Augmenting Gesture Recognition with Erlang-Cox Models To Identify Neurological Disorders in Premature Babies

Mingming Fan[1], **Dana Gravem MD**[2], **Dan M. Cooper MD**[2], **Donald J. Patterson**[1]
[1]Department of Informatics
[2]Institute for Clinical and Translational Science
University of California, Irvine
mingminf@ics.uci.edu, dana.gravem@uchospitals.edu, dcooper@uci.edu, djp3@ics.uci.edu

ABSTRACT
In this paper we demonstrate a Markov model based technique for recognizing gestures from accelerometers that explicitly represents duration. We do this by embedding an Erlang-Cox state transition model, which has been shown to accurately represent the first three moments of a general distribution, within a Dynamic Bayesian Network (DBN). The transition probabilities in the DBN can be learned via Expectation-Maximization or by using closed-form solutions. We test this modeling technique on 10 hours of data collected from accelerometers worn by babies pre-categorized as high-risk in the Newborn Intensive Care Unit (NICU) at UCI. We show that by treating instantaneous machine learning classification values as observations and explicitly modeling duration, we improve the recognition of Cramped Synchronized General Movements, a motion highly correlated with an eventual diagnosis of Cerebral Palsy.

Author Keywords
Gesture Recognition, Health, Sensors, User Modeling

ACM Classification Keywords
H.5.2 Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Algorithms, Human Factors, Measurement

INTRODUCTION AND RELATED WORK
There is emerging data that patterns of motion early in life can predict impairments in neuro-motor development. However, current techniques to monitor infant movement mainly rely on expert observer scoring, a technique limited by skill, fatigue, and inter-rater reliability. Consequently, we analyzed data collected by a lightweight, wireless, accelerometer system that measures movement and can be worn by premature babies without interfering with routine care. We hypothesized that we could improve the detection of CSGMs to a sufficient fidelity that it would be feasible to utilize the system in clinical care by reducing the amount of video that a clinician would need to review for positive diagnosis.

Figure 1. Baby being monitored in the NICU with accelerometers on each limb. The equipment in our study is wireless.

Medical Background
Over the past two decades, the incidence and survival of preterm births (infants born at less than 37 weeks of gestation) have increased dramatically [3]. Not surprisingly, long-term neurological complications are often associated with prematurity, such as cerebral palsy (CP) [4, 25] or the less severe category of minor neurological dysfunction [20] which has increased as well. The early assessment of physical activity patterns in premature babies is increasingly recognized as an essential step in identifying metabolic and/or neuromotor impairments and optimizing therapeutic approaches [9, 8, 42].

Given the fragility of this population and the constraints imposed on diagnostic procedures in premature babies, it is not surprising that early assessment of physical activity has proven to be quite challenging. Such measures depend almost exclusively on direct observation of infants, or on viewing post hoc, real-time videotape of infant activity [19]. Very little is known about the normal developmental patterns of physical activity in premature babies, and, as a consequence, identification of abnormalities can occur quite late. For example, it is generally agreed upon that the diagnosis of a condition like CP cannot be made definitively until a child

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.

is at least 4 years-old [25]. While sophisticated brain imaging such as MRI can provide additional anatomic mechanisms for conditions like CP [24], these approaches are not yet suited for screening and early diagnosis.

Gesture Recognition

While there has been a wide variety of work in gesture recognition in the computing field in recent years, there has been limited research using accelerometers to monitor infant movement. Factors such as accelerometer weight and size, which are not problematic for physical activity measurement in older children [14, 38], are major obstacles in premature babies. In our work we chose to use custom wireless accelerometers as the medium for detecting gestures due to their small size and high data density.

Broadly, gesture recognition can be done through a variety of media including video cameras [32, 29], touch screens, pointing devices [44], accelerometers [5, 23], forearm electromyography [39], fabric-embedded sensors [15, 11] and range-sensors [26]. Applications of gesture recognition include recognizing Activities of Daily Living (ADLs) like "setting the table" [35], recognizing the components of sign language (fenemes) [6], assisting in completing questionnaires [1] and for end-user gesture programming [2].

Machine Learning

Recognizing Gestures

There is also much work applying machine learning techniques to accelerometer data for gesture recognition. Hidden Markov Models are used for predefined gesture recognition [40]. An improved version of support vector machines (frame based SVM) was designed to improve recognition accuracy [45]. These gesture recognition techniques lack extensibility since they focus on a pre-defined gesture set. uWave extends it to more general user-defined personal gestures recognition [27].

Of particular interest to this work are methods for using machine learning to model explicit durations.

Modeling Duration in Markov Models

Frequently in activity and gesture recognition Hidden Markov Models are used to model the hidden state representing the occurrence of an event. This approach has been very successful in a number of studies [12, 35, 28] A single state discrete markov model assumes that the transition through the model will be distributed according to a geometric distribution. The probability of leaving the state after exactly t steps given the self-transition probability, a_{ii}, is

$$p(t) = {a_{ii}}^{(t-1)}(1 - a_{ii})$$

While this has worked in many cases, it falls short when the actual duration of activities is not distributed accordingly. In fact, intuition suggests that most activities are far more likely to be distributed very differently, perhaps uniformly or normally. When the time duration becomes an important component to the recognition task, a single markov model or a chain of markov models is inadequate.

Modeling Duration in Hidden Semi-Markov Models

If the underlying distribution is not geometric, a Hidden Semi-Markov Model may be a more appropriate choice for modeling the duration of an event. An HSMM has an exit distribution which depends on the amount of time that has been spent in the state so far and multiple observations can be made of the system while the model remains in the same state. Unfortunately a straight-forward application of the Baum-Welch algorithm to an HSMM is not possible and more computationally complex variants must be introduced. Nonetheless HSMMs have been broadly and successfully applied to areas including the recognition of Activities of Daily Living [16]. See [46] for a thorough survey.

Modeling Duration with Continuous Time Bayesian Networks

A different approach to modeling durations is represented by continuous time markov processes which represent transitions as rates rather than probabilities. Researchers have extended these ideas to collections of dependent processes which form Continuous Time Bayesian Networks. The advantage of such an approach is that the state of the system can be queried at any time period rather than just at discrete time-steps. Additionally they do not suffer from numerical representational problems when the duration of an activity is very long, but the discrete time sampling is very frequent. Exact inference in a CTBN is generally intractable, but approximate inference techniques have been developed [30]

Cramped Synchronized General Movements

Very few studies, have quantifiably measured infant limb movement and tied these movements with neurodevelopmental outcomes. Intriguingly, even using a large (4 g, 20 x 12.5 x 7.5 mm) commercially available accelerometer placed on a single upper extremity, Ohgi et al. [31], in pioneering work, found different patterns of spontaneous movements of premature infants with known brain injuries compared with controls. In addition, promising research has recently been conducted by Heinze et al. showing that a wired accelerometer could be used to differentiate between healthy babies and those at risk for CP [22].

A barrier in evaluating any new approach toward measuring infant movement is the dearth of metrics for comparison; however, in the case of the early diagnosis of CP, there exists a standardized, direct observation tool developed by Prechtl [36]. We designed our present experiments to compare our accelerometers with the Prechtl direct observational approach. In normal infants, Prechtl defined "general movements" (GMs) as elegant, smooth, variable in sequence, intensity and speed with a clear beginning and end. Prechtl also observed a unique abnormality of GMs that he named "cramped-synchronized" (CSGMs) in which the infant's limbs were rigid and moved nearly in synchrony. CSGMs have high predictive value for the development of CP. [37, 13, 17].[1] The literature shows that other motions are also correlated with CP. We did not investigate these alternatives, nor develop hypotheses for new signals such as audio cues (e.g., crying).

[1] An example video demonstrating a CSGM be seen here: http://archpedi.ama-assn.org/cgi/content/full/156/5/460/DC1

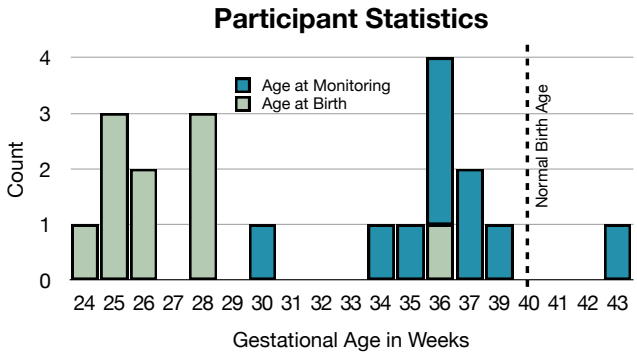


Figure 2. Histogram of age of participants at birth and at monitoring.

Abnormal:	0	7040	0	5838	3853	5444	2737	427

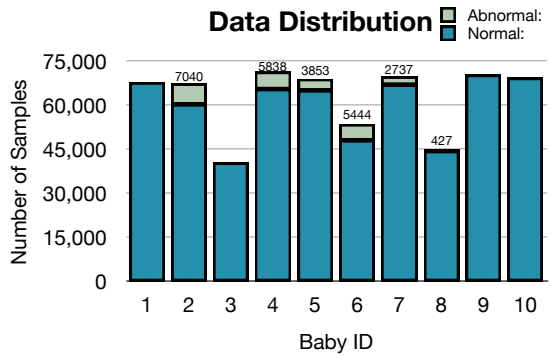


Figure 3. Distribution of samples after removing interventions. (The number of abnormal samples are shown above column)

Contributions of this paper

This paper adds to the literature on how to use accelerometer measurements to identify motion disorders in babies that are already classified as high-risk. The specific contributions of this paper are as follows:

- We build on previous work [41] by conducting a rigorous comparative ROC evaluation of the classes of machine learning methods previously published for detecting CSGMS.
- We demonstrate a lack of additional classification power in adding features which were developed to specifically represent motion symmetry observed in CSGMs.
- We demonstrate a technique for combining Erlang-Coxian (EC) distribution modeling with Dynamic Bayesian Networks.
- We evaluate EC methods using closed-form solutions, and compare it to other duration modeling methods.
- We conduct a cost-benefit analysis for an expert video scorer using our system in a clinical setting.

METHODOLOGY

Data Collection

Our experimental protocol was reviewed and approved by the Human Subjects Institutional Review Board at UCI. We identified potential participants by screening the medical records of infants in the NICU at the UCI Medical Center and recruited preterm infants with a gestational age at birth of between 23 and 36 weeks. Infants were excluded if they had mothers less than age 18 or if they had skin disorders which could preclude the attachment of the accelerometers to the skin. We recruited high-risk babies who had cerebral ultrasound abnormalities and low birth weight, both of which increase risk for CP. For this study the parents of 10 premature infants provided written informed consent and enrolled in the study. (Figure 1 shows an example of a baby being monitored by our equipment in the NICU, although not a participant in this study).

All infants were monitored and videotaped for 1 hour at 30-43 weeks corrected gestational age (see Figure 2) in their isolette wearing only a diaper and with all swaddling removed to allow for free limb movement. The ambient temperature of the isolette was adjusted and maintained according to the judgment of the NICU nurse.

A video camera was positioned with a mid-sagittal view of the infant above the isolette at a downward angle of 45 degrees to record motion for post hoc video scoring.

Four custom wireless accelerometers were used for data collection [34]. Each one measured 3 orthogonal axes of acceleration each of the 4 limbs. Devices were embedded in cloth bands that were placed around the wrists and ankles of the infants with a canonical anatomical orientation.

The accelerometers transmitted data that was sampled non-uniformly at approximately 19Hz in real-time to a computer located near by. The raw accelerometer data consisted of real valued samples of the 3 axes measuring the degree of acceleration due to gravity and changes in limb motion. The choice of a non-uniform sampling rate was a technical limitation of the devices, not a methodological choice. [2]

Expert Review

After each data collection session, the video data was transferred to a nurse trained in identifying CSGMs. The nurse was blinded to the patient information and accelerometer data and was asked to record the start and stop time for each observed CSGM. The annotations were then associated with the timestamp of the accelerometer readings. Although these readings became our ground truth and classification target, clearly the nurse was not able to accurately identify the movements to the accuracy of the data rate of the accelerometers. Start and stop boundaries were likely not exact as a result. Notably this process was similar to how a clinical evaluation would be conducted without any assistance from our system.

Cleaning Data

We followed several steps in order to clean the data and create features. First the video data and accelerometers streams

[2] The data set used in this paper is the same as reported in [41], however our data processing was done independently with corresponding differences seen in Figure 3.

had to be temporally aligned. This was done manually by comparing the motion in the video to the motion in the accelerometer stream. Second, during data recording, the nurses might move the babies a little if they needed care (For instance to change a diaper or attend to a medical concern) or adjust the sensors (e.g. sensors might slip off their limbs). All these activities were defined as "interventions", since it induces artificial errors in sensor data. So we manually reviewed all video to identify all possible interventions (e.g. starting and ending time) and removed the corresponding corrupted accelerometer data.

Feature Extraction

After cleaning, our data set was structured as shown in Figure 3. We represent the resulting raw data sample as a 14-tuple:

$$S_1 = (T, x_1, y_1, z_1, x_2, y_2, z_2, x_3, y_3, z_3, x_4, y_4, z_4, c)$$

where T is the timestamp of the sample, $x_1, .., z_4$ are 12 real numbers corresponding to 3 axes of 4 accelerometers. Each accelerometer corresponds to left arm, right arm, left leg, and right leg, respectively. The accelerometer readings vary between $-3g <= x_1, .., z_4 <= 3g$ and c is the ground truth indicating the presence of a CSGM.

Accelerometers measure a combination of gravity and motion. It was beyond the scope of this work to separate the acceleration caused by the baby's motion from gravity. This is a difficult task due to the changing pose of the baby. First we calculated the magnitude of the acceleration of each limb.

$$m_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$$

In order to get rid of low frequency noise, including gravity and sensor calibration drift, we subtracted the mean of a 10 second window centered at each sample, m_i.

From this data we calculated several basic motion features:

$$m_1, m_2, m_3, m_4, max(m_1, m_2), max(m_3, m_4)$$

$$max(m_1, m_3), max(m_2, m_4), max(m_1, m_2, m_3, m_4),$$

$$m_1 * m_2, m_3 * m_4, m_1 * m_3,$$

$$m_2 * m_4, m_1 * m_2 * m_3 * m_4$$

It is worth pointing out that the choices above support aggregations both across arms and legs, but additionally a novel feature for representing aggregation down the left and right side. This was motivated by a desire to capture the observed symmetric property of motion in CSGMs (left side body vs. right side body).

In addition, we also calculated temporal features for each one of the above 14 basic features, which were mean, max, min, standard deviation and z-score of a 2 second window centered on the current sample. Therefore, we end up with 84 features and one class label.

NON-TEMPORAL MODELING

We modeled our basic data using machine learning techniques from three different algorithmic classes, tree ensembles, boosted Naive Bayes and Support Vector Machines. In all the experiments that we report in this paper we conducted a full 10-fold cross validation on a baby-by-baby basis such that each experiment generated 10 models. Descriptions of our techniques follow.

Random Forest (RF) is an ensemble method which combines tree predictors of randomly selected features [7]. In our experiment, we used a random forest of 100 trees, each constructed while considering 7 random features (Out of bag error: 0.0004) without limiting the maximum depth.

AdaBoost is a powerful meta-algorithm, which can be used with other learning approaches to improve their performance [18]. We iterate AdaBoost with Naive Bayes (AdaBoost(NB)) for 10 rounds of iterations. During each iteration we reweight our learner with a pruning threshold of 100.

Support vector machines (SVM) are another state-of-art supervised learning technique which try to maximize the margin between classes. In our setting, we use the "radial basis function" (RBF) as our kernel because RBF can handle nonlinear attributes by nonlinearly mapping samples into a higher dimensional space, unlike a linear kernel [10].

In all non-temporal classification we ran our test cases using the "WEKA" toolkit [21]. In order to support comparison, we set the algorithms to generate probability estimations instead of binary labels.

To do performance analysis, we utilized "Receiver Operating Characteristic" (ROC) curves. ROC curves provide information about dynamics of true positive rate (TPR) and false positive rate (FPR), which can be used to evaluate the trade-offs between the relative cost of false positives vs. false negatives.

The first contribution of our paper is the analysis of the complete ROC curves shown in Figure 4. From these curves it is clear that the class of machine learning algorithm that is chosen to conduct the classification is not critical to the success of the algorithm. There is a slight bias toward AdaBoost(NB) when the cost of false positives is higher than false negatives. The overall measures of area under the curve (AUC) for SVM, RF and AdaBoost(NB) are 0.5937, 0.6264 and 0.6403, respectively. With regard to AUC, AdaBoost(NB) performs best while SVM is worst. A t-test validated that the differences between these algorithms were significant with $p << 0.01$.

Based on the nature of CSGMs, we hypothesized that adding features which captured symmetry in motion would be helpful in recognizing CSGMs more accurately. We conducted a comparison of models trained with and without them. The AUC for SVM, RF and AdaBoost(NB) with the additional features are 0.5756, 0.6266, 0.6388. AdaBoost(NB) still performed the best and SVM performed worst. These dif-

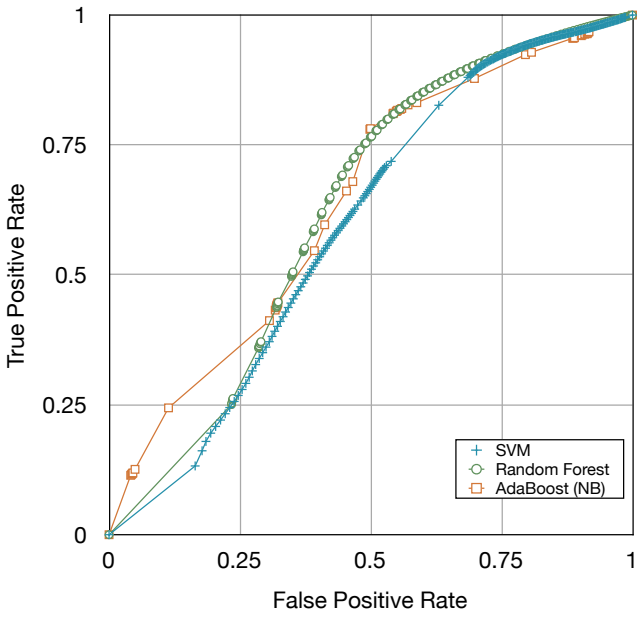


Figure 4. ROC curves of SVM, Random Forest and AdaBoost(NB)

ferences are again statistically significant. Comparing the AUC of two cases (with and without new features) for the same ML technique, we find that adding these new features, while causing a statistically significant difference, did not noticeably or consistently improve performance.

MODELING DURATION

CSGM vs. Non-CSGM

One of the key short-comings in previous work by Singh, et.al. [41] was a lack of any temporal modeling of the CSGMs. Using a simple sample based classifier didn't take into account any limits on switching CSGM guesses from one sample to the next. We improved this by specifically modeling the duration of a CSGM. We also chose to model the gap of time between CSGMs, which we called non-CSGMs.

There are 98 CSGM segments and 100 non-CSGM segments in all 10 babies' data. The distribution of their durations are shown in Figure 5 and Figure 6 (CSGM:$\mu = 14.5, \sigma^2 = 189.6$; non-CSGM: $\mu = 334.9, \sigma^2 = 636310$). Discrete Markov Chains assume geometric distribution of data, however, it's not clear from the data that the CSGM and non-CSGM distributions are best modeled this way. This caused us to find an alternative solution, which led us to consider a class of Phase-Type distributions called Erlang-Cox distributions.

Phase-Type Distributions

A Phase-Type distribution is the distribution of the absorption time in a continuous time Markov chain [33]. Approximating general distributions (in our case CSGM and non-CSGM duration distributions) by phase-type (PH) distributions is a popular technique in stochastic analysis, since PH distributions are often analytically tractable due to their Markovian properties.

It is known that a general distribution G is in a tractable subset of PH, called PH_3, iff its normalized moments satisfy $m_3^G > m_2^G > 1$, where m_3^G, m_2^G are the third and second normalized moments of distribution G. This is true in general for any non-negative distribution. In our case this holds as the third and second normalized moments for CSGMs are 3.01 and 1.89, while those for non-CSGMs are 13.57 and 5.30. Since $3.01 > 1.89, 13.57 > 5.30$, both CSGM and non-CSGM duration distributions belong to PH_3.

MERGING EC AND DBNS

An n-phase EC (Erlang-Coxian) distribution (Figure 7) is a convolution of an (n-2)-phase distribution and a 2-phase $Coxian^+$ distribution possibly with mass probability at zero. The set of EC distributions is general enough, however, that for any distribution, in PH_3, a minimal closed form solution can be derived for all the parameters in Figure 7 [33],

$$(n, p, \lambda_Y, \lambda_{X1}, \lambda_{X2}, p_X) \quad (1)$$

Although these parameters are calculated in closed-form, they do not have simple mappings to the data that we collected. Regardless, such a model is simply a Markov Model and it is possible to solve it using existing Baum-Welch and Viterbi solvers without any additional overhead. Yet at the same time one gets some of the benefit of the Hidden Semi-Markov Model's ability to model durations.

Because of this we decided to evaluate modeling CSGM and Non-CSGM durations with EC models embedded within DBNs. To accomplish this it is necessary first to take the parameters of the EC model above, which are expressed as rates, and convert them to probabilities based on the frequency of observations. Our sampling rate was approximately 19Hz. The next step was to embed the EC model into a Dynamic Bayesian network.

Our technique for accomplishing this can be seen in Figure 8. In this figure, each column represents a time step. The single observable variable is the gray circle which represents the confidence of the AdaBoost(NB) classification of the accelerometer data features taken at that time step. AdaBoost(NB) was chosen because it was the top performing non-temporal method we evaluated. The hidden variables in the system are the boxes $1..n$ (n from Equation 1) which each represent one state of the EC model at time t. The transition probability $f(\lambda)$ is the function of transition rate λ and time between samples, δ_t: $f(\lambda) = 1 - e^{-\lambda\delta_t}$. At each time-step every state of the EC model is rolled out. The probabilities of transitioning between states in the EC model are incorporated into the temporal transitions in the DBN. The transition parameters can be set by closed form calculations based on the observed CSGM durations shown in Figure 5. There is a deterministic dependency that causes the AdaBoost(NB) circle to be true if any of the states $1..n$ are true.

There is a parallel set of states that represent the transition through a non-CSGM event. All of these states are abstracted into the box labelled NONCSGM, which has a parallel structure as the DBN for the CSGM. Any transition out of the

phase Coxian$^+$ PH distribution is known to well represent any distribution that has high second and third moments (any distribution G that satisfies $m_2^G > 2$ and $m_3^G > (3/2)m_2^G$) [19]. However a Coxian$^+$ PH distribution requires more phases for approximating distributions with lower second and third moments. For example, a Coxian$^+$ PH distribution requires at least n phases to well represent a distribution G with $m_2^G \leq (n^2 + 1)/n$ for $n \geq 1$ (see Section 3). The large number of phases needed implies that many free parameters must be determined, which implies that any algorithm that tries to well represent an arbitrary distribution using a minimal number of phases is likely to suffer from computational inefficiency.

By contrast, an n-phase Erlang distribution has only two free parameters and is also known to have the least normalized second moment among all the n-phase PH distributions [1,16]. However the Erlang distribution is obviously limited in the set of distributions which it can well represent.

By combining the Erlang distribution with the two-phase Coxian$^+$ PH distribution, we can represent distributions with all ranges of variability, while using only a small number of phases. Furthermore, the fact that the EC distribution has a small number of parameters $(n, p, \lambda_Y, \lambda_{X1}, \lambda_{X2}, p_X)$ allows us to obtain closed from expressions for these parameters that well represent any given distribution in $\mathcal{PH}_2$.

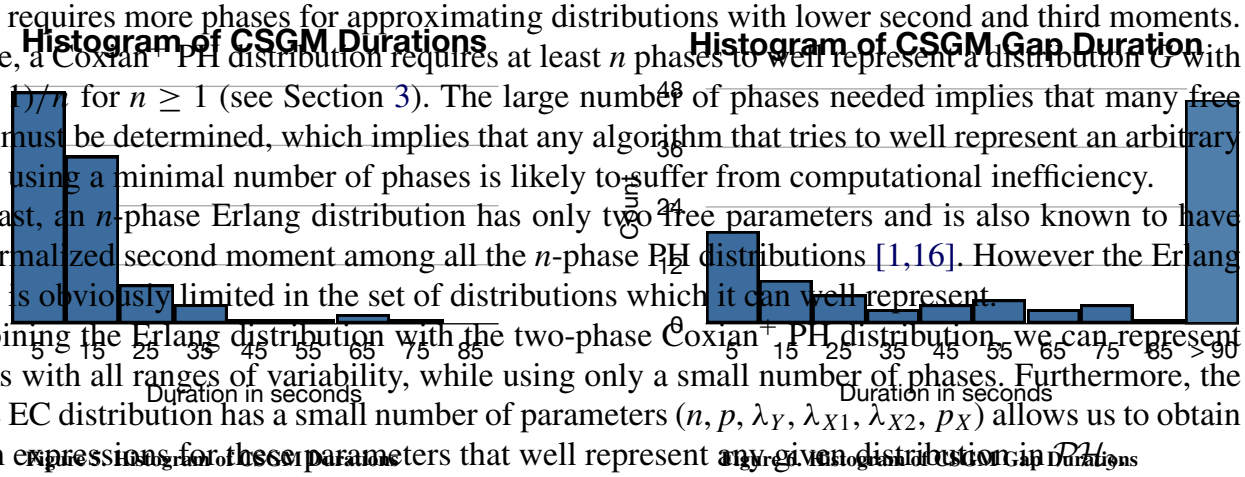


Figure 5. Histogram of CSGM Durations

Figure 6. Histogram of CSGM Gap Durations

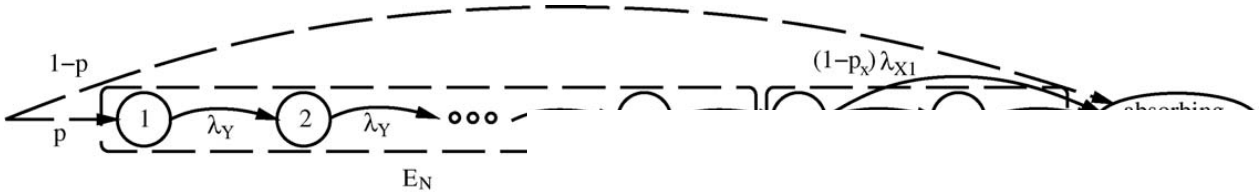


Fig. 2. The Markov chain whose absorption time defines an Erlang-N distr

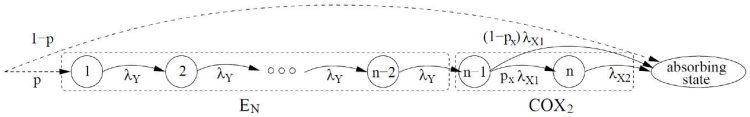
Figure 7. The Markov chain whose absorption time defines a two-phase Coxian$^+$ I states.

Model	AUC
AdaBoost(NB)	0.649
EC Model	0.7001
Smoothed AdaBoost(NB)	0.745
Smoothed EC Model	0.745

Table 1. Area under the ROC curve.

CSGM box goes into the first state of the NONCSG and vice versa.

In our experiments using this model we first set the eters of the model to the closed-form solutions, th allowed all the parameters to converge using Expec Maximization training.



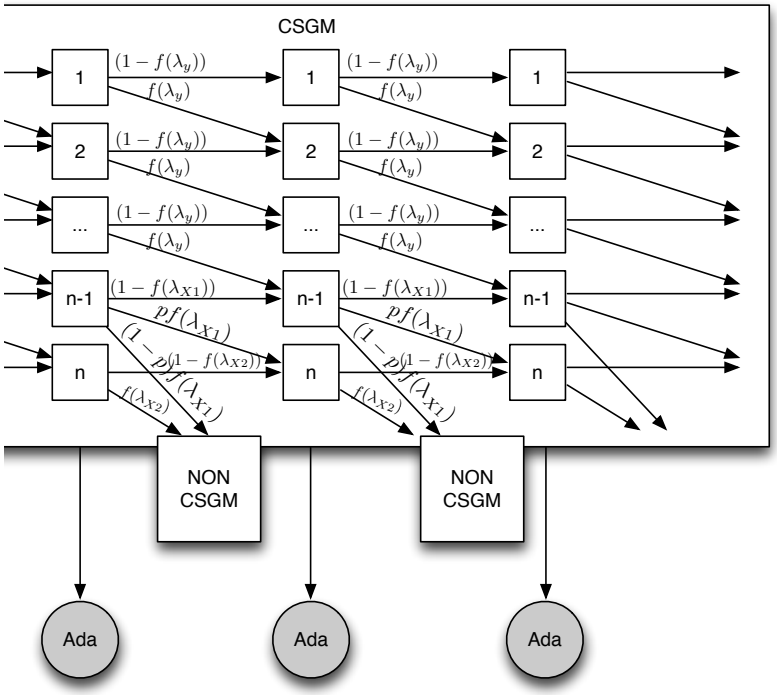


Figure 8. Dynamic Bayesian Network with EC embedded

RESULTS

With these modeling details in place it was possible t uate their performance. The overall goal of this next experiments was to model duration to improve reco accuracy by smoothing the classifications of the und machine learning algorithms.

Accuracy-Based Results

Figure 9 shows as a baseline the best performing non-t machine learning algorithm in blue, AdaBoost(NB). A out in Table 1 the area under the ROC curve was 0.64 like AdaBoost, a DBN does not typically produce cont values because its goal is to determine the maximum hood path through state space given the observations. The expected experimental result for each time step is to only

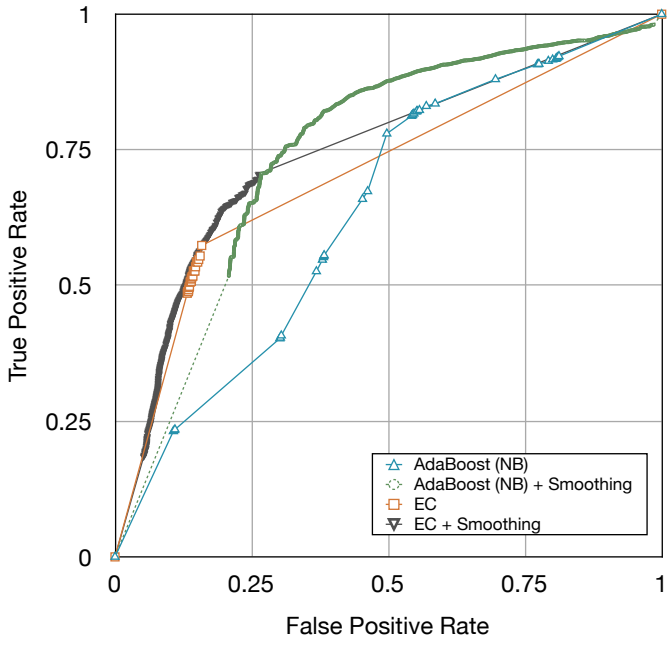


Figure 9. ROC curves generated by modeling duration and/or smoothing

deduce a binary CSGM or No CSGM classification. The resulting ROC curve would be degenerate. To create a better comparison, for each training fold of the EC Model we generated 35 models by sampling 7 babies out of 9 to train on. The we tested the 35 resulting models on the held out baby's data and averaged the resulting guesses at each time step. This data is shown on the figure in green. The relatively small number of possible average confidence values is reflected in the small number of tightly bunched points.

This results of the EC model confirmed our hypothesis that explicitly modeling duration produced improvements in accuracy. The resulting AUC was 0.7001.

However, a simpler way of smoothing the output of the machine learning guesses would be to average the confidence values from AdaBoost over a window of time. The green dotted line in Figure 9 ("AdaBoost(NB)+Smoothing") shows this result when using an averaging window of 14s, the mean CSGM duration. The Area-Under-the-Curve increases to 0.745.

Finally, we applied the same smoothing to the EC Model and found that it scored the same as the smoothed AdaBoost result. The resulting curve is shown in gray ("EC+Smoothing"). The results show different biases in that the smoothed EC model emphasizes low false positives and the window-based smoothing emphasizes low false negatives. Both results are much better than the classification that did not consider duration of CSGMs at all.

Cost-Based Results

As Ward et. al. point out, continuous activity recognition can be evaluated in a wide variety of ways which are not all time sample based [43]. In order to evaluate our results in a different way we referred back to the original clinical motivation for this work of providing a cheaper way to evaluate high-risk babies for CSGMs. The primary cost in the system currently comes from the time required for specially trained experts to review video data for evidence of CSGMs. In light of this, we conducted an analysis of the cost-savings our system could provide.

We propose that rather than viewing the full hour of data that was recorded for each baby, we only ask the expert to review the portions of the video data in which our accelerometer based assessment predicted the presence of a CSGM. This would focus the expert video reviewer on evaluating the most likely portions of the video in which a CSGM may be present.

In order to conduct this assessment we used the EC model to label the data of each baby. Cross-validation was conducted on a baby-by-baby basis in contrast to the work in [41] in which cross-validation was done by random sampling of individual samples. This resulted in the video being segmented into sequentially alternating estimates of CSGMs and non-CSGMS.

The ground truth also had sequentially alternating estimates of CSGMs and non-CSGMS. For each prediction, we divided the corresponding ground truth into segments at the end-points (see Figure 10). Each predicted segment became one sample. If a prediction of non-CSGM overlapped with a section of ground truth with no CSGMs, that was considered a *true negative*. If a prediction of CSGM contained or overlapped one or more sections of ground truth with a CSGM that was considered a *true positive*. This was based on the presumed clinical outcome that the expert video review would identify the presence of a CSGM in the predicted segment. Furthermore, if a predicted CSGM overlapped a ground-truth CSGM, then we did not allow the ground-truth CSGM to penalize the next sequential prediction of non-CSGM, as predicting the exact boundaries of a CSGM is not clinically relevant. If our system predicted that there was no CSGM present and there was a CSGM in the ground-truth we considered that a *false negative*. Finally, if we predicted a CSGM and there was no CSGM in the ground truth, that was considered a *false positive*.

The results were very encouraging. In Figure 11, the amount

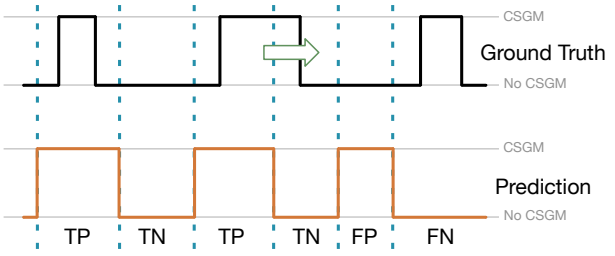


Figure 10. Scoring for the Cost Analysis

	1	2	3	4	5	6	7	8	9	10
	18.0%	11.9%	45.1%	4.3%	5.0%	27.2%	4.7%	40.1%	14.9%	5.7%

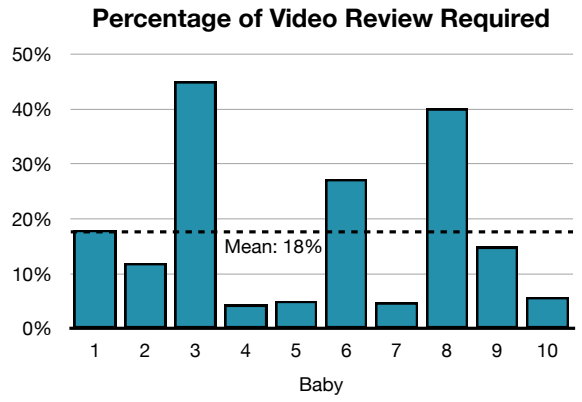


Figure 11. Time savings for each individual baby's video review

Measure	Value (Std)
Sensitivity(Recall) TP/(TP+FN)	0.72 (0.37)
False Positive Rate FP/(FP+TN)	0.43 (0.09)
Specificity TN/(FP+TN)	0.57 (0.09)
Precision TP/(TP+FP)	0.24 (0.30)

Table 2. Confusion matrix calculations.

of video that is required to be reviewed by an expert after classification by our system is shown. On average the video review time is reduced to 18% of the original, so instead of reviewing an hour of video, an expert is only required to watch 11 minutes.

Such a reduction in video viewing time, comes at a cost of some accuracy, however which is demonstrated by the trade-offs in various measures of the confusion matrix as shown in Table 2. In our data we identified 72% of the actual CSGMs in the data (Recall). 24% of the CSGMs that the expert video scorer would be required to view in our approach would be actual CSGMs (Precision). Notably, in the case of baby 8 (representing 10% of the participants), we narrowly missed catching any of the 6 CSGMs that were present in the data. This baby would have appeared to be healthy because our system did not present a portion of video with the CSGM present to the expert. The data collection for this baby had unusual transmission problems, however, which can be seen in Figure 3 and may be the source of the error.

CONCLUSIONS

In this paper we conducted a rigorous analysis of accelerometer data collected from high-risk babies in the Newborn Intensive Care Unit. Our goal was to support the automatic detection of Cramped Synchronized General Movements which are correlated with eventual diagnoses of Cerebral Palsy. To accomplish this we used expert video review as the ground truth and applied explicit duration modeling techniques to the data.

We showed results comparing the accuracy of three classes of machine learning techniques without temporal modeling. By conducting a complete Area-Under-the-Curve analysis we showed that AdaBoost applied to Naive Bayes classifiers is a highly accurate classifier. Although we expected that CSGMs would be more easily recognizable if symmetric features were explicitly given to the machine learning algorithm, we found that our method of doing this did not create an improvement in accuracy.

To improve accuracy further we demonstrated that explicitly considering durations of CSGMs was valuable. We showed a way of using Erlang-Cox models to build Dynamic Bayes Networks that can model the first three moments of any duration directly while still being tractable with well-understood algorithms such as the Viterbi algorithm. This approach has the added strength of being able to set the transition probabilities in closed-form.

We compared the EC/DBN model to an averaged window model applied to the AdaBoost algorithm and found that the two algorithms were equivalent in terms of AUC performance, but showed markedly different biases toward reporting false positives and false negatives.

Our study is limited in that it was based on 10 babies. In future work we will be extending the generalizability of the results by applying these techniques to a larger patient population. It should be emphasized that these babies were already at high-risk of developing CP as a result of our selection criteria. As a result, the babies are representative of that population, but not of a population of babies born at term. Our study was also limited in that it was attempting to model the expert video reviewer's annotations rather than a directly measurable quantity. Errors by the scorer would negatively effect our modeling techniques. We suspect that our data has higher reliability and more information content than the expert video scorer's rating, but we have yet to clearly demonstrate this.

Finally we showed that these techniques could have clinical impact. In current U.S. medical practice there is no established use of Prechtle's method to evaluate high-risk babies. A fundamental reason for this is because the cost of having an expert video scorer review an hour of video for evidence of CSGMs is prohibitive. Furthermore, while previous studies have demonstrated inter-rater reliability of the Prechtle method for discovering CSGMS and a high correlation exists between CSGMs and CP [42], there is no well-established therapy for mitigating CP in newborns. Nonetheless unless newborns can be diagnosed, developing a therapy will be difficult.

In our study, our abbreviated video review process identified at least one CSGM in 5 of the 6 children manifesting CSGMs. On average it required 11 minutes ($\sigma = 9$) per baby of video review for *all* babies to discriminate between true and false positives. In a statistically equivalent population of high-risk babies, this would equate to 83% recall of babies manifesting CSGMs. This approach simultaneously reduces the amount of expert video scoring by 82%, significantly lowering the cost of video review. Although this means that 17% of babies exhibiting CSGMs would be missed by our procedure, the cost of adding full video review to clinical

practice prevents the use of the Prechtle method: 0% of children with CSGMs are currently being detected now.

ACKNOWLEDGEMENTS

This research was supported by the National Institutes of Health (R01 HL-110163), the National Center for Research Resources and the National Center for Advancing Translational Sciences (UL1 TR000153). Additional thanks to Julia Rich and the NICU staff at UCI. The content is solely the responsibility of the authors.

REFERENCES

1. R. Amstutz, O. Amft, B. French, A. Smailagic, D. Siewiorek, and G. Troster. Performance analysis of an hmm-based gesture recognition using a wristwatch device. In *CSE*, volume 2, pages 303 –309. IEEE Computer Society, 29-31 2009.

2. D. Ashbrook and T. Starner. MAGIC: a motion gesture design tool. In *CHI*, pages 2159–2168, New York, NY, USA, 2012. ACM.

3. D. M. Ashton, H. C. r. Lawrence, N. L. r. Adams, and A. R. Fleischman. Surgeon general's conference on the prevention of preterm birth. *Obstetric Anesthesia Digest*, 30(2), 2010.

4. M. A. Babcock, F. V. Kostova, D. M. Ferriero, M. V. Johnston, J. E. Brunstrom, H. Hagberg, and B. L. Maria. Injury to the preterm brain and cerebral palsy: Clinical aspects, molecular mechanisms, unanswered questions, and future research directions. *Journal of Child Neurology*, 24(9):1064–1084, 2009.

5. L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In A. Ferscha and F. Mattern, editors, *Pervasive*, volume 3001 of *Lecture Notes in Computer Science*, pages 1–17. Springer, April 2004.

6. H. Brashear, T. Starner, P. Lukowicz, and H. Junker. Using multiple sensors for mobile sign language recognition. In S. Feiner and D. Mizell, editors, *ISWC*, page 45, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

7. L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

8. E. C. Cameron, V. Maehle, and J. Reid. The effects of an early physical therapy intervention for very preterm, very low birth weight infants: a randomized controlled clinical trial. *Pediatr Phys Ther*, 17(2):107–119, Summer 2005.

9. P. H. Casey, R. H. Bradley, L. Whiteside-Mansell, K. Barrett, J. M. Gossett, and P. M. Simpson. Effect of early intervention on 8-year growth status of low-birth-weight preterm infants. *Arch Pediatr Adolesc Med*, 163(11):1046–1053, 2009.

10. C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

11. J. Cheng, O. Amft, and P. Lukowicz. Active capacitive sensing: Exploring a new wearable sensing modality for activity recognition. *Pervasive Computing*, pages 319–336, 2010.

12. T. Choudhury, J. Lester, and G. Borriello. A hybrid discriminative/generative approach for modeling human activities. In L. P. Kaelbling and A. Saffiotti, editors, *IJCAI*, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

13. G. Cioni, F. Ferrari, C. Einspieler, P. B. Paolicelli, M. T. Barbani, and H. F. R. Prechtl. Comparison between observation of spontaneous movements and neurologic examination in preterm infants. *The Journal of pediatrics*, 130(5):704–711, 05 1997.

14. D. P. Cliff, J. J. Reilly, and A. D. Okely. Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0-5 years. *J Sci Med Sport*, 12(5):557–567, Sep 2009.

15. L. E. Dunne, S. Brady, R. Tynan, K. Lau, B. Smyth, D. Diamond, and G. M. P. O'Hare. Garment-based body sensing using foam sensors. In *AUIC '06: Proceedings of the 7th Australasian User Interface Conference*, pages 165–171, Darlinghurst, Australia, Australia, 2006. Australian Computer Society, Inc.

16. T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *CVPR (1)*, pages 838–845. IEEE Computer Society, 2005.

17. F. Ferrari, G. Cioni, C. Einspieler, M. F. Roversi, A. F. Bos, P. B. Paolicelli, A. Ranzi, and H. F. R. Prechtl. Cramped synchronized general movements in preterm infants as an early marker for Cerebral Palsy. *Arch Pediatr Adolesc Med*, 156(5):460–467, 2002.

18. Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In L. Saitta, editor, *ICML*, pages 148–156. Morgan Kaufmann, 1996.

19. F. Giganti, G. Cioni, E. Biagioni, M. T. Puliti, A. Boldrini, and P. Salzarulo. Activity patterns assessed throughout 24-hour recordings in preterm and near term infants. *Developmental Psychobiology*, 38(2):133–142, 2001.

20. M. Hadders-Algra, K. R. Heineman, A. F. Bos, and K. J. Middelburg. The assessment of minor neurological dysfunction in infancy using the touwen infant neurological examination: strengths and limitations. *Dev Med Child Neurol*, 52(1):87–92, Jan 2010.

21. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

22. F. Heinze, K. Hesels, N. Breitbach-Faller, T. Schmitz-Rode, and C. Disselhorst-Klug. Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy. *Med Biol Eng Comput*, 48(8):765–772, Aug 2010.

23. N. Kern, S. Antifakos, B. Schiele, and A. Schwaninger. A model for human interruptability: Experimental evaluation and automatic estimation from wearable sensors. In M. Smith and B. H. Thomas, editors, *ISWC*, volume 1, pages 158–165. IEEE, October 2004.

24. S. J. Korzeniewski, G. Birbeck, M. C. DeLano, M. J. Potchen, and N. Paneth. A systematic review of neuroimaging for cerebral palsy. *Journal of Child Neurology*, 23(2):216–227, 2008.

25. I. Krägeloh-Mann and C. Cans. Cerebral palsy update. *Brain and Development*, 31(7):537 – 544, 2009.

26. P. O. Kristensson, T. Nicholson, and A. Quigley. Continuous recognition of one-handed and two-handed gestures using 3d full-body motion tracking sensors. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 89–92, New York, NY, USA, 2012. ACM.

27. J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive Computing and Communications, IEEE International Conference on*, 0:1–9, 2009.

28. K. Lyons, H. Brashear, T. Westeyn, J. Kim, and T. Starner. GART: The gesture and activity recognition toolkit. *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments*, pages 718–727, 2007.

29. P. Mistry, P. Maes, and L. Chang. WUW - wear ur world: a wearable gestural interface. In D. R. O. Jr., R. B. Arthur, K. Hinckley, M. R. Morris, S. E. Hudson, and S. Greenberg, editors, *CHI Extended Abstracts*, pages 4111–4116, New York, NY, USA, 2009. ACM.

30. U. Nodelman, C. R. Shelton, and D. Koller. Expectation maximization and complex duration distributions for continuous time bayesian networks. In *UAI*. AUAI Press, 2005.

31. S. Ohgi, S. Morita, K. K. Loo, and C. Mizuike. Time series analysis of spontaneous upper-extremity movements of premature infants with brain injuries. *Phys Ther*, 88(9):1022–33, Sep 2008.

32. N. Oliver, B. Rosario, and A. Pentland. Graphical models for recognizing human interactions. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *NIPS*, pages 924–930. The MIT Press, 1998.

33. T. Osogami and M. Harchol-Balter. Closed form solutions for mapping general distributions to quasi-minimal ph distributions. *Perform. Eval.*, 63(6):524–552, 2006.

34. C. Park and P. Chou. Eco: Ultra-wearable and expandable wireless sensor platform. In *BSN*, pages 165–168. IEEE Computer Society, 3-5 2006.

35. D. J. Patterson, D. Fox, H. A. Kautz, and M. Philipose. Fine-grained activity recognition by aggregating abstract object usage. In K. Mase and B. Rhodes, editors, *ISWC*, pages 44–51, Osaka, Japan, October 2005. IEEE Computer Society.

36. H. F. Prechtl. State of the art of a new functional assessment of the young nervous system. an early predictor of cerebral palsy. *Early Hum Dev*, 50(1):1–11, Nov 1997.

37. H. F. Prechtl, C. Einspieler, G. Cioni, A. F. Bos, F. Ferrari, and D. Sontheimer. An early marker for neurological deficits after perinatal brain lesions. *The Lancet*, 349(9062):1361 – 1363, 1997.

38. A. V. Rowlands. Accelerometer assessment of physical activity in children: an update. *Pediatr Exerc Sci*, 19(3):252–266, Aug 2007.

39. T. S. Saponas, D. S. Tan, D. Morris, R. Balakrishnan, J. Turner, and J. A. Landay. Enabling always-available input with muscle-computer interfaces. In A. D. Wilson and F. Guimbretière, editors, *UIST*, pages 167–176, New York, NY, USA, 2009. ACM.

40. T. Schlömer, B. Poppinga, N. Henze, and S. Boll. Gesture recognition with a wii controller. In *Proceedings of the 2nd international conference on Tangible and embedded interaction*, TEI '08, pages 11–14, New York, NY, USA, 2008. ACM.

41. M. Singh and D. J. Patterson. Involuntary gesture recognition for predicting cerebral palsy in high-risk infants. In K. V. Laerhoven, K. ho Park, and H.-J. Yoo, editors, *ISWC*, pages 61–68. IEEE, Oct 2010.

42. A. J. Spittle, L. W. Doyle, and R. N. Boyd. A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life. *Dev Med Child Neurol*, 50(4):254–66, 2008.

43. J. A. Ward, P. Lukowicz, and H.-W. Gellersen. Performance metrics for activity recognition. *ACM TIST*, 2(1):6, 2011.

44. J. O. Wobbrock, A. D. Wilson, and Y. Li. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, UIST '07, pages 159–168, New York, NY, USA, 2007. ACM.

45. J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li. Gesture recognition with a 3-d accelerometer. In *Proceedings of the 6th International Conference on Ubiquitous Intelligence and Computing*, UIC '09, pages 25–38, Berlin, Heidelberg, 2009. Springer-Verlag.

46. S.-Z. Yu. Hidden semi-markov models. *Artificial Intelligence*, 174(2):215 – 243, 2010.