on in a pseudo-random order with a configuration that the total number of repeated trials per each direction or the total number of trials per task is the same across experiments. When a target is turned on, subjects are asked to move their hand to the target area as fast and as accurately as they can, and stay there until the current trial is ended, which is cued when the target light is turned off. An experimental set for a subject consists of these trials examined on one arm (i.e., two sets per each subject, after a successful test).

KINARM [11] is used as a robotic exoskeleton platform to enable a subject's flexion and extension movements of the shoulder and elbow with the arm projected on the horizontal plane. It also minimize effects of gravity during movements, by attaching fiberglass braces to the upper and lower segments of each arm. During each trial, various aspects of motor performance are recorded including hand position, tangential hand velocity, shoulder angles, and elbow positions. Figure 1 shows an example of the hand velocity selected from a typical stroke subject's data. A color coding scheme is used to group trials of the same direction together.

B. Feature Selection and Data Preprocessing

As KINARM processes the raw data signals at a rate of 1000 Hz (i.e., 1000 frames per second), it generates millions of attribute values for a single task. We extract four features that capture the main characteristics of the original data [12], as illustrated in Figure 2. Reaction time (R) is the time interval until the subject starts to move after the current target light is turned on. First peak velocity (F) is the first local maximum velocity, and the maximum velocity (M) is the global maximum velocity. For most of control cases, F is equal to M, which is not necessarily true for stroke subjects. Total movement time (T) is the time between the subject starts to move and stops.

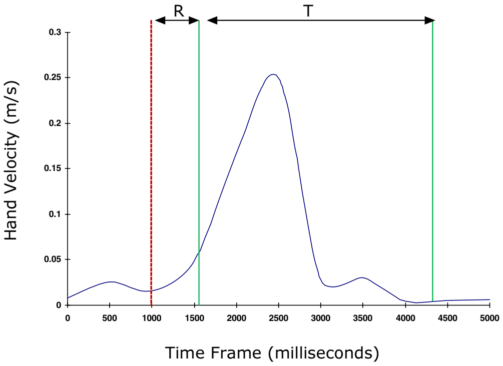


Fig. 2. An illustrative example of tangential hand velocity profile of a single trial. The leftmost (red) vertical line designates the time frame on which the target light is turned on. The following two vertical lines (green) show points when the subject starts moving (onset) and stops moving (offset), respectively. R (Reaction Time) is defined as the delay between the target on time and the movement start time. T (Total Movement Time) is the difference between movement onset and offset frames. F (First Peak Velocity) is the first local maximum velocity after the target is turned on. In this example, F is also the the maximum velocity over the entire trial period, M.

Experimental data was collected from 154 stroke sets and 108 control (reporting no previous neurological disorders) sets along with the corresponding clinical assessment information. Each set contains eight directions and eight trials are repeated per direction, so a normal set data contains 256 feature values. As we mentioned earlier, we adopt Chedoke-McMaster (we denote CM) Arm scale for the classification label, such that CM scale 7 (i.e., no apparent functional disabilities) is tagged as Y, otherwise as N. The problem of this original data is that only 36/262 sets are categorized as N class (i.e., functionally impaired), as many of stroke subjects data as well as all control subjects data belong to Y class. Therefore we bootstrap N class data with random perturbation up to $1.0 * \sigma$, where σ is the standard deviation of each feature over all trials in a set. So the preprocessed training data consist of 226 Y sets (not changed) and 252 N (216 added) sets, 478 sets in total.

C. Ensemble Networks

Figure 3 depicts the training steps of the ensemble classifier networks used in this work. First, the current training data $K = \{[R, F, M, T \mid C]^i, \ i = 1, ...|K|\}$ where the classification label $C \in \{Y, N\}$, is partitioned into 8 subgroups $[R_d, F_d, M_d, T_d \mid C]^i$ according to the target direction $(d = 1, ...8)$. Each subnetwork d gets this partial input patterns of length 32, and is trained to examine whether a pattern can be considered to represent the functionally intact subject or not. As the stroke subjects generally have deficits on only one arm, labeling result C may be different from the original

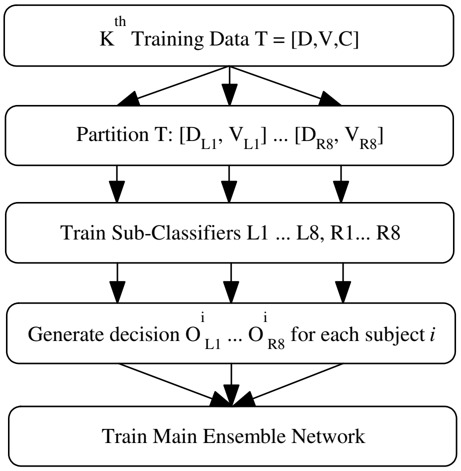

K th Training Data T = [D,V,C]
Partition T: [D L1, V L1] ... [D R8, V R8]
Train Sub-Classifiers L1 ... L8, R1... R8
Generate decision O i L1 ... O i R8 for each subject i
Train Main Ensemble Network

Fig. 3. The overall construction and learning procedure of the ensemble network. This figure shows one iteration case with K^{th} training set during cross validation, and the classification results specified in Table 1 are averaged estimations over ten times of 10-fold cross validation.

stroke/control label. Next, for each set i, the final output of the classifier is decided by the median output of O_d^i for all eight subnetworks. We choose median values rather than majority vote or weighted averages because there may be one or two outlier trials simply caused by neglect or fatigue. Finally, this procedure is repeated with ten times of 10-fold cross validation [13] to obtain a generalized performance expectation for the ensemble classifier.

III. Results

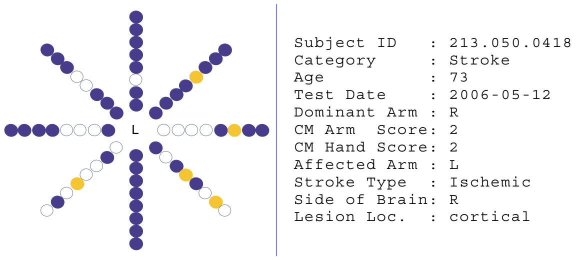
A. Classification Performance

Four different classification algorithms were considered for performance comparison: Decision Tree (C4.5) [14]; Naive-Bayes [15]; Support Vector Machines [16], [17]; and Logistic Regression Models [18]. Bagging [19] and boosting (AdaBoost.M1) [20] are combined with above classifiers in order to build an ensemble, and the option of ten iterations and a re-sampling ratio of 1.0 is applied to both schemes.

TABLE I

Classification Performance Comparison

Classifier Type	YY	NN	YN	NY	Error(%)
Decision Tree	183.8	216.9	42.2	35.1	16.2
Decision Tree + Boosting	189.0	229.0	37.0	23.0	12.6
Decision Tree + Bagging	196.9	222.6	29.1	29.4	12.2
NaiveBayes	177.4	190.7	48.6	61.3	23.0
NaiveBayes + Boosting	192.0	230.0	34.0	22.0	11.7
NaiveBayes + Bagging	174.2	191.1	48.6	61.3	23.6
SVM	210.2	241.7	15.8	10.3	5.5
SVM + Bagging	210.3	245.1	15.7	6.9	4.7
Log. Regression	210.4	241.4	15.6	10.6	5.5
Log. Regression + Bagging	211.0	241.0	15.0	11.0	5.4
Ensemble Networks	210.3	246.0	15.7	6.0	4.5

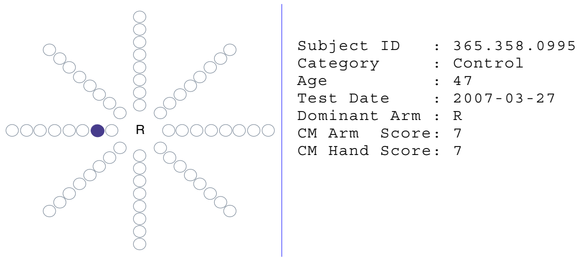


```
Subject ID    : 213.050.0418
Category      : Stroke
Age           : 73
Test Date     : 2006-05-12
Dominant Arm  : R
CM Arm  Score: 2
CM Hand Score: 2
Affected Arm  : L
Stroke Type   : Ischemic
Side of Brain: R
Lesion Loc.   : cortical
```

Fig. 4. An example of a trial map, taken form a typical stroke subject's data. The left figure illustrates the experimental results. Each circle represents a trial, and the direction and order of circles represent the target direction of the trial and the order of the trial tested in the same direction, respectively. An empty circle means that this trial looks like to be CM scale 7, while yellow or blue circle means that this trial seems to be in the N class. The right text box summarizes clinical assessment information corresponding to this experimental data.

Table I summarizes the results for the different classification algorithms. Column YY and NN correspond to the average number of task data pattern that were correctly classified as Y (CM score 7) and N (otherwise). Column YN and NY correspond to the average number of incorrect classifications as Y to N and vice versa. The error column shows the misclassification rate of each classifier in percentage, averaged over ten iterations of 10-fold cross validation procedures. The maximum possible number of YY and NN are 226 and 252 respectively, and the lowest error possible from blind estimation is 47.3 percent for this data. Decision tree algorithm gains about a 4 percent performance boost with both ensemble schemes, while NaiveBayes algorithm shows worse average performance with bagging. SVM and logistic regression combined with bagging turn out to be the best ensembles among compared classifiers. We do not report the testing result of boosting on both algorithms because zero training error occurred in some iterations. Our ensemble classifier outperforms all competitors, but the performance difference is not statistically significant. However, the error rate is significantly lower in general compared with our previous result [8], in which similar ensemble networks are used but the classification labels are either stroke or control.

Using the trained ensemble networks, we estimate the class label of each trial and visualize the all trial results into a map. Specifically, each t^{th} trial data of direction d, $pattern_{(d,t)} = [R_{(d,t)}, F_{(d,t)}, M_{(d,t)}, T_{(d,t)}|C]$ is fed into the corresponding sub-classifier d, as if all eight trial slots have the same $pattern_{(d,t)}$ data. The output of the sub-classifier are categorized as "CM scale 7" (output $>$ 0.6), "not sure" (0.3 $<$ output $<$ 0.6), or "not CM scale 7" (output $<=$ 0.3), and depicted as an empty, yellow, or blue circle, respectively. Two examples of this visualization tool are illustrated in Figures 4 and 5. The experimental results are shown on the left, and the corresponding clinical assessment information is shown on the right . Each circle represents a trial, and the position of a circle represents the direction and the order of these trials. For example, the map in Figure 5 tells that



```
Subject ID   : 365.358.0995
Category     : Control
Age          : 47
Test Date    : 2007-03-27
Dominant Arm : R
CM Arm  Score: 7
CM Hand Score: 7
```

Fig. 5. An example of a trial map, taken form a typical control subject's data. While Figure 4 clearly shows that the subject's left arm has functional problems to move towards almost all directions, this data illustrates that only one trial may have problem over 64 trials, possibly due to fatigue or neglect. This results are concurrent with the clinical assessment data (Chedoke-McMaster Arm and Hand score), shown in the right side.

the subject had some problems while doing the second trial while jointly moving elbow and shoulder towards the body center. In both figures, it can be easily identified that the experimental data match up with the clinical assessment data (CM hand and arm scores).

IV. Conclusion

In this paper, we introduced a visualization approach that can compactly represent a large-volume of experimental task data and facilitate consistency verification between machine-based examination data and manually collected clinical assessment information. In order to check the validity of experimental data, we adopt a neural network ensemble and train it to be matched up with the well-known Chedoke-McMaster scale. The classification performance was compared with other algorithms and ensemble schemes, including decision tree, naive-bayes with boosting, SVM, and logistic regression with bagging. In general, classifiers showed better performance than using stroke/control class labels, and our network outperformed the other ten classifier models. Trial maps summarizing experimental results were built upon this ensemble classifier, and we demonstrated that our maps are consistent with clinical data collected by physicians. The contribution of this work is to establish a visual tool useful useful for examining the experimental data at an abstract level, or to investigate discrepancy between data from two different sources. With extended use of robotic/mechanic devices for stroke rehabilitation field in recent years, this research area of automated assessment for stroke impairment and recovery is an emerging field. Our work contributes to a general framework for an automated clinical assessment system.

References

[1] S. H. Scott, "Apparatus for measuring and perturbing shoulder and elbow joint positions and torques during reaching", *Journal of Neuroscience Methods*, vol. 89, pp. 119-127, 1999.

[2] G. V. Dijck, J. V. Vaerenbergh, and M. M. V. Hulle, "Posterior probability profiles for the automated assessment of the recovery of stroke patients", In *Proceedings of the national conference on artificial intelligence (AAAI)*, vol. 22, issue 1, pp. 347-353, 2007.

[3] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley NY, 1987.

[4] S. B. O'Sullivan and T. J. Schmitz, *Physical Rehabilitation*, F.A. Davis Company, PA, USA, 2007.

[5] M. Vestling, B. Tufvesson, and S. Iwarsson, "Indicators for return to work after stroke and the importance of work for subjective well-being and life satisfaction", *Journal of Rehabilitation Medicine*, vol. 35, pp. 127-131, 2003.

[6] K. Salter, J. Jutai, L. Zettler, M. Moses, N. Foley, and R. Teasell, *Evidence-Based Review of Stroke Rehabilitation (EBRSR)*, Heart and Stroke Foundation of Ontario and Ministry of Health and Long-Term Care of Ontario, 2007.

[7] D. J. Gladstone, C. J. Danells and S. E. Black, "The fugl-meyer assessment of motor recovery after stroke: a critical review of its measurement properties", *Neurorehabilitation and Neural Repair* vol. 16, issue 3, pp. 232-240, 2002.

[8] J. Y. Jung, J. I. Glasgow, and S. H. Scott, " A Hierarchical Ensemble Model for Automated Assessment of Stroke Impairment", *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, accepted, 2008.

[9] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey, "On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex", *Journal of Neuroscience*, vol. 2, pp. 1527-1537, 1982.

[10] K. M. Graham, K. D. Moore, D. W. Cabel, P. L. Gribble, P. Cisek, and S. H. Scott, "Kinematics and kinetics of multi-joint reaching in non-human primates", *Journal of Neurophysiology* 89:2667-2777, 2003.

[11] K. Singh and S. H. Scott, "A motor learning strategy reflects neural circuitry for limb control", *Nature Neuroscience* vol. 6, issue 4, pp. 399-403, 2003.

[12] A. A. Zeid, *Parameter Analysis for Robotic Assessment of Impairments in Reaching Due to Stroke*, Master Thesis, Queen's University, Kingston, ON, Canada, 2007.

[13] R. Kohabi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 2, pp. 1137-1145, 1995.

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann San Mateo, CA, USA, 1992.

[15] P. Domingos and M. J. Pazzani, "Beyond independence: conditions for the optimality of the simple bayesian classifier", In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 105-112, 1996.

[16] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press Cambridge, MA, USA, 2001.

[17] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training support vector machines", *The Journal of Machine Learning Research*, vol. 6, pp. 1889-1918, 2005.

[18] A. Agresti, *Categorical Data Analysis*, Wiley-Interscience, 2002.

[19] L. Breiman, "Bagging predictors", *Machine Learning*, vol. 24, issue 2, pp. 123-140, 1996.

[20] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm", In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148-156, 1996.