Computer Vision to Automatically Assess Infant Neuromotor Risk

Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R. Pierce,
Daniel K. Bogen, *Member, IEEE*, Laura Prosser, Michelle J. Johnson, *Member, IEEE*,
and Konrad P. Kording

Abstract— An infant's risk of developing neuromotor impairment is primarily assessed through visual examination by specialized clinicians. Therefore, many infants at risk for impairment go undetected, particularly in under-resourced environments. There is thus a need to develop automated, clinical assessments based on quantitative measures from widely-available sources, such as videos recorded on a mobile device. Here, we automatically extract body poses and movement kinematics from the videos of at-risk infants (N = 19). For each infant, we calculate how much they deviate from a group of healthy infants (N = 85 online videos) using a Naïve Gaussian Bayesian Surprise metric. After pre-registering our Bayesian Surprise calculations, we find that infants who are at high risk for impairments deviate considerably from the healthy group. Our simple method, provided as an open-source toolkit, thus shows promise as the basis for an automated and low-cost assessment of risk based on video recordings.

Index Terms— Automatic testing, computer vision, movement science, neuromotor disorders, pediatric risk assessment, pose estimation, YouTube.

Claire Chambers, Nidhi Seethapathi, and Daniel K. Bogen are with the Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: chambers.claire@gmail.com; snidhi@seas.upenn.edu; dan@seas.upenn.edu).

Rachit Saluja is with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: rsaluja@seas.upenn.edu).

Helen Loeb is with the Center for Injury Prevention, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA (e-mail: loebh@email.chop.edu).

Samuel R. Pierce is with the Physical Therapy Department, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA (e-mail: pierces1@email.chop.edu).

Laura Prosser is with the Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104 USA, and also with the Division of Rehabilitation, Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA (e-mail: prosserl@email.chop.edu).

Michelle J. Johnson is with the Department of Physical Medicine and Rehabilitation, University of Pennsylvania, Philadelphia, PA 19104 USA, also with the Rehabilitation Robotics Laboratory, University of Pennsylvania, Philadelphia, PA 19104 USA, and also with the General Robotics, Automation, Sensing and Perception (GRASP) Laboratory, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: johnmic@pennmedicine.upenn.edu).

Konrad P. Kording is with the Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104 USA, and also with the Department of Neuroscience, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: kording@seas.upenn.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2020.3029121

I. INTRODUCTION

DEVELOPMENTAL disorders, including those caused by neuromotor disease, are the most common source of childhood disability, affecting 5-10% of children and 3.7 to 7.4 million American children [1], [2] and are often the cause of lifelong disability. Early intervention may improve outcomes in neuromotor disease, but requires accurate early identification of infants at risk for physical disability [3], [4]. For effective widespread early identification, easily-deployed automated risk assessment tools are needed to quantify infant movement in the first few months of life during the development of motor control.

Many factors describe a 'good' diagnostic test. A test's predictive ability is key. Beyond that, it is desirable that tests are based on quantitative measurements and a series of well-defined steps. This motivates the development of quantitative tests to supplement qualitative clinical judgment. Second, it is important to evaluate the accessibility of a test. Clinical assessments often involve expert judgment and expensive equipment, which are only accessible in highly-resourced environments. This impacts families of limited means and in low-resource countries, where the burden of disability is higher [5], [6]. Even in highly-resourced environments, a test that is accessible can enable continuous monitoring of disorder. The pipeline we introduce in this article provides a means to achieve such quantitative and accessible evaluation.

Many clinical tests have been developed to assess neuromotor risk. The General Movements Assessment and the Hammersmith Infant Neurological Examination have high sensitivity and specificity [7]. These tests are effective at detecting disorder early, at less than four months corrected age [3]. The Test of Infant Motor Performance has been shown to detect the changes in the movement patterns pre- and post- treatment for infants later diagnosed with cerebral palsy [8]. Despite their high accuracy, these clinical tests have shortcomings. First, they require expert administration [9], [10] and specialized training of licensed clinicians. Thus, these clinical tests are predictive, but are qualitative and have limited accessibility.

The past two decades have seen the development of sensor-based measurements of infant movement to quantify movement features. Wearable sensors and 3-D motion capture have been used to measure infant movement in laboratory settings [11]–[18]. Sensor-based measurements provide an objective and quantitative means for risk assessments. However, these measurements are restricted to laboratory and clinical

settings, thus hindering the ability to evaluate infants in their natural environments. Moreover, sensor-based methods have limited access as they are costly and time-intensive. Thus, sensor-based approaches to diagnostics can provide quantified assessment, but have limited deployability and accessibility.

In the last decade, scientists have developed systems for video-based assessments using optical flow which hold promise to be widely accessible as they can be implemented on mobile devices [19]–[21]. Such methods typically use frame differencing of a video to estimate movements by tracking the centroid of motion. This technique can be extended to the measurement of movement of each limb [19]. Using this approach, the amount of movement and the frequency of movement have predicted clinical outcomes [20], [22]. However, such frame difference metrics rely on centroid estimates that only measure centroidal movements as opposed to measuring individual joint or limb segment movements. The extraction of limb movement from optic flow also requires careful parameter selection and manual adjustment of the tracking algorithm. As an alternative, markerless methods have been developed for the tracking of infant movement [23], [24]. However, these methods require depth images, are computationally intensive and the models are likely to be overfit to the relatively small datasets on which they are trained. In sum, video-based diagnostics provide quantitative metrics that have the potential to be widely accessible, but to be successfully applied, they must be effective on easily obtained 2-D videos, extract movement features of individual body-parts, and require minimal manual tuning.

Here, we provide a first step toward making video-based neuromotor assessment widely available (Figure 1). We produce a 'normative' reference database of infant movements using 85 videos found online. We calculate the movements of the body parts using an existing pose-estimation algorithm, OpenPose [25], which we augmented using domain adaptation with our own labelled dataset of infant pose consisting of 9039 infant images. Using this normative database we can calculate how much each infant deviates from the typical movements of healthy infants using a single score, the Naïve Gaussian Bayesian Surprise [26]. When we then tested this system on a clinical population (N = 19) where the level of neuromotor risk was assessed by a clinician (low, moderate, and high), we found that Bayesian Surprise varied across participant groups. Thus, we have developed an open-source framework that calculates the Bayesian Surprise of movement features given a reference database of 'normative' infants.

II. METHODS

Ethical approval for this study was provided by the University of Pennsylvania Institutional Review Board (IRB protocol number 822487) and the Children's Hospital of Philadelphia Institutional Review Board (IRB protocol number 15-012110). The parents of the infants shown in Figures 1, 2, and 4 provided informed, written consent for those images be published.

A. YouTube Data

In order to develop a framework that predicts the neuromotor risk level of infant movement, it is helpful to have an estimate of what is expected or normal. By creating such a 'normative' database, we can then estimate the probability
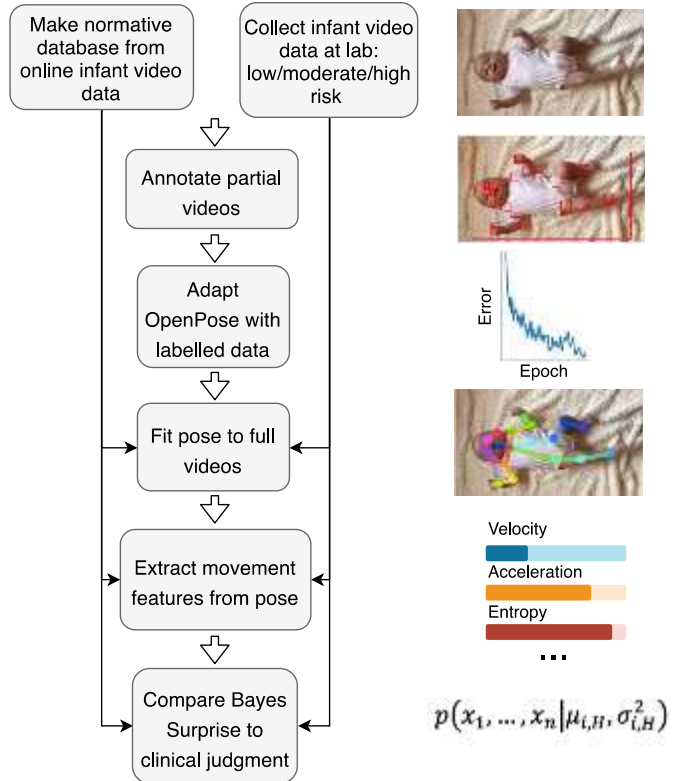

Fig. 1. Flowchart of the pipeline for computer vision-based neuromotor risk assessment. We created a normative database infant movement using videos found online (85 infants) and recorded infants at risk of neuromotor disease in a clinical setting (19 infants). Using video frames labelled with body-part landmarks from a subset of our video dataset, we adapted a pose estimator (OpenPose) to extract the pose of infants which we improved using domain adaptation. Using the adapted system, we then extracted pose from all videos. Next, from the pose data, we quantified kinematic features for each infant. Finally, our neuromotor risk prediction used Naïve Gaussian Bayesian Surprise that estimated the probability that each infant belonged to the reference population.

that the movements of a new infant belong to the normative distribution, thereby assessing the probability that the infant is healthy. In previous work, we have shown that online video databases are a useful source of human movement data [27]. In the present study, we built a normative database of infant movements based on video data of infants from YouTube, assuming that found videos represented healthy infant movement. This assumption is bolstered by the fact that each YouTube video was reviewed by two expert physical therapists to flag any infants whose movements or appearances suggested impairment. None of the videos included in our normative database obtained from YouTube were assessed by the experts to have impairments. Moreover, the curators who were searching for YouTube videos were asked not to include any videos where the video title or description suggested impairment. We used search terms such as 'one-month old', 'two-month old', etc. Search terms that resulted in suitable videos for inclusion in our normative database or for domain adaptation of our pose-estimation system are provided as supplemental data.

In our initial search, we identified 420 video segments featuring infants, where infants were non-occluded, where infants moved independently, and where most of the body was

present in the frame throughout the segment. We also selected videos that did not have significant camera movements. In our analysis, we included 95 of the 420 videos found that were over 5 seconds in duration, and where pose estimates were of sufficient quality for the extraction of basic kinematic variables, as evaluated by visual inspection (mean recording duration per infant = 39 s (SD = 33)). Details about the YouTube videos that were used in the study and URLs for the videos themselves are provided in a document along with this article (meta_data_youtube_dataset.csv). These details are the video URL and title, the search term used to find each video, timing of the video excerpt where the infant was non-occluded, a record of the videos that were annotated, and videos that were included in the healthy reference sample. Two physical therapists estimated the ages of infants (mean rated age in weeks (SD) = 9.67 (6.26), There was reasonable agreement between raters (inter-rater reliability: r =.76, p <.0001, n = 85). The age rating averaged across raters was used to classify infants as being more or less than 10 weeks of age (see Data analysis). When there was more than one video per infant, we included in the dataset the video with the longest duration. This database, containing video recordings of 85 different infants, provided a reasonably large 'normative' database of infant movements against which to compare the kinematic features extracted from infant video data collected at the lab.

B. Clinical Data

Infants tested in person were recruited through The Children's Hospital of Philadelphia (CHOP), University of Pennsylvania (Penn), and the local community. A subset of the infant data reported here were used in a previous study, which focused on the development of a method to track infant movement using a multiple view stereoscopic 3-D vision system [28]. Infants were screened to determine eligibility. Inclusion criteria for full-term infants (born at a gestational age of > 37 weeks) were the absence of any significant cardiac, orthopedic, or neurological condition. Infants born preterm (gestational age < 36 weeks) were recruited from the Newborn/Infant Intensive Care Unit (N/IICU) at CHOP. All infants were between the ages of 3 and 11 months. Infants who could walk were excluded. Parents of eligible participants provided written informed consent. The human subject ethics committee at Penn served as the IRB of record for this study (UPENN IRB # 822487). Data were collected from infants at CHOP, a local childcare facility, and at the Penn Rehabilitation Robotics Lab.

Before testing, an experienced pediatric physical therapist evaluated the infants' level of risk using the Bayley Infant Neurodevelopmental Screener (BINS). This screener includes specific tasks that experts administer to observationally score infants based on the Bayley Scales of Infant Development [29]–[31]. The experts did not have access to the Bayesian Surprise outputs of the algorithm we developed here. We assessed risk of neuromotor dysfunction as classified by the BINS test (low, moderate, high) computed at corrected age for preterm infants and chronological age for full-term infants.
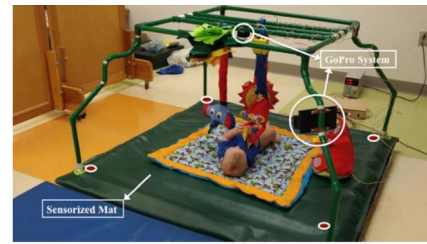

Fig. 2. Infant testing at CHOP in the PANDA gym. The infant was placed at the center of a sensorized mat and was recorded using GoPro cameras.

To assess neurodevelopmental risk, the BINS addresses several areas of ability: basic neurological functions/intactness (posture, muscle tone, movement, asymmetries, abnormal indicators); expressive functions (gross motor, fine motor, oral motor/verbal); receptive functions (visual, auditory, verbal); and cognitive processes (object permanence, goal-directedness, problem solving). The BINS consists of six item sets, each containing 11 to 13 items. Each item in the BINS is scored 'optimal performance' or 'non-optimal performance' using pre-defined rules. Optimal responses for a given item set are summed to give a raw score. For each item set, two previously established raw cut scores identify a given infant's level of risk for neurological impairment, resulting in three risk groupings: low risk, moderate risk, and high risk.

The Play And Neuro-Development Assessment (PANDA) Gym uses toys with sensors, cameras, and a mat structure which measures the center of pressure of the infant (Figure 2). The video data collected by the gym are pairs of color images extracted from high resolution (1920 x 1080) Go Pro Hero 4 Session videos. Two cameras were mounted on four 3-D printed stereo frames and the videos were captured at 30 frames per second, with one setup placed directly above the infant and the other positioned on the infant's right side, mounted on the side frame of the gym's platform. In the current work, we used data from one GoPro camera for each infant. The platform structure is lightweight, made with colored PVC tubing and a sensorized mat (4 × 4 ft), which is developed using a DragonPlate carbon fiber foam core board with four force sensors on each corner. Vinyl and foam padding cover the dragon board to make it comfortable for the infant. Before each infant trial, both systems undergo calibration. A GUI allowed data collection from all sensors and real time tracking of an infant's center of pressure.

The original purpose of the experiment was to examine the infants' interactions with sensorized toys, while their movements were recorded using the sensorized mat and GoPro cameras. Sensorized toys contained inertial measurement units (MPU-9150, InvenSense, San Jose, CA) to measure infants' toy interactions. The experiment included three toy conditions where different toys were hung above the infant (elephant, orangutan, lion, Figure 2) and a fourth baseline condition without the hanging toys (No-toy condition). Since the toys occluded the view of the infant, this caused the pose estimation to fail. Therefore, here, we restricted our analysis to the No-toy

condition, where the infant was non-occluded and moved spontaneously. The PANDA gym experiment provided video recordings that allowed us to examine infants' spontaneous movements.

In each trial, the infant was laid in a supine position inside the PANDA gym and video data were collected using the GoPro cameras. Preterm infants were 2 females and 8 males (mean corrected age (SD) = 14.39 weeks (6.92), mean chronological age (SD) = 24.84 weeks (4.77)). Full-term infants were 17 females and 4 males (mean chronological age (SD) = 26.28 weeks (9.88)). The data of four preterm infants and three full-term infants were excluded due to missing BINS score data. The data of two full-term infants were excluded because infants were sitting during the video, which prevented successful pose estimation. The data of a further three full term infants were excluded due to missing video data.

This left us with a final sample containing 19 infants (see Supplemental Table I at external figshare link). Preterm infants were 1 female and 5 males (mean corrected age (SD) = 17.64 weeks (4.82), mean chronological age (SD) = 26.74 weeks (4.54)). Full-term infants were 10 females and 3 males (mean chronological age (SD) = 24.65 weeks (9.44)) with 5 infants at low-risk, 9 infants at moderate risk, and 5 infants at high risk, as evaluated by the BINS score. There was a mean (SD) total recording duration of 287.57 s (161.02) per infant.

C. Pose Estimation

In order to extract pose information from videos, we used OpenPose, a pose-estimation algorithm [25]. OpenPose consists of convolutional neural networks that have been trained using labelled image data to identify 2-D human joint and limb positions from images. OpenPose extracts positions of nose, neck, ears, eyes, shoulders, elbows, wrists, hips, knees, and ankles from 2-D videos. On a system with multiple graphical processing units, OpenPose can run in real time (~30 frames/second). With this processing speed, OpenPose provides a viable method for performing motion capture from 2-D videos.

Due to differences between adults and infants in their appearance and pose, pose tracking using OpenPose was initially limited in performance. We therefore adapted OpenPose for infants. We created a dataset of infant images with labels of joint positions. We required that infants be non-occluded, and that their full body be in the frame. We included the first 10 s of 107 video segments collected from YouTube that were assessed to be appropriate to train on and the first 30 s of 17 out of the total 19 videos from the clinical dataset. The URLs and timing information of the YouTube video data used for domain adaptation is provided as supplemental data. Videos were selected in order to have a representative set of images for domain adaptation that describe infant pose, in terms of infant appearance, clothing, pose, video background, and image scale. We labelled the video frames with body landmarks using Vatic, a video annotation tool [32]. Videos were labelled to match the format of the Common

Objects in Context dataset [33], with labels for left and right eyes, ears, shoulders, elbows, wrists, hips, knees, ankles, and nose, and a bounding box around all points. We included all frames in the labelled dataset, resulting in a total of 36,030 images with labelled pose (32,417 in the training set, 3,613 in the test set). We initialized domain adaptation with pretrained weights from previous work [25] and updated weights by gradient descent for 50 iterations. This resulted in a pose estimator that could extract an infant's pose from 2-D video.

D. Data Analysis

Using the pose-estimation system adapted for infants, we extracted pose estimates from video recordings of 85 infants in the YouTube cohort and the 19 infants from the clinical sample. From this pose data, we then extracted kinematic features for each infant, then predicted their neuromotor risk using a Naïve Bayes classifier.

When evaluating the pose-estimation model, we used three performance metrics. We computed the root-mean-squared error (RMSE) between ground-truth data and pose estimates, normalized by bounding-box dimensions, for landmarks that were in both datasets. To obtain the RMSE, we computed the distance between each landmark detected by the algorithm and label from the ground-truth data, in x coordinates and y coordinates. To account for differences in scale, we normalized x and y distances by the width and height of the ground-truth bounding box around the infant. Finally, we computed the root-mean square of these individual errors from all labelled images, leading to one error score for the entire test dataset. We also evaluated precision, that is, the proportion of total pose estimates that were present in the ground-truth dataset, and the recall, that is, the proportion of ground-truth labels that were found by the pose estimator. These three metrics allowed us to evaluate the pose-estimation model before and after domain adaptation.

Pose-estimation data contained missing data, outliers in the form of false positive detections and additional noise around body landmarks. To obtain clean signals, we preprocessed time series data from each body landmark (Figure 3). We first removed missing data by applying linear interpolation to the raw time series for each body landmark. We then removed outliers by using a rolling-median filter with a smoothing window of 1 second. In order to obtain smooth signals, we then performed smoothing using a rolling-mean filter with a smoothing window of 1 second. Outlier removal and smoothing provided a cleaner signal for the extraction of kinematic variables.

To compute infant kinematics from YouTube data, we needed to compensate for camera properties. Our goal is to have a tool that compensates for some camera movements, because we want our pipeline to work on videos recorded on a mobile phone outside of a laboratory setting. Movement signals extracted from videos contain a combination of infant movement and camera movement. For example, a moving camera with respect to a non-moving infant would produce a movement signal. Also, a short distance between the camera and infant would produce greater movement signals than when
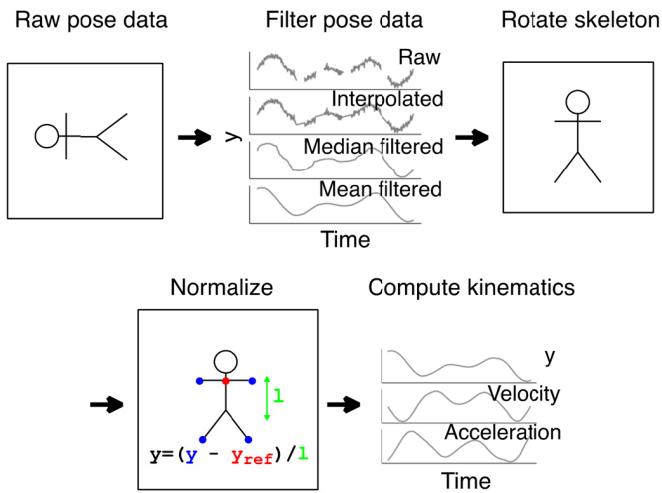
Fig. 3. Preprocessing of pose data. We took raw pose data for whole videos as input (frame coordinates of body landmarks). To filter the pose data, we interpolated the raw signal to replace missing data, then applied a rolling-median filter to remove outliers and finally, used a rolling-mean filter. This provided a smooth signal from which to compute derivatives. To ensure that we could compare infants recorded under different conditions (camera angles, video resolution, etc.), we then rotated and normalized body landmark coordinates in each frame. We rotated the upper-body landmarks with respect to the center of the shoulders and we rotated the lower-body landmarks with respect to the center of the hips. Next, we normalized the landmark coordinates within each frame, by subtracting a reference landmark (the neck) and dividing by a reference distance (the trunk length). Finally, based on pre-processed signals, we computed kinematic variables from selected body landmark coordinates and joint angles (position or angle (y), velocity, acceleration).

further away. To compensate for these differences, we first extracted 2-D joint angles from video frames, which are invariant to the size of the infant in the frame, and the infant's orientation in the frame (i.e. landscape, portrait). Second, we also computed landmark positions, rotated so that all infants from all frames were aligned (Figure 3). We rotated the upper-body landmarks relative to the angle of the left and right shoulders with respect to the midpoint between the shoulders. We rotated the lower-body landmarks relative to the angle of the left and right hip joints with respect to the midpoint between the hip joints. We then normalized landmark coordinates within each frame (Figure 3). We computed distance from a reference landmark (the neck) and divided by a reference distance on the body (trunk length). Measuring movement in body-centered coordinates allowed us to compensate for the effects of camera movement in YouTube videos. Our preprocessing methods compensate for camera angles of 0.004 radians median (range 0 to 0.04) and camera distance of 0.003 median in torso length units (range 0 to 0.125).

Our predictions of infant risk depended on kinematic variables. From each time series of body-landmark positions, we computed the velocity and acceleration at each time interval. Time-series data extracted from pose estimates provided data for computing kinematic features.

We selected a simple set of features based on movement of the arms and legs computed from knee angles, elbow angles, ankle positions, and wrist positions. Our normalization of pose estimates prevented us from examining movement of the shoulders and hips, so we did not include these as kinematic variables. In future, this could be solved by using a fixed camera for all recordings, which would allow us to attribute all recorded movement to the infant and remove the need for normalization. Where possible we computed the median and interquartile range (IQR), to avoid the effects of outliers. We chose features that represented postural information (absolute position and angle), variability of posture (variability of position and angle), velocity of movement (median absolute velocity), variability of movement (variability of velocity and acceleration), complexity (positional and angular entropy), and left-right symmetry of movement (left-right cross correlation of position and angle). As features of each extremity (left and right wrists and ankles), we included the median position in x and y coordinates (units of trunk length, l), IQR of position in x and y coordinates (l), median absolute velocity in x and y coordinates (units of trunk length per second, l/s), IQR of velocity in x and y coordinates (l/s), IQR of acceleration in x and y coordinates (l/s^2), left-right cross-correlation of position, and positional entropy. As features of each joint angle (left and right elbows, wrists, knees and ankles), we included the angular mean (degrees, deg.), angular standard deviation (deg.), median angular velocity (deg./s), IQR angular velocity (deg./s), IQR angular acceleration (deg./s^2), angular cross-correlation, and angular entropy. Entropy is defined as the negative sum of probabilities each multiplied by their logarithm. Generally, a peaked distribution is associated with low entropy and low complexity, while a broader distribution has high entropy or complexity.

To reduce the number of features, we averaged feature values across the left and right sides of the body, where applicable. Having established a basic set of features, we preregistered our feature set and the algorithm to calculate Bayesian Surprise (https://osf.io/hv7tm/). These kinematic features provided basic descriptors of infant movement on which to make predictions.

Existing assessments rely on clinicians visually identifying at-risk infants whose movements differ from the normal population. We aimed to replicate this form of assessment using quantification of multiple kinematic features from videos and a normative database which tells us what healthy infant movement looks like. In this approach, infants who deviate from the healthy reference population are identified as being at-risk. We estimated the probability of each infant belonging to the healthy reference population, represented by the YouTube cohort in this case, who are presumed to be healthy. We adopted the naïve Bayes approach, previously used for clinical assessments of walking performance [26]. Under the assumption of normally-distributed features and feature independence, the joint probability of features over the reference population is:

$$p\left(x_1, \ldots, x_n \mid \mu_{i,H}, \sigma_{i,H}^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i,H}^2}} e^{\frac{-(x_i - \mu_{i,H})^2}{\sigma_{i,H}^2}} \quad (1)$$

where x_i indicates the i-th feature value for a subject, and $\mu_{i,H}$ and $\sigma_{i,H}^2$ are, respectively, the mean and variance of that feature across the reference subjects.

The negative natural logarithm gives the Naïve Gaussian Bayesian Surprise [48]:

$$\Psi = -\sum_{i=1}^{n} \left(\frac{1}{2} \ln \left(2\pi \sigma_{i,H}^2 \right) + \frac{(x_i - \mu_{i,H})^2}{2\sigma_{i,H}^2} \right) \qquad (2)$$

The metric given by (2) can be interpreted as related to the log probability of a subject being part of the reference population. The Bayesian Surprise was normalized with respect to the reference population. Infants were compared within age brackets computed based on corrected age (less than 10 weeks, 10 weeks or older). We chose 10 weeks as a threshold so that there was sufficient data within each age bracket to serve as a normative group for comparison (47 infants less than 10 weeks, 38 infants older than 10 weeks). This provided a standard score: subjects who were more than two standard deviations from the mean would be classed as at-risk. As we do not use any parameters to weight our pre-registered features, we do not need to cross-validate for our Bayesian surprise calculation. Our approach to risk assessment classified infants based on how different their movement was to that of the reference population.

To further explore how combinations of movement features relate to clinically-assessed risk, we applied matrix decomposition to our set of 38 movement features from 104 infants. Matrix decomposition provides a way of describing infants' movement in terms of a lower number of features which are aggregates of the original set of 38 movement features. This method also allows us to summarize relationships between movement features. Singular value decomposition (SVD) is a matrix-decomposition method that finds 'latent variables' which describe important sources of variance in the data. Each latent variable is a linear combination of the original set of movement features, with a singular value that tells us how much of the variance it explains. For example, one latent variable could have a high weight for velocity and acceleration, and a low weight for symmetry and entropy; a second latent variable could have a high weight for symmetry and entropy, and a low weight for the remaining features, and so on. These groupings of features into one latent variable tell us that they vary together. Each infant has one score for each of the latent variables. For example, an infant with high velocity and acceleration, and moderate symmetry and entropy would have a high score for the first latent variable and a moderate score for the second latent variable. Thus, matrix decomposition allows description of infants' movement in terms of a lower number of features.

More formally, applied to our dataset of M infants by N movement features ($A_{M \times N}$), SVD approximates this matrix by the product of three matrices: $A = U \Sigma V^T$. The columns of the left matrix, U, span the column space of matrix A that characterizes the infants, later described as singular vectors. The columns of the right matrix, V, span the row space of matrix A that characterizes the movement features, later described as singular vector weights. Σ contains singular values that describe how important each latent variable is. We first normalized columns of matrix A, by subtracting the mean and dividing by the standard deviation. We then applied SVD. We examined the singular vectors of U as a function of risk and we examined how the original movement features are weighted in latent variables by examining singular vectors of V. SVD allows examination of infants and movement features in terms of a set of latent variables that describe the sources of variance in the dataset.

III. RESULTS

We developed a system to assess infants' neuromotor risk based on kinematic variables extracted from 2-D videos. We adapted an existing adult pose-estimation algorithm to detect infant movements. Then, we computed kinematic data extracted from clinically assessed infants and a normative infant movement database we curated from YouTube. Finally, we combined these features into a single estimate of risk for each infant.

We required a system that could reliably extract the pose of infants from videos. We therefore adapted an existing deep-learning based pose-estimation algorithm, Open-Pose [25], for infants. We performed domain adaptation for infant pose estimation by updating the network's weights with infant images and body landmark labels as input. Domain adaptation led to improved pose estimation in infants. We observe a lower distance between ground-truth labels and pose estimates after domain adaptation, as shown by the RMSE in bounding-box units. RMSE of 0.05 before domain adaptation decreases to 0.02 after domain adaptation (Figure 4A,B). We also observe an increase in precision from 0.89 before domain adaptation to 0.92 after domain adaptation (Figure 4C,D) and an increase in recall from 0.76 before domain adaptation to 0.94 after domain adaptation (Figure 4E,F). Therefore, through domain adaptation with labelled images of infants, we improved the performance of the pose-estimation model. This allowed us to extract skeletal information on each infant and track the positions of body landmarks (Figure 4 G,H). The adapted pose estimator allowed us to extract movement trajectories from videos of infants moving.

We compared the movement of each infant from the clinical cohort with the movement extracted from the reference-group sample based on 38 features that described posture, velocity, acceleration, left-right symmetry, and complexity (see methods, all pre-registered). Inspection of kinematic features from different risk populations reveals some subtle deviations from the reference sample based on neuromotor risk (Figure 5). While each of these features may not be strongly indicative of an infant's neuromotor risk when considered individually, an estimate which pools across features is likely to provide more robust predictions of risk.

Our approach to assessing neuromotor risk allowed us to combine many features into one estimate. We computed the Bayesian Surprise for each individual infant. When features are combined into one estimate of risk, the normalized Bayesian Surprise increases with clinician-assessed risk; Low Risk: mean z $= -1.62$, SD $= 1.18$; Moderate Risk: mean z $= -1.68$, SD $= 1.34$; and High Risk: mean z $= -2.94$, SD$=1.43$ (Figure 6). The proportion of infants classed as at-risk varies
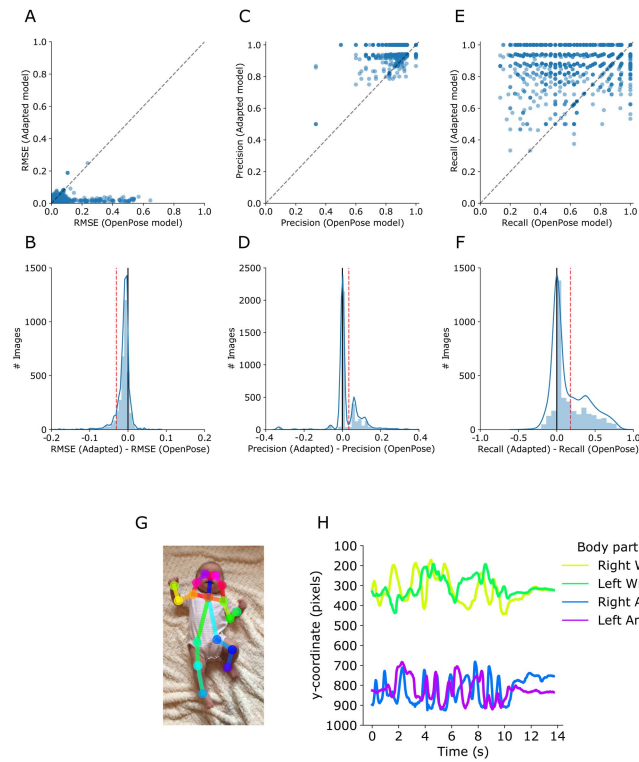
Fig. 4. Pose-estimation model performance and example outputs. (A) Scatterplot of the RMSE of the adapted pose-estimation model as a function of the RMSE of the OpenPose model before domain adaptation. Points show the RMSE for individual images in bounding-box units. In A, C and E, the dotted line shows the diagonal, where performance before and after domain adaptation are equal. (B) Distribution of the difference in model error before and after domain adaptation. For each model, a single RMSE score was computed from errors between individual landmarks and labels averaged across the whole test dataset. RMSE (Adapted) – RMSE (Original) is shown by the red dotted line. Solid black line shows a difference of 0, in B, D and F. The negative RMSE difference demonstrates improvement after domain adaptation. (C) Scatterplot of the precision of the adapted pose-estimation model as a function of the precision before domain adaptation. Points show the precision for individual images. (D) Distribution of the difference in precision before and after domain adaptation. Precision (Adapted) – Precision (Original) is shown by the red dotted line. (E) Scatterplot of the recall of the adapted pose-estimation model as a function of the recall before domain adaptation. Points show the recall for individual images. (F) Distribution of the difference in recall before and after domain adaptation. Recall (Adapted) – Recall (Original) is shown by the red dotted line. (G) Example of OpenPose outputs extracted from a video of an infant using our adapted pose-estimation system. (H) Image y-coordinates of the extremities for the same infant as in (G).
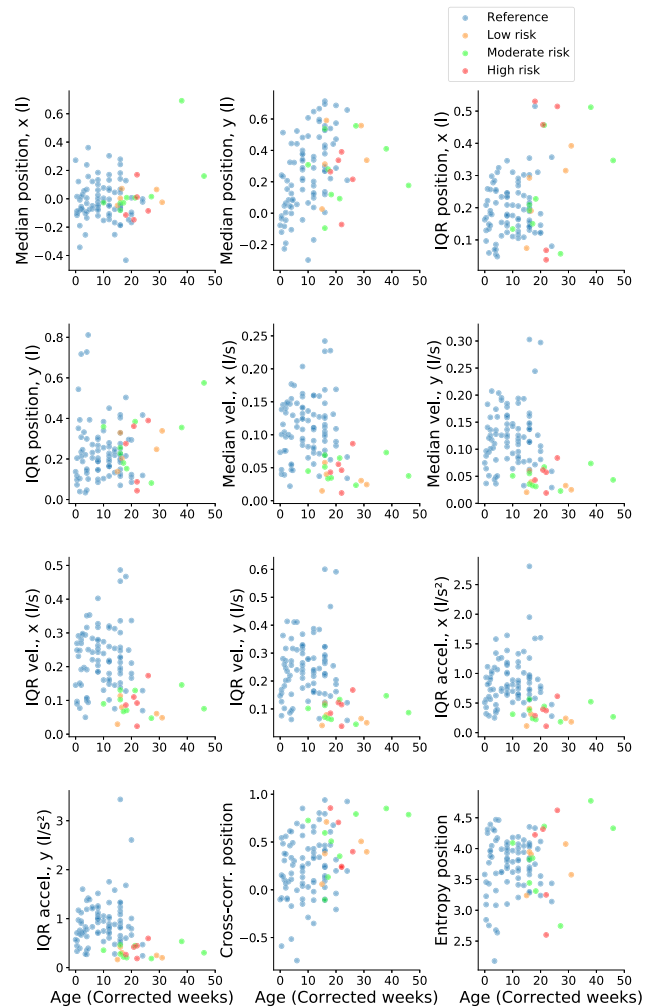

Fig. 5. Infant movement features. Kinematic features of the reference sample (blue), low-risk infants (orange), moderate-risk infants (green), high-risk infants (red) as a function of age in corrected weeks. Features are shown for the wrists: median absolute position (l), IQR of position (l), median velocity (l/s), IQR of velocity (l/s), IQR of acceleration (l/s²◁, left-right cross-correlation of position and entropy of position. Visualization of other features are provided as Supporting Information at the figshare link.

with participant group (Reference: P(Risk) = 0.07, Low Risk: P(risk) = 0.40, Moderate Risk: P(Risk) = 0.33, High Risk: P(Risk) = 0.8). A Kruskal-Wallis test showed a significant association between participant group and the Bayesian Surprise score ($\chi^2 (3) = 29.92$, p < 0.0001). We found significant differences between the reference population and the clinical risk groups using the Mann Whitney U test (Reference, Low Risk: U = 49, p < 0.05; Reference, Moderate Risk: U = 108, p < 0.005; Reference, High Risk: U = 9, p < 0.005, Bonferroni-corrected p-values).

Diferences among clinical risk groups were non-significant (Low Risk, Moderate Risk: U = 20, p=0.99; Moderate Risk, High Risk: U = 6, p=0.63; Low Risk, High Risk: U = 12, p=0.54, Bonferroni-corrected p-values). These results show differences between at-risk infants and our reference population based on a set of simple kinematic features, suggesting that it may be possible to predict clinician's assessments of infantile neuromotor risk through statistical comparison relative to a healthy reference population.

We explored how combinations of our set of 38 movement features relate to neuromotor risk. SVD finds linear combinations of movement features, or latent variables, that summarize the movement feature data using matrix factorization. In our data, the three most important latent variables account for 37%, 15%, and 12% of the variance respectively (Figure 7). We did not include other variables, since they accounted for small proportions of variance (<10%). The first latent variable mainly discriminates between the reference group and infants recorded at the lab, with at-risk infants being more extreme along this variable (Figure 7A). Well-represented movement features include median and variability of velocity, and
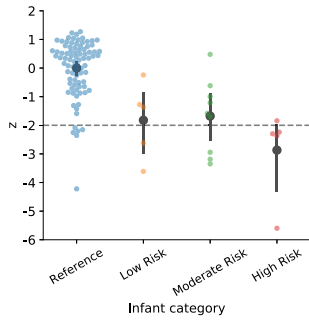
Fig. 6.   Normalized Bayesian Surprise as a function of subject group. The normalized Bayesian Surprise (z) is shown for the reference infant population, and at-risk infants recorded at the lab evaluated by clinicians using the BINS score (Low Risk, Moderate Risk, and High Risk). More negative scores indicate a smaller probability of belonging to the reference population, or higher risk. Points show individual data for each subject group. Individual data is overlaid with the mean for each group (error bars = 95% confidence intervals (CI)).

variability of acceleration (Figure 7B). The second latent variable discriminates between reference, low-medium risk infants and high-risk infants, with high-risk infants having more extreme values along this variable (Figure 7C). Variability of position, positional entropy, symmetry (cross-correlation), and postural variables from the lower body (mean knee angle, median ankle position) are well-represented (Figure 7D). Therefore, in our data, high-risk infants are atypical for this combination of features. The third latent variable does not show a clear pattern relative to the infant group (Figure 7E), but may relate to movement of the upper body relative to the lower body, as most positive weights describe lower body movement and most negative weights describe upper body movement (Figure 7F). Thus, our exploratory SVD analysis suggests that at-risk infants show different distributions of velocity and acceleration compared to the reference population. Also, high-risk infants in particular may have more extreme posture, variability of posture, and symmetry. These results show that at-risk infants in our sample have different movement patterns to the normal population, which can be described in terms of combinations of individual movement features. This analysis allows us to describe important sources of variance in the movement feature data and to explore the importance of combinations of movement features in determining risk assessments.

IV. DISCUSSION

We have developed a framework to identify infants at risk of neuromotor disorder based on video data. Our approach was to compare infants assessed in the laboratory at different levels of neuromotor risk with a normative dataset of infant movement extracted from online videos. We successfully adapted an existing pose-estimation system to perform motion capture on 2-D videos. Using pose estimation, we extracted kinematic features from videos. We found that the naïve Bayesian Surprise measure of risk, which pooled across features, varies across participant groups. We examined the main latent variables that describe the data, with the finding that velocity and acceleration distinguish infants from the reference
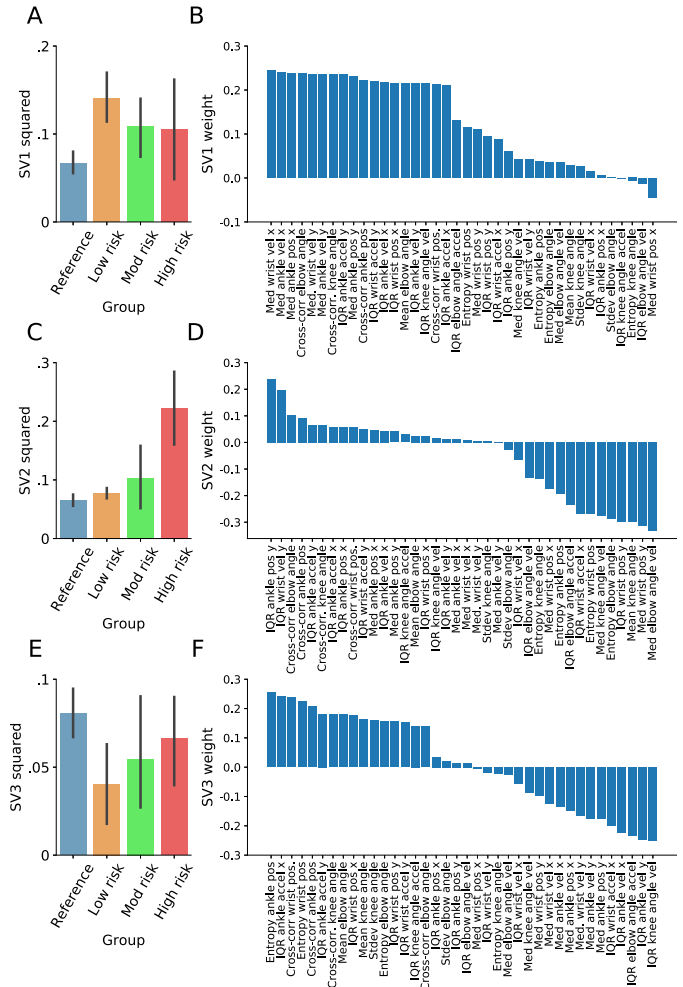

Fig. 7.   SVD analysis of movement feature data in terms of the three most important latent variables. (A), (C), and (E) show values from singular vectors, which describe the infants in terms of latent variables. (B), (D), and (F) show the weighting of movement features in each latent variable. (A) Mean eccentricity (error bars = 95% CI) along the first singular vector (SV1 squared), as a function of participant group. (B) Weighting of movement features in SV1 ranked in descending order. (C) Mean SV2 squared (error bars = 95% CI) as a function of neuromotor risk. (D) The weighting of movement features in SV2 ranked in descending order. (E) Mean SV3 squared (error bars = 95% CI) as a function of neuromotor risk. (F) Weighting of movement features in SV3 ranked in descending order.

population with at-risk infants, and that high-risk infants differ from reference in their posture, postural variability, and symmetry. Combinations of movement features are predictive of neuromotor risk.

Previous work has addressed kinematic markers of disordered infantile movement, with the finding that features of velocity and acceleration including skewness of velocity, maximum acceleration, and minimum speed predict infantile movement disorders [14], [15]. Other work has shown that infants with disordered movement show greater stereotypy [17], [18]. Our SVD analysis showed that at-risk infants may have different velocity and acceleration distributions relative to a reference sample of healthy infants, and that high-risk infants in particular show different patterns of posture, postural variability, symmetry of movement, and complexity. Our findings on the features that predict risk assessments show some

similarity with previous observations with roles for velocity and acceleration, as well as complexity and symmetry. In order to confirm our findings, it will be important to examine the kinematic markers of disordered movement using large clinical datasets.

Previous work on infant movement used measurements capable of extracting fine movement information, like accelerometers, motion capture using depth cameras, and optic flow from videos [14], [15], [22]. This allowed examination of the jerkiness of movement, previously found to be predictive of movement disorder [14], [15]. Pose estimation applied to low-quality videos currently does not afford the high precision needed to compute higher derivatives of movement trajectories. Here, because of the presence of outliers and noise in our pose-estimate data, we applied smoothing to the data which removed its fine detail. In future work, improvements in pose estimation and high-quality video datasets will make it possible to extract fine movement detail from pose estimates.

Another concern about our work here is the small number of subjects in the clinical group. Due to a range of constraints we were only able to analyze data from 19 subjects. We do find a significant difference of those subjects to the normative population. Features including velocity, acceleration and their variability discriminate between at-risk and reference infants. However, we find a significant difference even when comparing the normative population to the low-risk infants in the clinical cohort. Our SVD-based analysis (Figure 7) provides some insights into why we are seeing these differences. While our analysis was exactly as we had planned and preregistered, we cannot avoid the confound that other differences between the YouTube and clinical groups gave rise to the difference we observed. Also, while we do not find significant differences among the three clinical groups (low, moderate, and high risk), this could be due to the small number of infants in each individual risk category. Future work will need to analyze larger datasets, weigh the most predictive features, and correct for sources of variability which do not directly correspond to neuromotor risk. As such, the main contribution of our study is the provision of a pipeline and a set of computational tools that show promise on the small number of infants assessed here. Our pipeline should be used in future studies with larger populations, longitudinal follow-up, and a variety of neuromotor impairments to ascertain clinical efficacy.

In this work, we compare the movements of a reference YouTube infant dataset to that of infant movements observed in a clinical setting. As seen in Figure 5, the age ranges of these two populations are not exactly overlapping. To explore any age effect in our study, we recalculated the Bayesian Surprise scores so that each infant is only compared to infants who are within 10 weeks of age difference from itself. This resulted in 5 infants (2 low risk, 3 moderate risk) from the clinical population being omitted in this post-hoc calculation, not having any reference YouTube infants within a 10 week age range available for comparison. With this age-constrained comparison, we found significant difference between the reference population and the moderate and high risk clinical groups, but not with the low risk group using the Mann Whitney U test (Reference, Low Risk: U = 30,
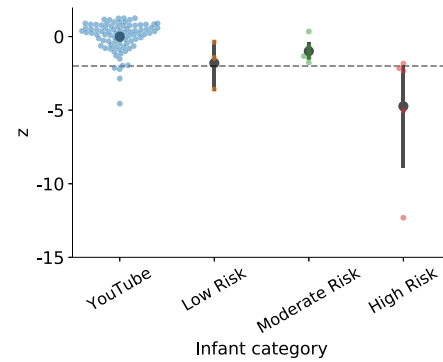

Fig. 8. Age-constrained Normalized Bayesian Surprise as a function of subject group. The normalized Bayesian Surprise (z) is shown for the reference infant population, and at-risk infants recorded at the lab evaluated by clinicians using the BINS score (Low Risk, Moderate Risk, and High Risk). In contrast to Figure 6, the comparison here for each infant is done only to infants who are within 10 weeks in age to explore any age effects on the results. More negative scores indicate a smaller probability of belonging to the reference population, or higher risk. Points show individual data for each subject group. Individual data is overlaid with the mean for each group (error bars = 95% confidence intervals (CI)).

$p = 0.07$; Reference, Moderate Risk: $U = 92$, $p < 0.05$; Reference, High Risk: $U = 11$, $p < 0.005$, Bonferroni-corrected p values). Thus, the moderate and high risk groups differed significantly, but the low risk group did not differ significantly from the other groups. The slight changes in results with this age-constrained recalculation of the Bayesian Surprise should be interpreted carefully as they may be dominated by the omission of 5 infants from an already small clinical population.

In this work, we provide a domain-adapted version of OpenPose, retrained on ground truth labels annotated manually on infant videos. As a few frames from each video are included in our training data, this work establishes within-video generalization: this drastically decreases the amount of manual annotation needed per video.

The pipeline we provide here uses 2-D approximation of 3-D movement from videos as a way to detect infant neuromotor risk. This approximation provides a useful proxy for diagnostic purposes for comparing an impaired population to a healthy one using easily available videos. Similar use of 2-D approximation of 3-D biological data is used in other clinical diagnoses such as X-ray imaging. However, before clinical interpretation of the 2-D movements into 3-D rehabilitation measures, there is a need for validation of pose estimation against 3-D sensor-based measurements. Past work has compared OpenPose estimates on adult movements to motion capture and reported mean absolute errors of less than 20 mm in ~50% data and less than 30 mm in 80% of data [47]. In the future, other studies could compare the pose estimates obtained from our pipeline on infant movements to 3-D motion capture to provide a better basis for clinical interpretability of the features obtained.

The method developed here depends on outputs from a computer vision algorithm. As such it depends strongly on the visibility of the infant. As well as limited ability to measure high frequency movements, the success of pose estimation can be limited by other factors: occlusions, orientation of the
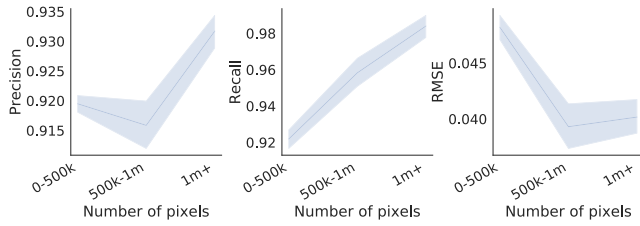
Fig. 9. Visualizing performance of the pose estimation model as a function of video resolution. In our data, we find that the performance of the domain-adapted pose estimation improves with video resolution. In this figure, we show how the RMSE, recall, and precision of the detected poses improve as a function of number of pixels in the original videos (video resolution). The figure above shows the mean and the 95% confidence interval of the mean.

infant, orientation of the camera, and camera movements [44]. To explore how our domain-adapted OpenPose model performs as a function of video quality, we visualized the RMSE, recall, and precision as a function of video resolution. As seen in Figure 9, and as is to be expected, the quality of pose estimation improves with video quality. Although the state-of-the-art computer vision algorithms can successfully extract body pose, they are less successful when small amounts of occlusion are present [49]. Because of this we did not include videos with occlusion by external objects or where the camera angle introduces occlusion of the body on itself (e.g. profile view). However, it can be argued that such recordings may not be ideal for clinical evaluations of infant movement either and thus equally apply to existing clinical tests. Having a clear view of the infant is a reasonable requirement for a diagnostic tool that depends on computer vision, but it does limit the types of videos that can be used for assessment. We argue that this is inherent to the approach at this time and that computer vision-based pose estimation tools still have a significant role to play in movement-based diagnostics.

In this work, we provided methods to quantify and automate infantile neuromotor risk assessments. This system could benefit from several additions. Firstly, future studies should use larger datasets to explore the optimal weighting between movement features that have the best ability to predict risk. Due to the limited dataset size here, we adhere to our pre-registered set of features without optimally weighting them. Secondly, infants move differently depending on the time of day and their emotional state [34]. For an automated movement test with high enough accuracy to be used, it will be important to take such variables into account. It is also important to take into account developmental changes in infant movement [35]. Development is variable across infants: an infant can be delayed in their motor development without having a neuromotor disorder [36]. Therefore, an algorithm that jointly infers an infant's developmental age and neuromotor risk promises to perform better than one that infers risk alone. Neuromotor states can change with time and future studies can explore time-dependence of neuromotor risk by adding features that capture time-dependence. Moreover, the YouTube videos we analyze here are relatively short and it is unclear how the video length may affect risk assessment. In future studies, longer videos could be used to explore the effect

of video length on risk assessment. The incorporation of several variables will be needed for high-accuracy predictions of neuromotor risk.

Our approach to infantile neuromotor risk assessments has made novel contributions. For our normative database, we collected movement data from close to 100 infants using markerless tracking applied to videos. Use of a normative database for comparison increases the robustness of risk assessments. Secondly, we used unsupervised methods to assess risk and to examine important movement features. The Bayesian Surprise metric compared movement of each at-risk infant to the normative database, weighting each feature by its uncertainty. Our SVD analysis describes the main sources of variance in the movement feature dataset. Therefore, our analysis is unlikely to be overfit to the data.

Markerless tracking of infant movement used here provides a number of advantages over existing clinical methods for neuromotor risk assessment. Firstly, video-based diagnostics will improve access to evaluations. A parent need only provide a simple smartphone video of their infant's movement to receive an assessment. Secondly, behavior observed in the laboratory and clinical setting may differ from the real-life setting [37], [38]. Movement variables collected in natural settings would provide a more ecologically-valid assessment, thus providing more meaningful predictions. Markerless tracking also has the potential for quantitative understanding of infant pathology. One of the most common movement disorders is cerebral palsy, a lifelong condition due to brain injury in infancy. Although subtypes and biomarkers have been described, it is poorly understood in terms of its causal determinants [39]. Markerless tracking may serve as a tool to provide a detailed description of movement pathology across a large population, providing better quantitative descriptions of disease subtypes [40]–[42]. Based on large-scale quantitative measurements of movement in different patient groups and in typical populations, one could model how motor-control processes differ between healthy and disordered infants [43]. Therefore, markerless tracking promises to improve accessibility to diagnostics, monitor naturalistic movements, and provide a quantitative understanding of infant neuromotor disorders.

We have developed a method to identify infantile risk of neuromotor disorder based on pose estimates extracted from 2-D videos. Such a method meets the requirements of objectiveness, as movements are assessed based on quantitative variables; and availability, since diagnoses would no longer require the opinion a trained specialist.

Our approach proposed here will improve as larger normative datasets are collected, and as pose-estimation algorithms better suited to movement science are developed [44]. We expect that, over the next decade, movement-based diagnostics from pose estimates will become a viable alternative to established tests such as the General Movements Assessment [45], [46].

V. USER GUIDELINES

We provide some guidelines for best use of the pipeline, code and data we provide along with this article. Our goal

is that the methods provided in this article should work on videos taken on a mobile phone. These guidelines are based on the results included in the paper and also on the experience of the authors from using this tool:

   i To ensure high quality of pose estimates, the users should use as high quality (frame rate and resolution) of video as possible.

   ii To ascertain across-video generalization of pose estimates, maintain similarity across videos in terms of quality, background, and infant clothing. Avoid clothing that occludes the hands and legs (e.g. onesies).

   iii If comparing movements across videos, ensuring the camera angles are similar will result in a more appropriate comparison.

   iv Avoid videos with excessive external occlusions. The algorithms is less likely to detect pose in the presence of occlusions.

   v The domain-adapted OpenPose model provided here has been retrained for supine infant movements. Some retraining may be needed for movements in other positions that have more self-occlusion (e.g. crawling, rolling, etc.). We provide steps in our code for how retraining can be performed.

   vi Avoid videos with excessive camera movements to get the best possible results. Our preprocessing measures here deal with some small camera movements as detailed in the methods section.

VI. Supplemental Information

Code and data referenced in the manuscript are provided at https://github.com/cchamber/Infant_movement_assessment/ and https://doi.org/10.6084/m9.figshare.8161430

Acknowledgment

References

[1] D. Rydz, M. I. Shevell, A. Majnemer, and M. Oskoui, "Developmental screening," *J. Child Neurol.*, vol. 20, no. 1, pp. 4–21, 2005.

[2] W. O'Hare, *The Changing Child Population of the United States: Analysis of Data from the 2010 Census*. Baltimore, MD, USA: Annie E. Casey Foundation, 2011.

[3] I. Novak *et al.*, "Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment," *JAMA Pediatr.*, vol. 171, no. 9, pp. 897–907, 2017.

[4] A. Herskind, G. Greisen, and J. B. Nielsen, "Early identification and intervention in cerebral palsy," *Develop. Med. Child Neurol.*, vol. 57, no. 1, pp. 29–36, Jan. 2015.

[5] *World Report on Disability*, World Bank, World Health Org., Geneva, Switzerland, 2011.

[6] M. Gladstone, "A review of the incidence and prevalence, types and aetiology of childhood cerebral palsy in resource-poor settings," *Ann. Tropical Paediatrics*, vol. 30, no. 3, pp. 181–196, Sep. 2010.

[7] D. M. Romeo, D. Ricci, C. Brogna, and E. Mercuri, "Use of the Hammersmith Infant Neurological Examination in infants with cerebral palsy: A critical review of the literature," *Develop. Med. Child Neurol.*, vol. 58, no. 3, pp. 240–245, Mar. 2016.

[8] 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," *Develop. Med. Child Neurol.*, vol. 50, no. 4, pp. 254–266, Apr. 2008.

[9] Y. Noble and R. Boyd, "Neonatal assessments for the preterm infant up to 4 months corrected age: A systematic review," *Develop. Med. Child Neurol.*, vol. 54, no. 2, pp. 129–139, Feb. 2012.

[10] M. Bosanquet, L. Copeland, R. Ware, and R. Boyd, "A systematic review of tests to predict cerebral palsy in young children," *Develop. Med. Child Neurol.*, vol. 55, no. 5, pp. 418–426, May 2013.

[11] C. Marcroft, A. Khan, N. D. Embleton, M. Trenell, and T. Plötz, "Movement recognition technology as a method of assessing spontaneous general movements in high risk infants," *Frontiers Neurol.*, vol. 5, pp. 1–9, Jan. 2015.

[12] N. Kanemaru *et al.*, "Specific characteristics of spontaneous movements in preterm infants at term age are associated with developmental delays at age 3 years," *Developmental Med. Child Neurol.*, vol. 55, no. 8, pp. 713–721, 2013.

[13] N. Kanemaru *et al.*, "Jerky spontaneous movements at term age in preterm infants who later developed cerebral palsy," *Early Hum. Develop.*, vol. 90, no. 8, pp. 387–392, Aug. 2014.

[14] L. Meinecke, N. Breitbach-Faller, C. Bartz, R. Damen, G. Rau, and C. Disselhorst-Klug, "Movement analysis in the early detection of newborns at risk for developing spasticity due to infantile cerebral palsy," *Hum. Movement Sci.*, vol. 25, no. 2, pp. 125–144, Apr. 2006.

[15] 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.*, vol. 48, no. 8, pp. 765–772, Aug. 2010.

[16] A. Berg, "Modellbasert klassifisering av spedbarns bevegelser (Model-based classification of infant movement)," M.S. thesis, Dept. Tech. Cybern., Norges teknisk-naturvitenskapelige Univ., Trondheim, Norway, 2008.

[17] H. Philippi *et al.*, "Computer-based analysis of general movements reveals stereotypies predicting cerebral palsy," *Develop. Med. Child Neurol.*, vol. 56, no. 10, pp. 960–967, Oct. 2014.

[18] D. Karch *et al.*, "Kinematic assessment of stereotypy in spontaneous movements in infants," *Gait Posture*, vol. 36, no. 2, pp. 307–311, Jun. 2012.

[19] A. Stahl, C. Schellewald, O. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerod, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 4, pp. 605–614, Jul. 2012.

[20] L. Adde, J. Helbostad, A. R. Jensenius, M. Langaas, and R. Støen, "Identification of fidgety movements and prediction of CP by the use of computer-based video analysis is more accurate when based on two video recordings," *Physiotherapy Theory Pract.*, vol. 29, no. 6, pp. 469–475, Aug. 2013.

[21] R. Støen *et al.*, "Computer-based video analysis identifies infants with absence of fidgety movements," *Pediatric Res.*, vol. 82, no. 4, pp. 665–670, Oct. 2017.

[22] L. Adde, J. L. Helbostad, A. R. Jensenius, G. Taraldsen, K. H. Grunewaldt, and R. Støen, "Early prediction of cerebral palsy by computer-based video analysis of general movements: A feasibility study," *Develop. Med. Child Neurol.*, vol. 52, no. 8, pp. 773–778, Feb. 2010.

[23] M. D. Olsen, A. Herskind, J. Bo Nielsen, and R. R. Paulsen, "Model-based motion tracking of infants," in *Computer Vision* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8927. 2015, pp. 673–685.

[24] N. Hesse *et al.*, "Learning an infant body model from RGB-D data for accurate full body motion analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*, 2018, pp. 792–800.

[25] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–9.

[26] L. Lonini, N. Shawen, K. Scanlan, W. Z. Rymer, K. P. Kording, and A. Jayaraman, "Accelerometry-enabled measurement of walking performance with a robotic exoskeleton: A pilot study," *J. NeuroEng. Rehabil.*, vol. 13, no. 1, pp. 1–10, Dec. 2016.

[27] C. Chambers, G. Kong, K. Wei, and K. Kording, "Pose estimates from online videos show that side-by-side walkers synchronize movement under naturalistic conditions," *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0217861.

[28] S. S. Shivakumar *et al.*, "Stereo 3D tracking of infants in natural play conditions," in *Proc. Int. Conf. Rehabil. Robot. (ICORR)*, Jul. 2017, pp. 841–846.

[29] K. Gücüyener *et al.*, "Use of the bayley infant neurodevelopmental screener with premature infants," *Brain Develop.*, vol. 28, no. 2, pp. 104–108, Mar. 2006.

[30] G. P. Aylward, *The Bayley Infant Neurodevelopmental Screener*. San Antonio, TX, USA: Psychological Corporation, 1995.

[31] G. P. Aylward and S. J. Verhulst, "Predictive utility of the bayley infant neurodevelopmental screener (BINS) risk status classifications: Clinical interpretation and application," *Develop. Med. Child Neurol.*, vol. 42, no. 1, pp. 25–31, Feb. 2007.

[32] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vis.*, vol. 101, no. 1, pp. 184–204, Jan. 2013.

[33] T. Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[34] E. Oyerinde *et al.*, "A novel approach to parameterizing infant emotion and movement in relation to cognitive & motor development," in *Proc. Rehabil. Eng. Assistive Technol. Soc. Korea*, 2018, p. 106.

[35] J. Law, M. Lee, M. Hülse, and A. Tomassetti, "The infant development timeline and its application to robot shaping," *Adapt. Behav.*, vol. 19, no. 5, pp. 335–358, Oct. 2011.

[36] W. M. G. R. S. Group, "WHO Motor Development Study: Windows of achievement for six gross motor development milestones," *Acta Paediatr.*, vol. 450, pp. 86–95, Apr. 2006.

[37] U. Bronfenbrenner, "Toward an experimental ecology of human development.," *Amer. Psychologist*, vol. 32, no. 7, pp. 513–531, 1977.

[38] J. N. Ingram and D. M. Wolpert, "Naturalistic approaches to sensorimotor control," *Prog. Brain Res.*, vol. 191, pp. 3–29, Jan. 2011.

[39] T. D. Sanger, "Is cerebral palsy a wastebasket diagnosis?" *J. Child Neurol.*, vol. 23, no. 7, pp. 8–726, 2008.

[40] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, and M. Bax, "A report: The definition and classification of cerebral palsy April 2006," *Dev. Med. Child Neurol.*, vol. 49, no. 109, pp. 8–14, 2007.

[41] T. D. Sanger *et al.*, "Definition and classification of negative motor signs in childhood," *Pediatrics*, vol. 118, no. 5, pp. 2159–2167, Nov. 2006.

[42] R. T. Katz and W. Z. Rymer, "Spastic hypertonia: Mechanisms and measurement," *Archi. Phys. Med. Rehabil.*, vol. 70, no. 2, pp. 144–155, 1989.

[43] S. H. Scott and K. E. Norman, "Computational approaches to motor control and their potential role for interpreting motor dysfunction," *Current Opinion Neurol.*, vol. 16, no. 6, pp. 693–698, Dec. 2003.

[44] N. Seethapathi, S. Wang, R. Saluja, G. Blohm, and K. P. Kording, "Movement science needs different pose tracking algorithms," 2019, *arXiv:1907.10226*. [Online]. Available: http://arxiv.org/abs/1907.10226

[45] H. F. R. Prechtl, "General movement assessment as a method of developmental neurology: New paradigms and their consequences," *Develop. Med. Child Neurol.*, vol. 43, no. 12, pp. 836–842, Feb. 2007.

[46] K. Wei and K. P. Kording, "Behavioral tracking gets real," *Nature Neurosci.*, vol. 21, no. 9, pp. 1146–1147, Sep. 2018.

[47] N. Nakano *et al.*, "Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras," *bioRxiv:842492*. Jan. 2019, Art. no. 842492.

[48] Z. Ghahramani, "Information theory," in *Encyclopedia of Cognitive Science*. Hoboken, NJ, USA: Wiley, 2006.

[49] I. Sárándi, T. Linder, K. O. Arras, and B. Leibe, "How robust is 3D human pose estimation to occlusion?" 2018, *arXiv:1808.09316*. [Online]. Available: http://arxiv.org/abs/1808.09316