A Hierarchical Ensemble Model for Automated Assessment of Stroke Impairment

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

***Abstract*— Assessment of sensory, motor and cognitive function of stroke subjects provide important information to guide patient rehabilitation. As many of the currently used measures are inherently subjective and use course rating scales, here we propose a hierarchical ensemble network that can automatically identify stroke patients and assess their upper limb functionality objectively, based on experimental task data. We compare our neural network ensemble model with ten combinations of different classifiers and ensemble schemes, showing that it significantly outperforms competitors. We also demonstrate that our measure scale is congruent with clinical information, responsive with changes of patients motor function, and reliable in terms of test-retest configuration.**

I. Introduction

Stroke (cerebrovascular accident) is defined as the sudden loss of neurological function caused by the interruption of blood flow to the brain that lasts more than 24 hours [1]. It is the most common cause of disability and affects approximately 700,000 people each year in the United States [2]. Only about ten percent of stroke survivors can fully recover, and impairment often includes upper-limb hemiparesis resulting in a substantive reduction in the quality of life post-stroke [3].

Objective and quantitative assessment of stroke interventions is crucial to prognosis for recovery and to selection of rehabilitation strategies [4]. Several outcome measures in stroke rehabilitation 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 how to perform the various assessment techniques. But the assessment results may suffer from reliability problems such as inter-rater discrepancy and from poor responsiveness [5]. As human-involved clinical assessments are inherently subjective, robotic / mechanic devices have been developed and adopted for assessment purposes recently [6], [7].

We collected experimental data of stroke and control subjects from a robotic device (KINARM) that can monitor various aspects of a subject's upper limb movement during a given task. Relatively little attention has been paid to the automated assessments in stroke recovery, partly due to lack of the appropriate devices and difficulty in building a large-scale patient database. Our general hypothesis is that robotic devices could be useful for the next generation clinical assessment.

In this context, our goals are 1) to build an effective classifier that can identify stroke subjects; and 2) to produce a responsive assessment score, only based on experimental task data. The hierarchical ensemble network used here consists of a set of artificial neural networks each of which deals with only a partition of input pattern, and the main network on top of these subnetworks. We compare the performance of this model with ten other ensemble schemes to examine if this model works better than other classifiers. Secondly, assessment scores are calculated from the collected output of sub-classifiers. These results are compared with Chedoke-McMaster scores [4] which are measured by clinicians.

II. Methods

A. Experimental Task

Fig. 1. Examples of the raw hand trajectory data. Left figure shows a typical stroke subject's hand movements during reaching tasks from the center point to eight different targets, and the right figure depicts a typical control subject's hand movements. Color coding is used in order to designate trials towards the same target direction.

KINARM (Kinensiological Instrument for Normal and Altered Reaching Movements, BKIN Technologies, Kingston, ON) 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, and to minimize effects of gravity during movements [8].

The experiment we describe here involves an unloaded reaching task [9], [10], designed to measure the generic motor performance of upper extremities with no mechanical perturbations applied. Stroke and control (reporting no previous neurological disorders) subjects are instructed to reach with their arm (left or right, one at a time) from a

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

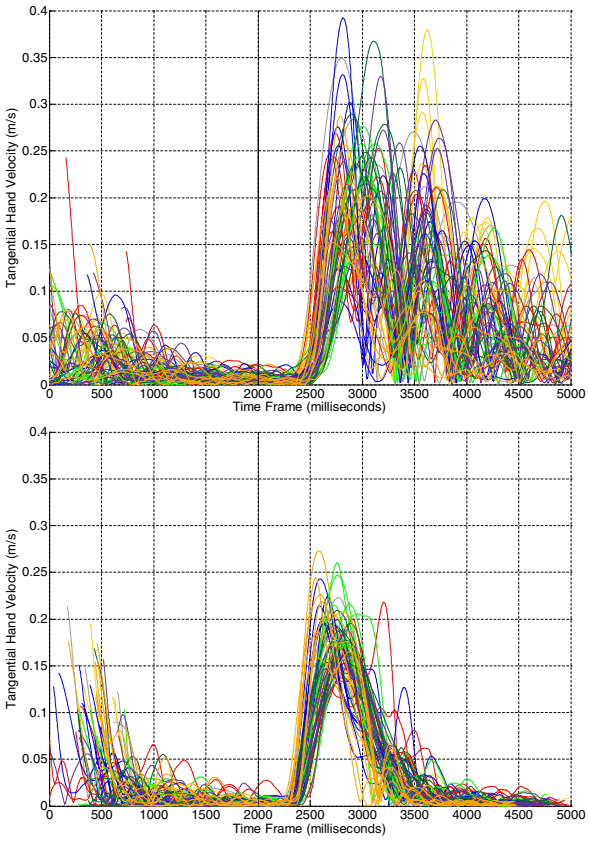

Fig. 2. Examples of the raw tangential hand velocity data. For each trial, velocity data are illustrated as if the target light is turned on at time frame 2000. Upper and lower graph show changes of hand speed during reaching task from the same stroke patient (upper) and control subject (lower) with Figure 1. Most of control subjects end trials within three to four second period, while many of stroke patients cannot finish a trial within this period, as shown in the upper graph.

given center point to one of the eight fixed peripheral targets $(0°, 45°, ..., 315°)$ when the target light is turned on, as fast and accurate as possible. There is no restriction on the minimum/maximum velocity and subjects are instructed to stay their testing arm in the target area and to keep their hand at the target until the target light is extinguished. The order of illuminated targets are selected in a random manner, but with a configuration that the total number of repeated trials per each direction would be the same. An experimental session of a subject consists of this set of trials examined on both arms.

During each trial, various aspects of motor performance are recorded including hand position (we denote as P), tangential hand velocity (V), shoulder angles, and elbow positions. In this work, we choose the first two features into consideration in order to reduce the input space. Figures 1 and 2 illustrate an example of these two features. Feature data are taken from a typical stroke session (left column) and a control session (right column), and a color coding scheme is used to group trial data of the same direction together.

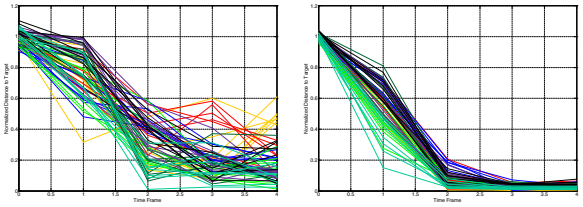

Fig. 3. Examples of the preprocessed data used to train our ensemble model. Left and right figures show the median distance to the target $\bar{D}$ of the corresponding stroke subject and control subject shown in Figure 1, before normalization.

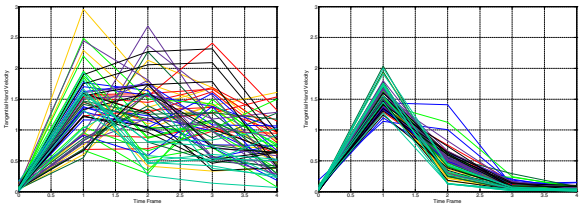

Fig. 4. Examples of the preprocessed hand velocity data used to train our ensemble model. The left and right figures depict median hand velocity over 500 millisecond intervals, taken from the same stroke (left) and control (right) subject's data in Figure 2, before normalization.

B. Feature Selection and Data Preprocessing

The task data were collected from 59 control sessions and 93 stroke sessions. Some of stroke and control subjects took multiple number of sessions over time and this information is used for reliability and responsiveness analysis in the following section. As KINARM captures the raw data signals at a rate of 1000 Hz (i.e., 1000 frames per second) and trial durations vary from three seconds to ten seconds at maximum, a single session of 128 trials (8 trials per direction x 8 directions x left/right arm) has 1.1 x 10^6 $\sim$ 3.8 x 10^6 feature values.

In order to cut this raw data into a feasible input space, the following preprocessing steps are applied to each session data. First, we replace the hand position P with the distance to the current target, D. Second, the median values $[\bar{D},\ \bar{V}]$ are calculated on 500 millisecond intervals after the target light is turned on. Third, for each trial, the first five $[\bar{D},\ \bar{V}]$ tuple are selected to represent the corresponding trial data. If a trial is finished earlier than this 2.5 second period, then $[0.001,\ 0.001]$ is plugged into the remaining slots. Finally, all feature values are normalized into the $[0.0,\ 1.0]$ range, and any missing trial is replaced with five $[0.999,\ 0.001]$ values denoting that the subject did not move at all throughout a trial. Figures 3 and 4 show the preprocessed data examples, taken from the same session data as the earlier Figures 1 and 2. After preprocessing, each input pattern (session) has a fixed set of 1280 feature values.

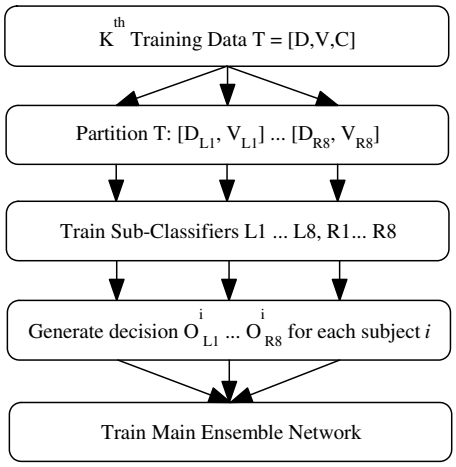

Fig. 5. The overall construction and learning procedure of our hierarchical ensemble network. This chart shows one iteration case with K^{th} training set, and the classification results specified in Table 1 are averaged estimations over ten times of 10-fold cross validation.

C. Hierarchical Ensemble Networks

Figure 5 depicts the training steps of the ensemble classifier networks used in this article. First, the current training data $T = \{[\bar{D}, \bar{V}, C]^i | i = 1, ...|T|\}$ where the classification label $C \in \{0, 1\}$, is partitioned into 16 subgroups $[\bar{D}_{s_d}, \bar{V}_{s_d}, C_{s_d}]$ according to the arm tested ($s = L \mid R$) and the target direction ($d = 1, ...8$). Each subnetwork s_d gets the partial input patterns $[\bar{D}_{s_d}, \bar{V}_{s_d}]$ and is trained to examine if each trial pattern belongs to a stroke subject or a control. A feedforward network with five hidden nodes is used for each subnetwork, trained by gradient-descent with a momentum backpropagation algorithm. As stroke predominantly affects on only one of arms, the stroke subjects' data are used in training subnetworks only if s is the affected arm Next, for each subject i, the final output of subnetwork classifier $O^i_{s_d}$ is decided by the median output of $s_d \in [0.0\ 1.0]$ for all eight trial patterns given to the subnetwork. We choose median values rather than majority vote or weighted average because there may be one or two outlier trials simply caused by neglect or fatigue but dominate the output values otherwise. Third, the main neural network classifier is trained based on these 16 $O^i_{s_d}$ input patterns in order to produce an estimation of this subject i being a control subject (= 1.0) or not (= 0.0). Finally, this whole procedure is repeated with ten times of 10-fold cross validation [11] to obtain a generalized performance expectation for this ensemble classifier.

III. Results

A. Classification Performance

Four different classification algorithms were considered for performance comparison: Naive-Bayes [12], Decision Tree (C4.5) [13], Support Vector Machines (RBF kernel) [14], [15], and Logistic Regression Models [16]. Bagging [17] and boosting (AdaBoost.M1) [18] are combined with above classifiers in order to build an ensemble, and the option of ten iterations and a resampling ratio of 1.0 was applied to both schemes.

TABLE I
Classification Performance Comparison*

Classifier Type	SS	CC	CS	SC	Error(%)
Decision Tree	74.3	38.4	20.6	18.7	25.9
Decision Tree + Boosting	79.0	42.0	17.0	14.0	20.4
Decision Tree + Bagging	76.1	47.8	11.2	16.9	18.5
NaiveBayes	72.3	46.1	12.9	20.7	22.1
NaiveBayes + Boosting	80.0	40.5	18.5	13.0	20.7
NaiveBayes + Bagging	76.0	45.0	14.0	17.0	20.4
SVM	77.5	45.5	13.5	15.5	19.1
SVM + Bagging	80.0	47.7	11.3	13.0	16.0
Logistic Regression	72.6	48.6	10.4	20.4	20.3
Logistic Regression + Bagging	76.8	48.9	10.1	16.2	16.0
Hierarchical Ensemble NNs	80.2	52.1	6.9	12.8	12.9

* $|S|$ = 93, $|C|$ = 59

Table 1 summarizes the results. Column SS and CC correspond to the average number of task data patterns that were correctly classified as stroke and control, respectively. Column CS and SC correspond to the average number of incorrect classifications as control to stroke, and vice versa, respectively. The right-most error column shows the misclassification rate for each classifier in percentage, averaged over ten iterations of 10-fold cross validation procedures. As the number of stroke ($|S|$ = 93) and control ($|C|$ = 59) data are not equal, the lowest error possible from blind estimation is 38.8 percent for this data set. For all classifier types, the ensemble versions show better performance than a single classifier. SVM and logistic regression combined with bagging turn out to be the best ensembles among compared classifiers, but the misclassification rate of these ensembles are significantly (t-test, $p < 0.01$) higher than that of our hierarchical ensemble model.

The error rate of about thirteen percent would possibly be considered high in other problem domains, but we believe that it is close to the optimum result we can get from real patient data in a medical domain. For example, an elderly control subject may perform more poorly than a younger, mild stroke patient. In contrast, some stroke patients recover completely over time without chronic upper limb deficits, thus eventually become indistinguishable from controls.

B. Reaching Assessment Score

Based on the final sub-classifier outputs $O^i_{s_d}$ for each subject i, we construct an outcome index that measures motor performance of each arm. More specifically, the reaching assessment score of subject i is determined as a sum of O^i values,

$$score^i = [\sum_{d=1}^{8} O^i_{Ld} \sum_{d=1}^{8} O^i_{Rd}]$$

The maximum score of $[8.0\ 8.0]$ means that the subject shows no upper limb deficits during the experimental task, whereas the minimum score of $[0.0\ 0.0]$ implies the opposite.

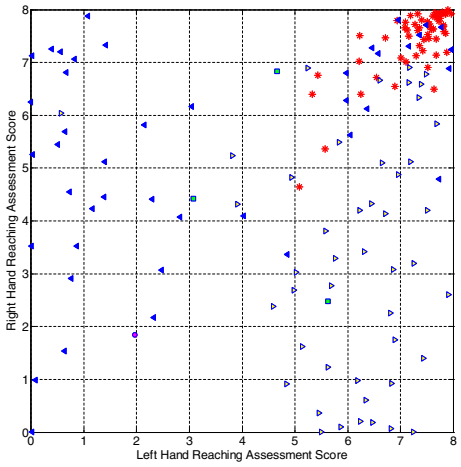

Fig. 6. Left and right reaching assessment scores are shown in two-dimensional plane. Each marker represents one session data. Stars, left triangles, right triangles, circles and rectangles specify control subjects, left arm affected stroke patients, right arm affected stroke patients, both arm affected by stroke, and stroke subjects of whom affected arm information was not available, respectively. The affected arm information is independently reported through clinical assessments, not estimated from task data or our ensemble network results.

Figure 6 illustrates the result combined with the affected arm information, independently collected through clinical examinations. This figure demonstrates that our measure matches up with clinical data, in a sense that most control subjects data are plotted in the upper right corner, and left (right) arm affected stroke subjects data are plotted in the left (right) side in general, respectively.

Next, we check the responsiveness (i.e., sensitivity to changes within patients over time [4]) of our measure by comparing the reaching scores from the same subject with Chedoke-McMaster scores, which are measured by physicians. Figure 7 plots the changes of the reaching score while the corresponding Chedoke-McMaster scores remain the same during various time intervals. Each line segment depicts the result of two sessions done by the same subject, and the x-axis represents the time interval between two sessions. Most lines move towards performing better than the previous assessment, showing that our measure are responsive in both short term (e.g., a few days) or relatively long term (6 months or more) periods in the rehabilitation process. Further, seven stroke patients and two control subjects have two sessions that are examined on the same day, in order to check if changes occur under repeated experiments. Pairwise t-test result in score pairs ($p^L = 0.5108,\ p^R = 0.1025$) which show that these two test results are not much different (i.e., cannot reject the null hypothesis that test data are drawn from the same population at the five percent significance level), implying that our score measures are reliable under test-retest configurations.

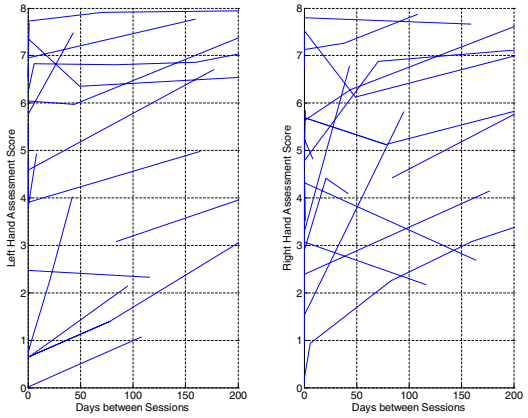

Fig. 7. Plots of reaching score changes over time. Each line segment represents scores of two sessions done by the same subject, over the given time intervals specified in x-axis. The first session date is set as day zero for all subjects. Note that only changes while the corresponding clinical-based Chedoke-McMaster scores remain fixed are shown in this figure.

IV. Conclusion

In this work, we introduced a hierarchical ensemble network model used in identifying stroke patients and assessing their upper limb functionality based on experimental task data. The classification performance was compared with other algorithms and ensemble schemes including SVM, logistic regression with bagging, and naive-bayes with boosting, and our network outperformed all ten compared classifier models. The reaching assessment score proposed here is calculated from the partial output of sub-network classifiers. We demonstrated that this outcome measure coincides with clinical assessment information, and can capture the changes of functionality over time, while the currently adopted clinical score did not. The contribution of this work is to establish an automated assessment tool with a reliable scoring measure. We consider this model as an initial step towards a general framework of a robust assessment system. Such a system would complement clinical-based examinations for stroke rehabilitation.

V. Acknowledgement

We thank K. Moore and H. Bretzke for technical help. This work was supported by NSERC and Canadian Institutes of Health Research (CHIR) MOP 81366.

References

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

[2] W. Rosamond *et al*, "Heart Disease and Stroke Statistics—2007 Update", *Journal of The American Heart Association*, Circulation vol. 115, e69-e171, 2007.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] D. Nozaki, I. Kurtzer, and S. H. Scott, "Limited transfer of learning between unimanual and bimanual skills within the same limb", *Nature Neuroscience*, 9(11):1364-1366, 2006.

[8] 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.

[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] 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.

[12] 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.

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

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

[15] C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

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

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

[18] 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.