The First Body Sensor Network Contest

Summary and Results

Roozbeh Jafari, University of Texas at Dallas
John Lach, University of Virginia

The First Body Sensor Network Contest was held in conjunction with the *International Conference on Body Sensor Networks* in May 2011. It was organized in response to the recognition that the practical value of body sensor networks (BSNs), which have potential to revolutionize a variety of fields, from medical patient monitoring to recreational sports training, is hindered by the fact that the signal processing for the collected sensor data is often designed only for specific datasets, BSN platforms, and applications. Due to inevitable variations in hardware, software, and experiment design, it is extremely challenging for BSN researchers and developers to reproduce and verify or demonstrate improvement over results reported by others. Additionally, due to variations in sensor placement and orientation, it might be difficult to later extend and reuse existing datasets. All this suggests an urgent need to develop signal processing techniques capable of handling multiple, possibly non-standard, datasets collected on various patient populations using disparate BSN platforms following a variety of experimental protocols. This goal can be best realized through sharing of existing data and tools in the research community. In addition, data sharing can significantly strengthen the value of statistical signal processing results, provide important insight into data collection and storage issues not raised using individual tools, and eventually help shape a data standard for the community.

The competition was therefore developed to:
- Promote data sharing and reuse in the BSN field.
- Explore the issues of signal processing applied to multiple non-standard datasets.
- Encourage development of light-weight and robust signal processing and decision making techniques that guide the user to gain better access to the relevant information.
- Help shape an eventual data standard.

This document provides a description of and results from the contest that was conducted in May 2011. Seven teams from around the world participated, and much can be gleaned from the experiences of the organizers and participants as the BSN field moves forward.

Contest Preparation

In order to provide the necessary dataset diversity, the competition combined data from three research groups from the University of Virginia, the University of Texas at Dallas and the University of California, Berkeley. During the competition, the organizers released two sets of data for training and testing the participant's software. This included the raw sensor data collected from the various sensor systems, as well as ground truth annotations to help guide the participants' signal processing development.

The first set of data included:

1. Multiple repetitions of the key movements of interest (See Table 1).
2. Ground truth labels for the movements of interest and unknown movements.
3. Description of the sensing systems, sensor positioning and orientation.

Table 1: Movements of Interest

№	Movement Name
1	Sit to Stand
2	Stand to Sit
3	Sit to Lie
4	Lie to Sit
5	Turn Counterclockwise 90 deg
6	Turn Clockwise 90 deg
7	Pick a Book up from the Floor
8	Place a Book on the Shelf
9	Step Forward (1 step)

The second set of data consisted of 30-45 second walking samples. Walking samples included two types of variations. First, we varied the experimental setup properties –some data were collected on a treadmill, while other data were collected during free walking. The second variation was the way the individual walked – speed, inclination, shoe type, surface type, and consistency of conditions for individual trials.

Participants were given a goal to prepare signal processing software to effectively handle three distinct tasks across the diverse datasets:

1. In a given signal, detect movements of interest and their beginning and end annotations (within +/-0.5 seconds of the ground truth). The contestants were given information about the dataset of origin for this task (i.e. data collection platform, configuration, placement, etc.).
2. Given a sample of walking data, detect the average stride time of the data sample, again with dataset origin information.
3. Given an annotated sample, identify whether it represents the 'sit-to-stand' action, but *without* the dataset origin information.

Expected Early Challenges

- The first challenge was selecting a set of tasks for the competition. Throughout the process, we had multiple goals in mind. First, the selected tasks needed to be representative of challenges encountered in the process of data sharing. The aim was set on the issues of sensor node type, sensor node placement, and variability in movement performance. Second, we wanted to be very careful not to introduce any artificial problems that would complicate the technical implementation of the solution without any impact to the data sharing goal. Finally, we aimed to make the tasks practical and realistic. We did not want to create a set of tasks that may be interesting from the mathematical perspective, but irrelevant in the practical setting.
- The second challenge was movement of interest selection. The contest goals included both detection and segmentation of movement of interest variations in a continuous data stream. As a result, it was essential to pick movements that have little resemblance so the contestants can focus on the variations of the data. It was important to collect a movement set that can be performed in a sequence without introducing any additional *unknown* movements. Finally, since one of the tasks we outlined was movement detection, it was important to make sure that there is no obvious implied ordering to the selected dataset.

- The third challenge was data collection coordination. While the data for some of the tasks had the associated dataset included, we wanted to make sure that at least on the semantic/visual level different datasets are consistent with each other. We wanted to stress the similarity in the data, even in the presence of significant variations.
- The final challenge was data annotation. It was important for us to provide the best ground truth we could; as a result much of the annotation work has been done manually.

Unexpected Early Challenges

While we put a lot of thought into the contest preparation, in the process of the contest we discovered some challenges that we did not expect:

- A major setback, which we discovered somewhat late into the competition, was lack of sample testing sets that teams could use to evaluate the performance of their approaches. We originally anticipated the teams would use a portion of the training data for testing. However, this did not include the format differences between the training and testing datasets. We did not expect that due to the nature of the test tasks, the format of the training data would have to be altered. While we tried to keep those changes to a minimum, and eventually posted a trial test set, that cause some trouble for some of the teams.
- Originally, we expected that the majority of the teams would be present at the BSN Conference in Dallas and planned for the test phase to be conducted in person. We later realized that only a few teams would be present and had to adjust the rules during late stages of the competition to account for that.
- The training data and documentation had some issues. In order to be able to track changes in the posted files, we kept a strict log of changes to the uploaded data along with the time associated with the changes made.
- There were many minor issues that teams brought to our attention. To address those issues, we setup a mailing with the entire contest organizer team as recipients. We tried to keep response time to a minimum allowing quick clarifications about the posted data and contest rules. We also setup an FAQ page on the contest website to keep track of the major questions that may be of interest to all of the teams.

Contest Proceeding

Once we realized that the majority of the teams would not be able to make it to the BSN Conference in Dallas, we adjusted the rules to allow for remote participation. As a result, we asked the teams to provide us a 2-hour time frame during one of three days Friday – Sunday. Each team was scheduled to receive the test data at the beginning of the 2 hour period, and needed to forward us the results at the end of it.

The data for each of the tasks was organized as follows:

1. We collected continuous data of subjects' movements from the list in Table 1. To simplify the detection of the movement beginning and end annotation, we introduced a bit of a pause between movement performances (e.g. stand-to-sit -> brief pause -> sit-to-lie). That was done to accommodate for the training on individual samples, and make the task of movement annotation more interesting. The number and type of movement varied between the test trials.
2. We collected observations of multiple subjects walking (both on the treadmill and free walk). The test trials included some variations from the original training data: type of shoes, type of walking surface, walking speed, walking inclination.

3. We extracted individual trials from the data collected in the first two tasks with a mix of "sit-to-stand" and other movements from all three data sources.

Expected Challenges

- The timeframe of the proceeding was driven by the schedule of the BSN Conference. Some of the teams were not ready for this solution to our scheduling problem. The offered days were selected mainly during the weekend, which limited teams' access to the office/computer equipment during those days.
- We settled on the format of the test data only a few days before the proceeding of the conference, which left little time for the teams to prepare/modify their software accordingly. That caused for some modification during the time of the contest, which was not our intent.

Contest Evaluation

For each of the tasks, we ranked teams based on performance. The top performing team in each task received 5 points toward the final score. The second best team for each task received 3 points toward the final score. The third performing team received 1 point toward the final score. At the end, the teams have been ordered in descending order of the final score to find the winners.

Task 1 Evaluation

For the first task, the test dataset ground truth included:

1. Annotation of the start for each movement
2. Annotation of the end for each movement
3. Movement type (see possible movements in Table 1)

Contestants had to provide a list of detected movements that, just like ground truth, included beginning and end annotations and the movement type.

First, we verified whether the teams properly produced their outputs, such as verifying that teams did not report two *start* or two *end* annotations in a row (which was specified as being against the rules).

For the evaluation of the first task, we traversed the list of *start* ground truth annotations, and checked if the teams detected the beginning of any movement within 0.5 seconds of the ground truth entries. If they did, we verified whether the *end* annotation matched the ground truth as well. If that was the case, we verified if the movement was properly labeled. If all the conditions held, a team gained a point toward their final score on this task. Once we evaluated all individual movement samples, we tallied the scores and ranked teams based on their first task performance.

Task 1 Evaluation Challenges

- We originally considered counting both False Positives and False Negative errors (in many practical applications one may have more weight than the other). However, we decided against it in favor of simplicity. In the end, we only considered the False Negative errors, meaning that if teams detected phantom movements between actual movements, we did not penalize them for it. This worked out pretty well, because in reality, if teams detected an additional movement between the real movements, they

were not able to properly address the annotations points (and lost points even without an explicit penalty).
- In order to guarantee the quality of evaluation, we went over the evaluation results manually, which took a considerable amount of time

Task 1 Evaluation Results

- This is the task teams had the most trouble with. Only a few teams did exceptionally well, many teams were able to detect a few movements correctly, and a few teams were not able to detect any movements correctly.
- Some teams made a simplifying assumption that the end of one movement was also the beginning of the next one. That was not the case in our data, which resulted in some of the teams scoring no points in the first task.
- Some teams made a simplifying assumption and trained individual binary classifiers, instead of classifiers that aim to differentiate a larger movement set. It caused trouble, since "not-movement" often coincided with the way signals of other movements looked, causing confusion in the classifier performance.

Task 2 Evaluation

For the second task, the test dataset ground truth included:

1. Average stride time for each walking sample.

For each of the movements, we calculated the error in the average stride time evaluation. We added the error of all trials together and ordered teams in increasing order of the error sum.

Task 2 Evaluation Challenges

- For the evaluation of the second task we considered two types of annotations. First, we considered annotation of the steps of one foot manually. The other option was to consider a full gait cycle, assuming that steps of both feet take the same amount of time. We settled on the second annotation type because we provided the sensor readings of only one of the legs, and no specific step annotations were given during training.
- Some teams did not realize that for the first two tasks they could make individual classifiers for each of the datasets, and designed a unified classifier. That strongly reduced the quality of their classification.
- For the second task, some teams were not ready for the types of variations introduced in the test data and were not able to reliably find the average stride time. In fact, all teams did best on the third dataset that introduced the least amount of variation to walking.

Task 2 Evaluation Results

- All teams performed best in the second task. The error was fairly small for all teams, with best teams having almost no error at all.

Task 3 Evaluation

For the third task, the test dataset ground truth included:

1. Movement type of each sample

2. Movement dataset for each sample

For each of the movements, teams were awarded 1 point if they correctly identified a specific movement. Teams were then ranked in descending order.

Task 3 Evaluation Challenges

- For this task, we wanted to make sure that the movement type would not be obvious based on the movement length. We looked through the movement set (used in the previous tasks) and selected a set of 'sit-to-stand' and 'not-sit-to-stand' movements that had somewhat consistent lengths.
- We did not remove obvious differences in data format (number of sensing axes for example) from this data. We kept those, so that teams would be able to identify properties of the different datasets and leverage these properties in their algorithms. This decision is also in line with the reality of data sharing challenges.

Task 3 Evaluation Results

- The third task was mostly hit or miss, with teams either performing exceptionally well or very poorly. We expected the results to be more evenly distributed.

Overall Evaluation

Eight teams originally registered for the contest. After the team registration and before the actual contest, three teams let us know that they will not be able to participate in the contest because they ran out of time, and were not able to prepare a solution. As a result, only five teams were considered in the final evaluation of the results.

Table 2 contains the normalized results, with each column corresponding to an individual task. We normalized each column with respect to the best team performance. During the first task evaluation, Team A performed significantly better than other teams. The other teams had trouble with proper annotations, causing a low detection ratio. For the second task, while the numbers look well distributed, all of the team performed well. The error did not exceed a second on average for all teams. Finally, for the third task Team A and Team B recognized the majority of unknown trials, while other teams were not able to recognize them.

Table 2 Contest Evaluation

	Movements Detected (Normalized)	Error (Normalized)	Trials Detected (Normalized)
Team A	1	0.63	0.94
Team B	0	0.45	1
Team C	0.02	1	0
Team D	0.01	0.3	0
Team E	0.22	0.31	0.63

Acknowledgments

We would like to thank the following groups for their contributions to the organization and administration of this contest.

Participating Teams

Team Name	Organization
Warwick Neural Engineering	University of Warwick
	University of Southampton
National Taiwan University	National Taiwan University
William Mary	College of William Mary
SIIT	Sirindhorn International Inst. of Tech.
	Thammasat University
UCI	University of California, Irvine
Sensor Networks and Applications Research	Graduate School of the Chinese Academy of Sciences
WASNLab	University of Parma
The Pantographs	CREATE-NET

Data Set Contributors

- Vitali Loseu, University of Texas at Dallas
- Jeff Brantley, University of Virginia
- Ruzena Bajcsy, Victor Shia, Posu Yan, and Allen Y. Yang at University of California, Berkeley for their continuous support and providing their data and expertise
- Dr. Jay Han and Dr. Ted Abresch at University of California, Davis for agreeing to support the contest and sharing their data

Advisory Committee

- Insup Lee, University of Pennsylvania
- Majid Sarrafzadeh, University of California, Los Angeles
- Jack Stankovic, University of Virginia
- Guang-Zhong Yang, Imperial College London