sensors

MDPI

Review

Video-Based Automatic Baby Motion Analysis for Early Neurological Disorder Diagnosis: State of the Art and Future Directions

Marco Leo [1,*], Giuseppe Massimo Bernava [2], Pierluigi Carcagnì [1] and Cosimo Distante [1]

[1] Institute of Applied Sciences and Intelligent Systems (ISASI), National Research Council of Italy, Via Monteroni Snc, 73100 Lecce, Italy; pierluigi.carcagni@cnr.it (P.C.); cosimo.distante@cnr.it (C.D.)
[2] Institute for Chemical-Physical Processes (IPCF), National Research Council of Italy, Viale Ferdinando Stagno d'Alcontres 37, 98158 Messina, Italy; giuseppe.bernava@cnr.it
* Correspondence: marco.leo@cnr.it

Abstract: Neurodevelopmental disorders (NDD) are impairments of the growth and development of the brain and/or central nervous system. In the light of clinical findings on early diagnosis of NDD and prompted by recent advances in hardware and software technologies, several researchers tried to introduce automatic systems to analyse the baby's movement, even in cribs. Traditional technologies for automatic baby motion analysis leverage contact sensors. Alternatively, remotely acquired video data (e.g., RGB or depth) can be used, with or without active/passive markers positioned on the body. Markerless approaches are easier to set up and maintain (without any human intervention) and they work well on non-collaborative users, making them the most suitable technologies for clinical applications involving children. On the other hand, they require complex computational strategies for extracting knowledge from data, and then, they strongly depend on advances in computer vision and machine learning, which are among the most expanding areas of research. As a consequence, also markerless video-based analysis of movements in children for NDD has been rapidly expanding but, to the best of our knowledge, there is not yet a survey paper providing a broad overview of how recent scientific developments impacted it. This paper tries to fill this gap and it lists specifically designed data acquisition tools and publicly available datasets as well. Besides, it gives a glimpse of the most promising techniques in computer vision, machine learning and pattern recognition which could be profitably exploited for children motion analysis in videos.

Keywords: baby motion analysis; early diagnosis; neurodevelopmental disorders; machine learning; deep learning

check for updates

Citation: Leo, M.; Bernava, G.M.; Carcagnì , P.; Distante, C. Video-Based Automatic Baby Motion Analysis for Early Neurological Disorder Diagnosis: State of the Art and Future Directions. *Sensors* **2022**, 22, 866. https://doi.org/10.3390/s22030866

Academic Editor: Raffaele Gravina

Received: 12 December 2021
Accepted: 20 January 2022
Published: 24 January 2022

1. Introduction

Neurodevelopmental disorders (NDD) are impairments of the growth and development of the brain and/or central nervous system. They encompass several conditions, including intellectual developmental disorders, communication disorders, Autism Spectrum Disorder (ASD), Attention Deficit Hyperactivity Disorder (ADHD), specific learning disorders, and motor disorders. The prevalence has been increasing during the last two decades and it has been recently proven that preterm infants have an increased risk of neurodevelopmental disorders [1]. The human brain takes about forty years to reach its full-blown adult configuration, and then, assessments need to be age-specific, that is, the assessment techniques and assessment criteria should be adapted to the age-specific properties of the infant brain [2]. Anyway, there is a growing appreciation that the origins of these disorders are at the earliest stages of brain development, even prenatally [3]. The early origin of the neurodevelopmental disorders would potentially allow their early detection, and hence, an early onset of intervention, that is, intervention in a time window characterized by high neural plasticity. Gold standard methods exist in clinical practice for early diagnosis of NDD and they have been very well described in a recent survey paper [2]. Concerning motor assessments, three methods are commonly used to predict outcome: general movement assessment (GMA) [4], the Test of Infant Motor Performance (TIMP) [5] and the Infant Motor Profile (IMP) [6]. The GMA method provides an assessment of the spontaneous movements, called General movements (GMs), that are present from early fetal life onwards until the end of the first half a year of life. GMs are complex, occur frequently, and last long enough to be observed properly but, if the nervous system is impaired, GMs lose their complex and variable character and become monotonous and poor. Normally, they are the predominant movement patterns in an awake infant at 3 to 5 months [7]. IMP method evaluates motor abilities, movement variability, ability to select motor strategies, movement symmetry, and fluency. The IMP consists of 80 items and is applicable to children from 3 to 18 months. The TIMP method assesses the posture and selective control of movement needed by infants under four months of age for functional performance in daily life. Many studies demonstrated that cerebral palsy (CP) or autistic (ASD) children show disturbances of movement that could be detected clearly at the age of 4–6 months, and sometimes even at birth [8]. Attention Deficit Hyperactivity Disorder (ADHD) was found to be predicted in the first years of life by delays in gross motor milestones, abnormal GMs and less motor maturity [9–11]. Nearly half of the pediatric chronic pain patients suffer from comorbid mental health disorders, including mood and anxiety disorders, autism and ADHD. As a consequence, also monitoring pain in infants can help to predict NDD [12]. An overview of early behavioural markers for neurodevelopmental disorders in the first three years of life can be found in [13].

In the light of the aforementioned clinical findings and prompted by recent advances in hardware and software technologies, several researchers tried to introduce automatic systems (which should overcome differences in movements assessments of raters with various levels of experience [14]) to analyse the baby's movements, even in cribs.

Most of the existing technologies for baby's motion analysis are based on contact physical sensors [15]. They include force sensors, accelerometers, gyroscopes, extensometers, inclinometers, goniometers, electromyography [16]. Physical sensors assure high temporal resolution and very high accuracy but their use is discouraged by the sparsity of spatial data, difficulties to get consistent positioning and possible modifications of the behaviours to be observed. Alternatively, active/passive visual markers can be positioned on the children and acquired by optical devices. Their use is discouraged by the difficulties to get consistent positioning and then by long set-up times [17]. In fact, wearable physical sensors may alter baby movement productions. Under this perspective, markerless video-based approaches, which leverage data contents acquired from RGB cameras or Depth devices (e.g., Microsoft Kinect/Intel RealSense, etc.) in an ecological way (i.e., without any additional elements in the scene), become much more attractive for a reliable assessment of movements in children [18]. On the other hand, they require much more computational efforts and this brought researchers to initially concentrate on specific or easily detectable movements (i.e., exhibiting periodicity), for example, for spotting in real-time the occurrences of anomalies and providing prompt warnings to parents or healthcare staff [19]. Typical examples are systems that detect the presence or absence, respectively, of periodic movements of parts of the body—e.g., the limbs in case of clonic seizures and the chest/abdomen in case of apneas [20]. Other examples refer to tools for Neonatal Intensive Care Units (NICU) for detecting discomfort moments [21] and to estimate respiratory rate [22–24]. Most recently, methodological and technological progress in computer vision, machine learning and pattern recognition introduced disrupting improvements in automatic human activity recognition in videos (even relying on data acquired by handheld devices, such as smartphones [25]) and enabled the development of various automated applications in different fields, such as security and surveillance, healthcare, sports, home automation and recommender systems [26,27]. The baby's motion analysis in videos for the early diagnosis of NDD benefited from this scientific fervor as well. Some of the related works dealing with GMA issues have been recently summarized in three systematic searches of papers [28–30] whereas those addressing the specific clinical task of detecting ASD features have been reviewed in [31]. The aforementioned survey papers are very interesting but they gave a task-specific (e.g., GMA or ASD assessment) view of the exploited computer vision and machine learning methods. This makes it difficult to understand how proposed approaches can be transferred to other tasks, involving different disorders and ages. A more global and structured vision of the problem would certainly be desirable to incentive research and its applicative effects but, unfortunately, to the best of our knowledge, the literature lacks a manuscript giving a broader overview of the video-based analysis of children motion for assessment of NDD. This paper fills this gap by providing a survey on advanced computational methods for early NDD diagnosis starting from temporal sequences of 2D/3D data. Besides, available software tools and public datasets are also considered. It gives also a glimpse of the most promising techniques in computer vision, machine learning and pattern recognition which could be exploited for children motion analysis in videos. The rest of the paper is organized as follows: Section 2 introduces a taxonomy to classify existing approaches in the scientific literature for movement assessment in children for early NDD diagnosis. Then, Section 3.2 introduces existing tools for data acquisition, collection and labelling whereas Section 4 discusses approaches for assessment of movements in newborns, infants and toddlers. The subsequent Section 5 provides a glimpse of the very latest methodologies for movement analysis which could be transferred in the considered domain of the early detection of NDD in children. Finally, Section 6 concludes the paper.

2. Taxonomy

Different categorisations of works dealing with the analysis of movements of infants for early NDD diagnosis can be carried out. The most straightforward one is based on the technology used to acquire input video data: RGB cameras, depth sensors or both. Acquisition devices can be handheld such as smartphones or fixed such as surveillance cameras. Another categorisation option relies on the acquisition conditions. Some works consider setups in which the children are in a hospital or an NICU. In these cases, children are in cribs, generally in the supine position, and then, acquisition devices and algorithms are designed to work properly according to this. Self-occlusions and other issues should be addressed to properly acquire the whole-body shape (for example, due to wearing diapers or sensors for monitoring vital parameters). Sometimes additional constraints are required to improve their robustness such as bedsheets of uniform colour and so on. Other works deal with children monitoring in more unconstrained environments such as homes and treatment centres. Acquisitions were carried out while adults are close to the child who can interact with the objects in the scene (e.g., puppets, tablets, etc.). Occlusions often happen and also lighting conditions change, sometimes even during the same acquisition session. Children can assume any possible posture (seating, standing, lying, etc.).

Alternative taxonomies can be pivoted either on the computer-vision/machine learning tasks actually addressed (data collection and labelling, static pose estimation, spatiotemporal modelling of motion, action recognition) or on the high-level clinical tasks pursued (ASD, CP or ADHD detection, pain quantification, monitoring rehabilitation/training sessions, etc.). All the aforementioned categorisations are certainly valid but, since the assessment techniques and criteria have to be adapted to the age-specific properties of the infant's brain, in this paper, the categorisation will be carried out considering the age of involved children. It is the main element that pilots the architectural and algorithmic choices: acquisition setups depend on the acquired ability to walk; motion patterns are related to the age (according to previously mentioned assessment theories) and as consequence algorithms have to capture age-specific motion features. In the light of the above, in this paper, three different age ranges are considered to categorise related works in the literature: newborns (up to 2 months old), infants (2 months to 1-year-old) and toddlers (from 1 to 4 years old).

Figure 1 summarizes the possible categorisation options (the chosen one is circled in red).

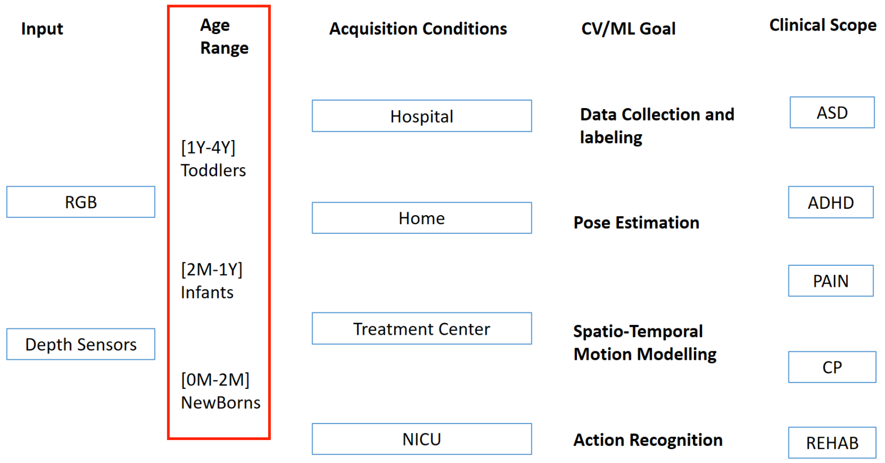

Figure 1. Introduced taxonomy.

3. Data Acquisition, Collection and Labelling

The first steps towards the design of any video-based framework, for automatic motion analysis, are the setup of the acquisition devices, the secure collection and the privacy-preserving storage of data and, eventually, their annotation by experts in order to feed supervised machine learning algorithms. Under these premises, in the following subsections, existing acquisition tools and the most relevant annotated benchmark datasets specifically designed for baby motion analysis are described.

3.1. Acquisition/Recording Tools

It is quite difficult to find tools properly designed and set up for video-based analysis of children's motor performance. In particular, two systems are described in the following. In chronological order, the first one is the AVIM system [32], a monitoring system developed in C# language using the OpenCV image processing library and specifically designed for an objective analysis of infants from 10 days to the 24th week of age. It acquires and records images and signals from a webcam and a microphone but also allows users to perform both audio and video editing. Very useful functionalities are the possibility of adding notes during the recording and to play/cut/copy and assess on-the-fly the sequences of interest. Besides, it extracts from the image the 2D position of the body segments to help the study of the movements according to amplitude, average speed and acceleration. The body analysis can concentrate either on the lower body, based on three points only, or on the full body by taking into account 8 points (right shoulder, left shoulder, left hand/wrist, base of the sternum, pubis/genitals, tight foot/ankle, left foot/ankle). It is worth noting that in both modalities (lower body or full body), all the points are manually placed and then tracked over time in order to extract some motion parameters according to the clinical literature are automatically extracted [33]. Some acoustic parameters (and related statistics) are automatically estimated as well (fundamental frequency, first two resonance frequencies of the vocal tract, kurtosis, skewness and time duration of each cry unit).

The second device deserving a mention is MOVIDEA [34] which has been designed for semi-automatic video-based analysis of infants' motor performance. It includes a camera placed 50 cm above the child, at chest height, and software designed to extract kinematic features of limbs of a newborn (up to 24 weeks old) at home while lying on a bed, upon a green blanket. A Graphical User Interface completes the system and it allows the software operator to interact with the system. At first, the operator has to identify the limb by selecting the central point of the region of interest (i.e., hand, foot). The system then tracks the selected point frame by frame using the Kanade–Lucas–Tomasi algorithm [35] and movement features of extracted trajectories are compared with the reference ones for the identification of pathological motion patterns [36].

3.2. Publicly Available Datasets

In general, there is a lack of publicly available benchmark databases specifically built for children movement analysis. This is mainly due to privacy and security considerations that pose restrictions on ethics approval making this way very difficult to train robust models from scratch. Anyway, some exceptions exist. For the purpose of this survey, three different types of publicly available databases can be mentioned depending if they are oriented to build up models aiming at:

- Estimating the child pose;
- Comparing normative behaviours to those of monitored children in order to suggest further investigations;
- Recognizing atypical behaviours in order to directly get an NDD diagnosis.

Concerning pose estimation in newborns, the babyPose dataset [37] contains data relevant to preterm children's movement acquired in NICUs. The data consist of 16 depth videos recorded during the actual clinical practice. Each video consists of 1000 frames (i.e., 100 s). The dataset was acquired at the NICU of the Salesi Hospital, Ancona (Italy). Each frame was annotated with the limb-joint location. Twelve joints were annotated, i.e., left and right shoulder, elbow, wrist, hip, knee and ankle.The database is freely accessible at http://doi.org/10.5281/zenodo.3891404 (accessed on 10 December 2021).

A benchmark dataset for a standardized evaluation of systems for pose estimation in infants has been made available for research purposes at http://s.fhg.de/mini-rgbd (accessed on 10 December 2021), namely the Moving INfants In RGB-D (MINI-RGBD dataset) [38] data set. It contains images of infants up to the age of 7 months lying in supine position, facing the camera. It has been created using the Skinned Multi-Infant Linear body model (SMIL) [39], a system able to build realistic infant body sequences (with both RGB and depth images) and to provide also a precise ground truth 2D and 3D joint positions. The dataset was bolstered by a recent study in [40] proving that movement assessment from videos of computed 3D infant body models is equally effective compared to the rating on conventional RGB videos. In particular, the MINI-RGBD dataset consists of 12 sequences of continuous motions (640 × 480 resolution at 25 FPS), each 1000 frames long. The sequences are divided into different levels of difficulty: lying on back, moving arms and legs, mostly besides the body, without crossing, (ii) medium: slight turning, limbs interact and are moved in front of the body, legs cross, and (iii) difficult: turning to sides, grabbing legs, touching the face, directing all limbs towards camera simultaneously.

A hybrid synthetic and real infant pose is the so-called SyRIP dataset [41]. The dataset includes a diverse set of real and synthetic infant images, which benefits from (1) appearance and poses of real infants in images from the web (from YouTube and Google Images), and (2) the augmented variations in viewpoints, poses, backgrounds, and appearances by synthesizing infant avatars. The dataset is available at https://github.com/ostadabbas/Infant-Pose-Estimation (accessed on 10 December 2021).

Concerning 'normative' reference datasets, the MIA dataset (https://vrai.dii.univpm.it/mia-dataset (accessed on 10 December 2021)) consists in the state vector ("in movement" or "not in movement" for each of the 4 limbs), along with the corresponding timestamp, derived from depth measurements collected by an RGB-D sensor placed perpendicularly above the child (a male hospitalized in an NICU) lying in a supine position on the crib, at a distance of 70 cm normally directed to the subject. Unfortunately, no video was provided for privacy reasons but, anyway, the provided state vector could be used for training models to be subsequently tested on data extracted from videos.

Recently, a 'normative' reference database of infant movements has been created using 85 videos found online [42]. Two physical therapists estimated the age of the infants. Estimated mean age was 9.67 weeks and standard deviation 6.26 weeks. Using this normative database, OpenPose tool [43] and a Gaussian estimator, the authors calculated how much 19 high-risk children deviate from the typical movements of healthy infants. Code and data referenced in the manuscript are provided at https://github.com/cchamber/Infant_movement_assessment/ (accessed on 10 December 2021).

The most used publicly available dataset of videos for Autism Diagnosis is named Self-Stimulatory Behaviours (SSBD dataset) [44]. It consists of videos of children exhibiting self-stimulatory ('stimming') behaviours commonly used in autism diagnosis. These videos, posted by parents/caregivers on public domain websites, are collected and annotated for the stimming behaviours. These videos are extremely challenging for automatic behaviour analysis as they are recorded in uncontrolled natural settings. The dataset contains 75 videos with an average duration of 90 s per video, grouped under three categories of stimming behaviours: arm flapping, head banging, and spinning. The dataset, the terms of use and the Matlab script file to generate baseline results (making use of a combination of STIP, HIG/HOF and Bag-of-Words, plus SVM for classification) are available at https://rolandgoecke.net/research/datasets/ssbd/ (accessed on 10 December 2021).

The Multimodal Dyadic Behaviour (MMDB) dataset [45] is a unique collection of multimodal (video, audio, and physiological) recordings of the social and communicative behaviour of toddlers (aged 15–30 months). The dataset contains 160 sessions of 5-minute interaction from 121 children. All multimodal signals are synchronized, including 2 frontal view Basler cameras (1920 × 1080 at 60 FPS), an overhead view Kinect (RGB-D) camera, 8 side view and 3 overhead view AXIS cameras (640 × 480 at 30 FPS), an omnidirectional and a cardioid microphone, 2 wireless lapel microphones, 4 Affectiva Q-sensors for electrodermal activity and accelerometry, worn by both the adult and the child.

Another dataset commonly used to train and test machine learning algorithms to classify ASD-related behaviours is the one introduced by Tariq et al. [46]. This dataset was collected through a mobile web portal built up by some researchers at Stanford University and it contains 116 short home videos of children (age range 2–4 years old) with autism and 46 videos of typically developing children.The de-identified data have been made available at the following GitHub repository and include the primary dataset and the validation dataset: https://github.com/qandeelt/video_phenotyping_autism_plos/tree/master/datasets (accessed on 10 December 2021).

Recently, the DREAM Dataset [47] has been also made publicly available at https://snd.gu.se/sv/catalogue/study/snd1156/1/1# (accessed on 10 December 2021). It consists of behavioural data recorded from 61 toddlers diagnosed with autism spectrum disorder collected during a large-scale evaluation of Robot Enhanced Therapy (RET). The public release of the dataset comprises body motion, head position and orientation, and eye gaze variables, all specified as 3D data in a joint frame of reference. In addition, metadata including participant age, gender, and autism diagnosis (ADOS) variables are included.

Finally, depth videos templates of autistic repetitive behaviours (e.g., hands on the face, hands back, tapping ears, hands stimming, hand moving front of the face, toe walking, walking in circles, etc.) were collected in the dataset 3D-Autism Dataset (3d-AD) [48]. Each action has been repeated at least 10 times with non-autistic people. The depth maps have been captured at a rate of 33 frames per second with a Kinect-v2 camera.

Table 1 shows the main properties of the aforementioned publicly available datasets.

Table 1. Main properties of publicly available datasets. N stands for Newborns, I for Infants and T for Toddlers, Misc for Miscellaneous, NA for Not Applicable, U for Unknown.

Database	Contents	Frame Size	Age Range	Info	Frames	Labels
BabyPose [37]	16 Videos	640 × 480	N	Depth 8 bit/16 bit	16,000	12 Body Landmarks
MINI-RGBD [38]	12 Videos	640 × 480	I	RGB/D	12,000	25 Body Landmarks
SyRIP [41]	Images	Misc	I	RGB	2000	17 Body Landmarks
Dataset [42]	85 Youtube Video URLs	Misc	I	RGB	NA	18 Body Landmarks

Table 1. *Cont.*

Database	Contents	Frame Size	Age Range	Info	Frames	Labels
SSBD [44]	75 Youtube Video URLs	Misc	NA	RGB	U	Behaviors
MMDB [45]	160 Videos	Misc	T	Multimodal	U	ASD Diagnosis
Tariq [46]	162 Videos	Misc	T	RGB	U	Behaviors
DREAM [47]	3121 Videos	NA	T	Depth	NA	3D Skeleton Gaze ADOS scores
3d-AD [48]	100 Videos	512 × 424	T	Depth	U	Behaviors

4. Methods and Systems for Movement Assessment

Searching for the considered topic, 20 works were found in the literature, quite equally distributed in the three age ranges (7 for newborns, 7 for infants and 6 for toddlers). It is worth noting that most of the works dealing with newborns leverage traditional computer vision methods (5 of 7). Concerning movements analysis in infants, this trend persists but only 3 of 7 works did not use deep learning-based approaches. Finally, when involving toddlers, the majority of the works made use of deep learning strategies (only 2 of 6 did not). In Figure 2, a pie chart representing the impact of deep learning in each age range is reported. It is possible to observe that when dealing with newborns, most of the solutions made use of ’traditional approaches’, i.e., based on handcrafted features and/or shallow classifiers. The percentage of deep-learning-based solutions increased when infants were involved and became predominant while handling toddlers. This is not surprising since most of the approaches made use of pre-trained models (on adults), and then, it is straightforward to use them on walking children instead of on children in supine pose (in a bed).

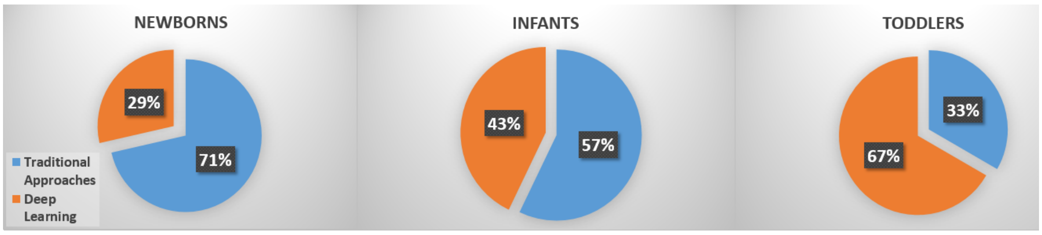

Figure 2. Percentages of published papers with respect to Age Range taxonomy.

Another preliminary consideration could be made about acquisition settings. In Figure 3, the percentages of published papers with respect to setup taxonomy (home, hospital, etc.) is reported. Methods dealing with newborns were mainly applied in hospital/NICU settings with only one work handling just synthetic data. Home settings were partially used for infants and in most of the experiments on toddlers. Once again, this is not surprising since as children grow, it becomes necessary to assess them in environments where their behaviour is not conditioned by the context.

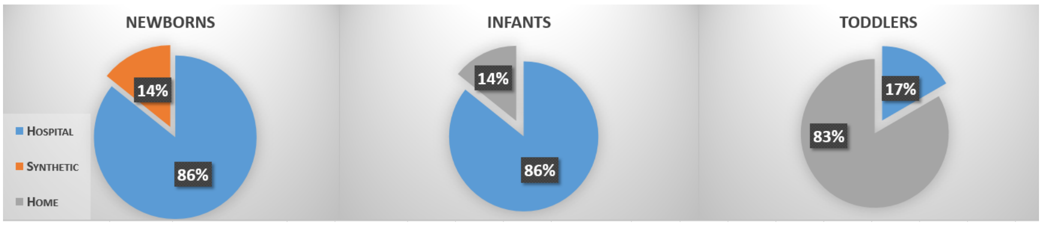

Figure 3. Percentages of published papers with respect to setup taxonomy (home, hospital, etc.).

The following subsections will detail the related works found in the literature.

4.1. Newborns

Newborn usually refers to a baby with an age from pre-term to approximately 9 weeks post-term. According to [49], that is the so-called 'writhing' age. Movements in this writhing age can be categorised as Writhing Movements (WM), Poor Repertoire (PR), Cramped Synchronised (CS) or Chaotic Movements (CM). These categorisations represent various levels of quality to the typical movements during this age period. The presence of persistently cramped synchronized writhing movements, followed by the absence of fidgety movements, is the strongest predictor for poor neurodevelopmental outcomes [4]. In particular, it has been reported that observation of poor repertoire pattern seems to be associated with minor neurological dysfunctions [50] and cramped-synchronized movements is highly predictive of the development of cerebral palsy [51]. Current studies indicate that the GMA is the most sensitive and specific test available to allow early detection of CP [52]. These kinds of assessments are carried out especially in the *NICUs*, and they need to accurately identify those infants most at risk. Early video-based works in this research area used fast recognition of key points to find the 3D positions of body joints in single depth images [53,54]. The tuning of model parameters exploited poses that are not typical for infants and their specific motions, but a series of synthetically generated baby-like poses was also added in order to reduce that bias. Tests were carried out on babies at the age of 3 months who were always filmed from above so that the main body axis is displayed vertically in the camera image.

In [55], a deformable part-based model was exploited to detect the body parts (by skeletonization) of children aged from 2 weeks to 6 months. Then, angles for joints are computed and tracked temporally in a video sequence to describe movements. Although the authors evaluated the accuracy in the estimation of joint positions and movements encoding without a specific clinical application, they asserted that it was specifically designed to evaluate the patient's poses and movement during therapeutic procedures (e.g., Vojta techniques [56]) aimed at early diagnosis of cerebral palsy, spinal scoliosis, peripheral paralysis of arms/legs, hip joint dysplasia and various myopathies.

Authors in [57] introduced a system relying on a multimodal recording setup consisting of two HD cameras, two Kinect and sensors for pressure measurement. Attributes used in audio/video analysis, such as optical flow, zero-crossings rate, harmonics-to-noise ratio and jitter are computed for children in their first 4 months of life and compared with those of children (of the same age) having a typical development and with diagnosed conditions of interest. To this end, logistic regression from multidimensional data was exploited.

The study in [58] attempted to automatically detect writhing movements instead. The study involved newborns on the second and third days of life. Different feature extraction strategies and traditional machine learning algorithms were exploited for writhing movement detection. Based on automatically detected writhing movement percentages in the videos, infants are classified as having a good level of writhing movements or as having a poor repertoire, i.e., a lower quality of movement in relation to the norm.

Even if the use of deep learning techniques in this application context is not easy, due to the lack of annotated data, recently, some solutions based on convolutional neural networks (CNN) have nevertheless been proposed. An approach to preterm infants' limb pose estimation that features spatiotemporal information to detect and track limb joints from depth videos with high reliability was proposed in [59]. The depth camera (model Astra Mini S-Orbbec (https://orbbec3d.com/astra-mini-series/ (accessed on 10 December 2021)), with a frame rate of 30 frames per second and image size of 640 × 480 pixels) is positioned at 40 cm over the infant's crib in order to not hinder health-operator movements.Limb-pose estimation is performed using a deep-learning framework consisting of detection and regression CNNs for rough and precise joint localization, respectively. The CNNs are implemented to encode connectivity in the temporal direction through 3D convolution. Assessment of the framework was performed through a comprehensive study with sixteen depth videos acquired in the actual clinical practice from sixteen preterm infants.

The proposed solutions can be exploited for diagnostic support, e.g., to classify abnormal limb movements.

Finally, authors in [60] proposed five deep-learning-based frameworks to classify infant body movement based upon the pose-based features which consisted of histogram representations describing different aspects of the extracted pose-based features. The final aim was to automatically label observed movements as indicative of typically developing infants (Normal) or that may be of concern to clinicians (Abnormal). Synthetic data from the MINI-RGBD dataset were used and an accuracy of over 90% was achieved by using histograms of joint displacement and orientation as features and 1D convolutional neural network architectures exploited as reported in Figure 4.

Table 2 Summarizes works dealing with movements assessment in newborns.

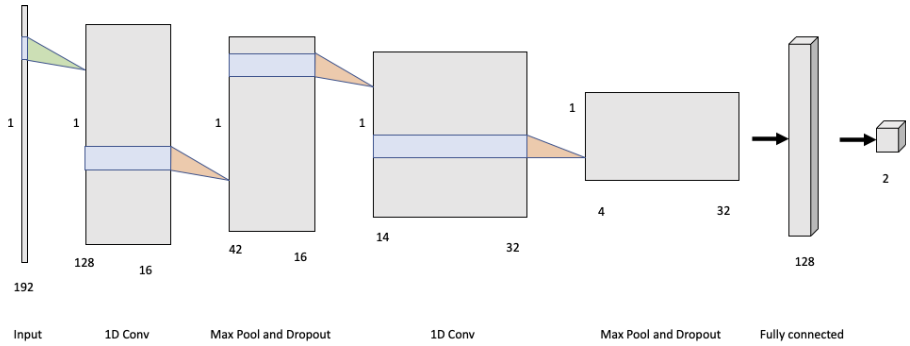

Figure 4. 1D convolutional neural network architecture exploited in [60] for labelling observed movements as indicative of typically developing infants (Normal) or that may be of concern to clinicians (Abnormal).

Table 2. Summarization of works dealing with movements assessment in newborns.

Work	Setup	Input	CV/Ml Task	Clinical Scope
[53,54]	Hospital	Depth	Pose estimation by Keypoints recognition	General
[59]	NICU	Depth	Limb Pose by 2 CNN 2CNN (detection + regression)	General
[55]	Hospital	RGB	Deformable part models	General
[57]	Hospital	Multimodal	Optical Flow + audio features Logistic regression	Normal/Abnormal
[58]	Hospital	RGB	Limb Motion Description by SVM, RF, LDA	WM vs. PR
[60]	NA	Synthetic	Histograms + CNN	Normal/Abnormal

4.2. Infants

After 9 weeks of age, in the proper development of the infant, writhing movements are replaced by fidgety movements that are present continuously in an awake infant. They involve the whole body and are circular movements of small amplitude and variable acceleration and they disappear after 15–20 weeks, along with the appearance of voluntary movements. As a consequence, four types of movements can then appear at this age: writhing movements (WMs), fidgety movements (FMs), poor repertoire (PR) and cramped-synchronized (CS). Studies using GMA suggest that the absence of fidgety movements between 9 and 15 weeks is the best criterion for early identification of CP and other developmental disorders [61].

To this aim, the system proposed in [62] calculated magnitude, balance and rhythm of movements by video analysis. Movements are detected by background subtractions and frame differences, whereas their classification is performed, based on the clinical definition, by using a feedforward-type neural network that includes Gaussian mixture models in a log-linearised form, enabling the estimation of the probabilistic distribution of a given sample dataset. This way, the system automatically classified the input motion images of infants during GMs into one of the four above mentioned types (WM, FM, PR or CS). Nineteen infants, including some with LBW (Low Birth Weight), were recorded either at home or hospital while standing in a crib. The movement of infants is measured using a video camera fixed directly above and parallel to the crib surface, which is covered by a unicolour fabric spread.

Automated detection and classification of presence vs. absence of FMs was the aim in [63] instead. A dataset of 2800 five-second snippets was annotated by two well-trained and experienced GMA assessors. Using OpenPose, the infant's full pose was recovered from the video stream in the form of a 25-points skeleton. This skeleton was used as an input vector for a shallow multilayer neural network in order to discriminate fidgety from non-fidgety movements.

The abovementioned works aimed at classifying movement types. However, higher-level reasoning can be introduced to directly scoring the risk of a disorder in observed infants. For example, in [64], a machine-learning model for early Cerebral Palsy (CP) prediction based on infant video recordings in NICUs has been proposed. The model was designed to assess in videos, acquired by a commercially available digital video camera, the proportion (%) of CP risk-related movements. The first processing step in the model consists of a time-frequency decomposition, by multivariate empirical mode decomposition (MEMD) and Hilbert–Huang transformation [65], of the movement parameters extracted for six body parts (arms, legs, head, and torso). Subsequently, each 5-second period in the video was clustered into 5 composite scores which were used in a linear discriminative analysis to classify movements typically found in children with or without CP. The model was developed and tested on video recordings from a cohort of 377 high-risk infants at 9–15 weeks corrected age to predict CP status and motor function (ambulatory vs. non-ambulatory).

Instead of using handcrafted features and a decision tree, in [66], image features were taken from Layer 8 of VGG19 [67], passed through a max-pooling layer and normalized before being input to a Long short-term memory (LSTM) [68] layer for classification of the image sequence. The classification outcomes label each sequence as containing normal or abnormal movements. The model, constructed using a transfer learning approach, is represented in Figure 5. Experimentally, on a dataset of videos taken from 80 CP and 135 normal subjects, it has been proved that it can classify normal videos with great confidence but struggles with intermittent classes. Classification of these borderline classes is also a difficult task for experts. How much video is required before an LSTM can identify the presence of CP is a key research question. It is hypothesized that, as a minimum, the videos should be no shorter than required for an 'expert' to make a positive identification.

Infants evaluated as at high-risk of CP, with an age of 12–24 weeks post-term, were examined in [69]. Parents and families were asked to video-record their baby through the In-Motion-App by a smartphone. Infants in videos were assessed by a motion tracker algorithm that consists of a convolutional neural network trained on 7-body points on about 15K video frames on high-risk infants. The final goal was to predict CP. This has been the first automatic infant body point tracker tested on video recordings from hand-held smartphones.

Some recent work addressed also the problem of making the proposed frameworks fully interpretable, i.e., providing an automatically generated visualization capable of relaying pertinent information to the assessor. To this aim, the framework in [70] takes video as the input and analyses the movement of individual body parts to determine if FMs are present or absent, subsequently identifying normal or abnormal general movements from segments of the sequence. The 2D skeletal pose is detected on a per-frame basis using OpenPose, hence, each pose is divided into different body parts and each body-part sequence is processed by a specific branch to learn a part-specific spatiotemporal representation (using LSTM). Finally, the outputs from all the individual body-part streams are concatenated and fed to the classifier. The label predicted by the classifier is then returned to the user as text message printed on the original video by a specific visualization module.

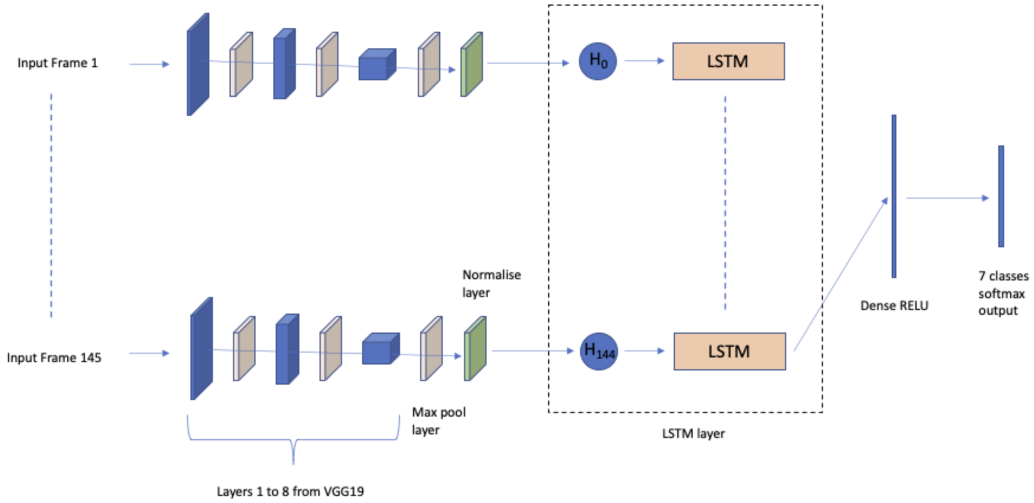

Figure 5. Normal or abnormal movement classification by means of VGG for feature extraction and LSTM for temporal modelling as proposed in [66].

Infants tend to shake their heads, extend their arms/legs, and splay their fingers when they experience pain. Therefore, body movement is considered the main indicator in several pediatric scales [71].

For example, in [72], infants having an average gestational age of 36 weeks were recorded before, during and after an acute episodic painful procedure. It was then proved that the amount of body motion presents a good indication of the infant's emotional state. The amount of motion in each video frame was computed by summing up the motion's image pixels and it was used as the main feature for classification by a threshold (pain-related movement or no pain-related movement).

Table 3 Summarizes works dealing with movements assessment in infants.

Table 3. Summarization of works dealing with movements assessment in Infants.

Work	Setup	Input	Method	Classification Goal
[62]	Home/Hospital	RGB	Motion Feature + Gaussian mixture network	4 type of mov. WMs/FMs/PR/CS
[64]	Hospital	RGB	Motion + MEMD + HT + Decision Tree	CP risk
[63]	Hospital	RGB	OpenPose+NN	FMs
[72]	Treatment Center	RGB	Amount of Motion	Pain Level
[66]	Home/Hospital	RGB	VGG9+LSTM	FMs
[70]	Home/Hospital	RGB	OpenPose+LSTM	FMs
[69]	Home	Smartphone	CIMA-Pose	CP risk

4.3. Toddlers

The term toddlers refers to children who have recently learnt to walk, i.e., having 1–2 years of age. Differently from newborns and infants, when toddlers are involved, acquisition conditions are more challenging since generally the child is acquired while freely moving and adults (and even other agents such as robots) are present, with the child being one of the smallest actors in the scene. On the other side, distinctive features of NDD can be more evident and a computer-based diagnosis becomes more feasible and reliable. In other words, for toddlers, the computer-aided clinical goal is to distinguish between children with and without NDD. Reliability and feasibility depend on the acquisition environment that could be domestic or some rehabilitative centre.

The problem of analyzing actions performed in a pediatric rehabilitation environment has been addressed in [73]. Automatically recognizing actions can help to assess infants' mobility skills and to understand if and how adults and socially assistive robots can promote their mobility. The paper proposes a multiview action classification system based on Faster R-CNN and LSTM networks that fuses information from different views by using learnable fusion coefficients derived from detection confidence scores. Pretrained deep features and detection annotations for training were used. The system is view-independent, learns features that are close to view-invariant, and can handle new or missing views at test time. The approach was tested on a small dataset (2 subjects, 10–24 months old) and on an extended dataset (6 subjects, 8–24 months old) for four action classes (crawl, sit, stand and walk).

Some interesting works concentrated on the possibility to disambiguate typically developing vs. autistic subjects using an ML approach operating on video sequences of simple executing gestures.

In [74], a Random Forest classifier [75] was trained to distinguish between Typical/Atypical Development and Autism Spectrum Disorder/Speech and Language Conditions. An interesting part of the study determined the impact of each video's annotations on the classifier's predicted label for that video by a unified approach inherited by the game theory.

Videos of a child's daily activities at home were analysed in [76]. Common categories of daily activities include "play alone", "play with others", "mealtime", and "parents' concerns". At first, the authors re-trained a 2D Mask R-CNN network to make it more robust in identifying children pose. A nonlinear state estimation technique was then exploited to predict the locations of missing key points. Finally, behavioural features were extracted from key points trajectories over a short time frame and they were exploited to feed a binary classifier to distinguish between atypical vs. typical characteristics.

In [77], a computer vision classifier for detecting head banging in home videos has been proposed. The solution uses the well-known scheme (see Figure 5) with a time-distributed convolutional neural network (CNN) in which a single CNN extracts features from each frame in the input sequence, and these extracted features are fed as input to a long short-term memory (LSTM) network. The solution achieved a 90.77% F1-score on video clips from the SSBD dataset.

The aforementioned scheme (CNN+LSTM) has been exploited in [78] for building a baseline in recognising 4 repetitive actions (spinning, arms flapping, hand action, and head banging actions) that are a potential indication of ASD disorder. Besides, the authors introduced an innovative tool which follows a Bag-of-Visual-Words configuration as reported in Figure 6. It firstly performs a person detection and tracking module by the YOLOv3 detector, feature extraction by Histogram of Optical Flow (HOF), and data clustering by K-means. Then, each video input is coded in visual words that are finally classified by MLP. Inputs are videos of 3-years old children and they are captured by parents in daily living settings. The best algorithmic pipeline (among those tested in the paper) achieved 78% of accuracy. Anyway, the proposed pipeline had several drawbacks and it did not provide satisfying outcomes in natural settings. A reason is that only 2D pose/appearance descriptors and weak models for spatiotemporal information were involved. Besides, the number of videos for tests was limited (i.e., 141) and, in most of them, there was the issue of shaking camera held by the parents during the recordings. Finally, authors in [79] applied machine learning methods (Logistic Regression with varying regularization penalties, SVM Classifiers, attempting a wide range of kernels and hyperparameters and decision trees) on ratings of different indicative features of autism from home videos.

Table 4 summarizes works dealing with movements assessment in toddlers.

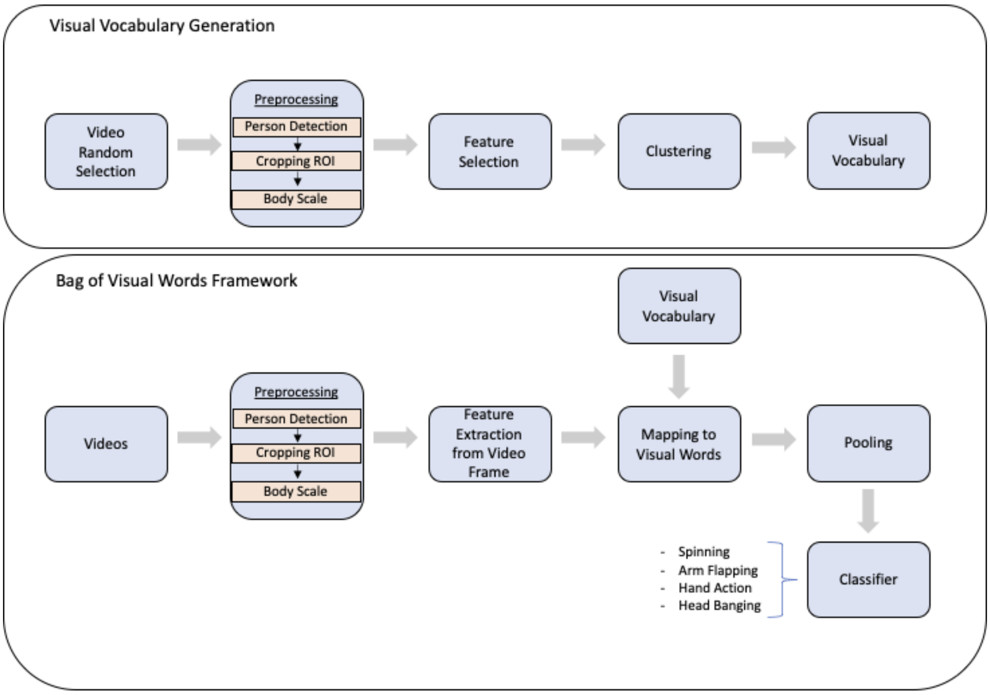

Figure 6. The innovative tool proposed in [78]. It follows a Bag-of-Visual-Words configuration for recognising 4 repetitive actions that are a potential indication of ASD disorder.

Table 4. Summarization of works dealing with movements assessment in Toddlers.

Work	Setup	Input	Method	Goal
[74]	domestic (Tariq dataset)	RGB	Random Forests	Typical/Atypical
[73]	Rehabilitation Environment	Multiview RGB	Faster R-CNN + LSTM + learnable fusion coefficients	4 daily actions
[76]	Domestic	RGB	2D Mask R-CNN + particle filter +CNN classifier	Atypical/Typical Trajectories
[78]	Domestic	RGB (from YouTube)	YOLOv3 + HOF + K-means K-means + MLP	4 repetitive Actions
[77]	Domestic (SSBD dataset)	RGB	CNN + LSTM	ASD/Typical
[79]	domestic (Tariq dataset)	RGB	Various regressors Classifiers	ASD Features Rating

5. Recent Advances in Human Motion Analysis

As described in the previous sections, the approaches for assessing children's movements rely on well-consolidated machine learning techniques. Some of them still rely on classical strategies (hand-crafted features and shallow neural networks) whereas the most recent and performing ones exploit deep learning for feature extraction or for providing end-to-end solutions. Anyway, research in machine learning is running ahead very fast [80], and thus, it could be of interest here to have a glimpse on the very latest methodologies for movement analysis which could be transferred in the considered domain of the early detection of NDD in children.

There are several possible research directions that could be pursued to improve existing frameworks aiming at computer-based early diagnosis of NDD by analysing video data. Among all, the tasks that have been attracting great attention from the machine learning and computer vision community are:

- Motion feature extraction;
- Human pose estimation;

- Extraction significant motion segments/temporal action localization;
- Human image completion;
- Action recognition and action quality assessment;
- Humans-objects interaction prediction/understanding;
- Spatiotemporal video representation;
- Interpretablilty of involved AI models.

Improving motion feature extraction could be the first pathway to explore for improving computer-aided diagnosis. To this aim, a rich and robust motion representation based on spatiotemporal self-similarity has been recently proposed [81]. Given a sequence of frames, the method represents each local region with similarities to its neighbours in space and time, enabling this way the learner to better recognize structural patterns in space and time. Code is available at https://github.com/arunos728/SELFY (accessed on 10 December 2021). Alternatively, the trainable neural module (MotionSqueeze) for effective motion feature extraction proposed in [82] could be exploited. Inserted in the middle of any neural network, it learns to establish correspondences across frames and convert them into motion features, which are readily fed to the next downstream layer for better prediction. Code is available at https://github.com/arunos728/MotionSqueeze (accessed on 10 December 2021).

In addition, improving human pose estimation could help the technologies for NDD diagnosis in children. To this aim, it is undoubtedly important to improve the localization quality of the regressed key point positions and this could be achieved by a multi-branch structure for separate regression (i.e., each branch learns a representation with dedicated adaptive convolutions and regresses one key point) as proposed in [83]. The code and models are available at https://github.com/HRNet/DEKR (accessed on 10 December 2021). An effective regression-based human-pose recognition method could be also carried out by building cascade transformers as suggested in [84] whose code has been made available at https://github.com/mlpc-ucsd/PRTR (accessed on 10 December 2021). Pose estimators still suffer from severe performance drop on corrupted images, and thus, some authors proposed to overcome this drawback by an adversarial data augmentation method together with a knowledge distillation module applied to transfer clean pose structure knowledge to the target pose estimator [85]. Code available at https://github.com/AIprogrammer/AdvMix (accessed on 10 December 2021). Operating on either pixel-level or key point-level transitions is good for analysing local and short-term motions of human bodies but not to handling higher-level spatial and longer-lasting temporal structures well.

Since children's body is often partially occluded, useful approaches could be those solving human image completion, which tries to recover the human body part with a reasonable human shape from the corrupted region. In [86], a framework for recovering the human body parts by a reasonable topological structure of the human body has been introduced. The paper proposes a structure and texture memory bank to introduce more additional priors as compensation for the corrupted region.

Skeleton-based action recognition is a very common solution also in children movement analysis. A substantial improvement of this strategy has been recently proposed in [87] where a temporal-then-spatial recalibration method, named memory attention networks (MANs), has been deployed using a temporal attention-recalibration module and a spatiotemporal convolution module.

In recent years, a number of end-to-end approaches based on 2D or 3D convolutional neural networks (CNN) have emerged for video action recognition, achieving state-of-the-art results on several large-scale benchmark datasets. An in-depth comparative analysis of available approaches on video data framing adults is available in [88]. How they can impact the movement analysis of children and the recognition of their actions is less debated instead [89]. Furthermore, their impact in the specific context of early diagnosis of NDD is totally missing. In the following, some examples of end-to-end (E2E for short) deep learning-based methods that can potentially impact the NDD diagnosis are reported. An approach with a novel temporal-spatial pooling block for action classification, which can learn pool discriminative frames and pixels in a certain clip, has been recently proposed in [90]. Similarly, in [91], an efficient spatiotemporal human action recognition framework for long and overlapping action classes has been proposed. Fine-tuned pre-trained CNN models were exploited to learn the spatial relationship at the frame level whereas an optimized Deep Autoencoder [92] was used to squeeze high-dimensional deep features. A Recurrent Neural Network (RNN) with LSTM [93] was used to learn the long-term temporal relationships.

However, RNN suffers from non-parallelism and gradient vanishing; hence, it is hard to be optimized, and then, encoder-decoder frameworks based on transformers [94] are becoming popular. For example, in the solution introduced in [95], the encoder attached with a task token aims to capture the relationships and global interactions between historical observations. The decoder extracts auxiliary information by aggregating anticipated future clip representations. Therefore, a transformer can recognize current actions by encoding historical information and predicting future context simultaneously. The code is available at https://github.com/wangxiang1230/OadTR (accessed on 10 December 2021). Following this research trend, a Video Transformer (VidTr) model with separable attention for video classification has been proposed [96]. Compared with commonly used 3D networks, these frameworks are able to aggregate spatiotemporal information via stacked attention and provide better performance with higher efficiency [97]. Multiscale Vision Transformers (MViT) for video and image recognition could be another important architecture to test for children movement analysis. It connects the seminal idea of multiscale feature hierarchies with transformer models [98]. Code is available at https://github.com/facebookresearch/SlowFast (accessed on 10 December 2021). A pure-transformer-based model for video classification, drawing upon the recent success of such models in image classification has been proposed in [99]. The model extracts spatiotemporal tokens from the input video, which are then encoded by a series of transformer layers. Code was released at https://github.com/google-research/scenic/tree/main/scenic/projects/vivit (accessed on 10 December 2021).

Training temporal action detection in videos requires large amounts of labelled data, yet such annotation is expensive to collect. This could be more and more challenging in the case of children, and even more if NDDs have to be observed. Incorporating unlabelled or weakly-labelled data to train action detection models could help reduce annotation costs. In [100], authors designed an unsupervised foreground attention module utilizing the conditional independence between foreground and background motion and put it in a Semi-supervised Action Detection (SSAD) task.

In a real-world scenario, human actions are typically out of the distribution from training data, which requires a model to both recognize the known actions and reject the unknown. Different from image data, video actions are more challenging to be recognized in an open-set setting due to the uncertain temporal dynamics and static bias of human actions. This is even more true in the case of children and in particular when we want to identify subtle behavioural differences. To overcome this issue, some researchers [101] formulated the action recognition problem from the evidential deep learning perspective and proposed a novel model calibration method to regularize the training and to mitigate the static bias of video representation through contrastive learning [102]. Code and pre-trained models used in [101] are available at https://www.rit.edu/actionlab/dear (accessed on 10 December 2021).

Another task that could be of interest in analyzing children's movement is related to the automatic Action Quality Assessment, i.e., analysing/quantifying how well an action (either spontaneous or voluntary) was performed. Assessing action quality is challenging since it has to rely on just subtle differences while performing. Mapping these differences, when found, in reliable scores is a difficult task as well. Regression strategies are commonly used to tackle this problem but they suppose to be able to extract a reliable quality score from a single video, ignoring the ineluctable large inter-video variations even when the action is performed by the same person. The consideration that relations among videos can provide important clues for more accurate action quality assessment during both training and inference inspired the work in [103]. The authors reformulated the problem of action quality assessment: instead of learning unreferenced scores, they aimed at regressing relative scores with reference to another video that has shared attributes (e.g., category and difficulty). Small intervals are considered in order to build a coarse-to-fine approach. In other words, they proposed a differential approach followed by a grouping strategy in order to achieve effective action scoring.

The PyTorch implementation of the method is available at https://github.com/yuxumin/CoRe (accessed on 10 December 2021).

Unfortunately, some children actions have a similar appearance and they require complex temporal-level relation understanding to be well analysed. Hence, a way to overcome this drawback could be to get an effective spatiotemporal video representation that could help to disambiguate them. Representing video structure as a space-time graph and discovering the discriminative sub-graphs is the solution proposed in [104]. This also leads to the elegant views of how to perform end-to-end learning of the discriminative sub-graphs, and how to nicely present the complexity of different actions in the reasoning process, which are problems not yet fully understood. The recent paper [105] studies the problem of learning self-supervised representations on videos. It presents a contrast-and-order representation framework for learning self-supervised video representation that can automatically capture both the appearance information within each frame and temporal information across different frames. Self-supervised video representation learning methods have been also addressed in [106] by two tasks to learn the appearance and speed consistency, respectively. Nevertheless, temporal modelling still remains challenging for action recognition in videos. To mitigate this issue, new video architectures with a focus on capturing multi-scale temporal information for efficient action recognition are being proposed. For example, in [107], an efficient temporal module that leverages a two-level difference modelling paradigm, assessing local and global motion, respectively, on short-term and long-term motion modelling, has been recently introduced. Code at https://github.com/MCG-NJU/TDN (accessed on 10 December 2021).

A list of the works mentioned in this section is shown in Table 5: the leftmost column indicates the referring works, the central one points out which tasks, among those involved in frameworks for the early diagnosis of NDD, have been improved. Finally, the methodological contributions that brought to the knowledge advancement are highlighted in the rightmost column.

Table 5. Recent works on human motion analysis.

Work	Improved Task	Key Contribution
[81]	Motion Features Extraction	Spatiotemporal self-similarity
[82]	Motion Features Extraction	MotionSqueeze module
[83]	Pose Estimation (Key points Positioning)	Multi-branch regression
[84]	Pose Estimation (Key points Positioning)	Cascade Transformers
[85]	Pose Estimation (Key points Positioning)	Adversarial algorithms
[86]	Human Completion	Topological Structure/Memory Bank
[87]	Skeleton-Based Action Recognition	Memory Attention Networks
[90]	Action Recognition	Temporal-Spatial pooling block
[91]	Action Recognition	CNN+Autoencoder+LSTM
[101]	Action Recognition	Contrastive Learning
[100]	Action Recognition	Semi-supervised Action Detection

Table 5. *Cont.*

Work	Improved Task	Key Contribution
[95–99]	Action Classification	Transformers
[103]	Action Quality Assessment	Contrastive Regression
[104]	video representation	Space-Time Graph
[105,106]	Video Representation	Self-supervised learning
[107]	Temporal Modeling	Two-level Motion Modeling
[108]	Motion Segment Extraction	Hierarchical Framework
[109]	Temporal Action Localization	E2E anchor free method
[110]	Temporal Action Localization	Anchor-Constrained Viterbi
[111]	Temporal Action Localization	Memory Network
[112]	Temporal Action Localization	Multi-Label Action Dependency layer
[113]	Human Object Interaction	Transformer /Cascade detector
[114]	Human Object Interaction	Graph Networks

One of the main issues when using machine learning approaches, especially in the medical domain, is to understand the rationale behind the prediction. This general problem is referred to in the literature as interpretable AI. Models are often seen as 'black boxes' in which the underlying structures can be difficult to understand. There is an increasing requirement for the mechanisms behind why systems are making decisions to be transparent, understandable and explainable. In medical scenarios, the cost of a simple prediction error could be significantly high, and thus, the reliance on the trained model and its capability to deliver both efficient and robust data processing must be guaranteed. Therefore, understanding the behaviours of machine learning models, gaining insights into their working mechanisms, and further generating explainable deep learning models have become essential and fundamental problems. Hence, another research line could be explaining the AI, and, in particular, making clearer how deep learning works for movement analysis when children are involved. To this aim, including visualization and uncertainty estimation can improve acceleration, robustness and stability making automatic tools actually exploitable in the clinical practice for early diagnosis of NDD. To this end, a starting reading could be the recent survey paper on explainable AI [115].

Another relevant issue to be furtherly investigated in order to improve computer-based diagnosis of NDD in children is the Motion Segment Extraction, which aims at detecting the temporal location of significant motion in the scene. To this aim, the authors in [108] incorporated higher-level reasoning of motion primitives by introducing a hierarchical motion understanding framework. They demonstrated also how to detect and extract significant motion segments that can be a crucial point in many diagnosis tasks. Code is available at https://sumith1896.github.io/motion2prog (accessed on 10 December 2021). A very strictly related issue is the temporal action localization, which is an important yet challenging task in video understanding. Typically, such a task aims at inferring both the action category and localization of the start and end frame for each action instance in a long, untrimmed video. They can rely either on pre-defined anchors, generated by different levels of supervision [110], or on anchor-free end-to-end trainable basic predictor [109]. Anchor-based methods generally provide a large number of outputs and require a heavy tuning of locations and sizes corresponding to different anchors. Instead, recently introduced anchor-free methods are lighter and get rid of redundant hyper-parameters. The code of the anchor-free approach proposed in [109] is available at https://github.com/TencentYoutuResearch/ActionDetection-AFSD (accessed on 10 December 2021).

Temporal action localization has been also addressed by an Expectation-Maximization (EM) framework that comprises Hidden Markov Models, MLP and self-supervised learning for action-level temporal feature embedding [116]. This way, it relaxes assumptions about the lengths of latent actions. Alternatively, in [111], an Action Unit Memory Network for weakly supervised temporal action localization was proposed. Two attention modules were designed to adaptively update the memory bank and to learn action units. Similarly, in [112], an attention based Multi-Label Action Dependency layer was introduced to improve action localization performance. The layer consists of two branches: a Co-occurrence Dependency Branch and a Temporal Dependency Branch to model co-occurrence action dependencies and temporal action dependencies. Code is available at https://github.com/ptirupat/MLAD (accessed on 10 December 2021). Finally, the problem of human–object interaction detection is very important when observing children's behaviours (especially for toddlers). It has been effectively addressed by a unified model (exploiting a transformer unit and a cascade detection over multi-scale feature maps) to jointly discover the target objects and predict the corresponding interactions in [113] and by Asynchronous-Sparse Interaction Graph Networks in [114]. Code available at https://github.com/scwangdyd/ (accessed on 10 December 2021) and https://github.com/RomeroBarata/human_object_interaction (accessed on 10 December 2021) respectively.

6. Conclusions

This paper summarizes the most relevant works on movement analysis in young children (aged 0–3) employing mainly machine learning techniques and starting from image/video data. The work was motivated by the observation that existing review papers dealt with technologies based on physical sensors. Actually, a few works concentrated on baby motion analysis from input video data and they collected only papers dealing with general movement assessment (GMA) issues. This paper addresses the more general problem of motion assessment for early diagnosis of neurodevelopmental disorders (NDD) in the first 3 years of life.

From the methodological scouting, it emerged that the approaches relying on handcrafted features and shallow classifiers were mainly exploited for fast recognition of key points (e.g., random ferns [117]) to find the positions of body joints or to analyse movement, e.g., by optical flow. Alternatively, the same task was sometimes achieved by OpenPose framework. Other well-established deep learning strategies were used in a preliminary step as well (e.g., for detecting body parts using U-Net [118]). Finally, deep learning was also exploited for the final classification outcomes (general or clinical-specific) through properly introduced architectures. In particular, the modelling of temporal dependencies is the main task assigned to deep architecture such as LSTM.

Besides, a glimpse of recent advancements in computer vision and machine learning has been provided in order to pave the way towards more effective solutions for the addressed issue. In particular, it emerged that improving deep architectures for motion feature extraction (i.e., by an additional MotionSqueeze module) could be an effective pathway to explore for improving computer-aided diagnosis. Human pose estimation has a great potential to be improved as well, for example, by multi-branch structures for a separate regression of key points.

Author Contributions: Conceptualization, M.L. and G.M.B.; writing—original draft preparation, M.L.; writing—review and editing, M.L., G.M.B., P.C.; supervision, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Larsen, M.L.; Wiingreen, R.; Jensen, A.; Rackauskaite, G.; Laursen, B.; Hansen, B.M.; Hoei-Hansen, C.E.; Greisen, G. The effect of gestational age on major neurodevelopmental disorders in preterm infants. *Pediatr. Res.* **2021**. [CrossRef]
2. Hadders-Algra, M. Early Diagnostics and Early Intervention in Neurodevelopmental Disorders—Age-Dependent Challenges and Opportunities. *J. Clin. Med.* **2021**, *10*, 861. [CrossRef] [PubMed]

3. Kundu, S.; Maurer, S.V.; Stevens, H.E. Future Horizons for Neurodevelopmental Disorders: Placental Mechanisms. *Front. Pediatr.* **2021**, *9*, 653230. [CrossRef] [PubMed]
4. Einspieler, C.; Prechtl, H.F.; Ferrari, F.; Cioni, G.; Bos, A.F. The qualitative assessment of general movements in preterm, term and young infants—Review of the methodology. *Early Hum. Dev.* **1997**, *50*, 47–60. [CrossRef]
5. Campbell, S.K.; Kolobe, T.H.; Osten, E.T.; Lenke, M.; Girolami, G.L. Construct validity of the test of infant motor performance. *Phys. Ther.* **1995**, *75*, 585–596. [CrossRef] [PubMed]
6. Heineman, K.R.; Bos, A.F.; Hadders-Algra, M. The Infant Motor Profile: A standardized and qualitative method to assess motor behaviour in infancy. *Dev. Med. Child Neurol.* **2008**, *50*, 275–282. [CrossRef] [PubMed]
7. Einspieler, C.; Prechtl, H.F. Prechtl's assessment of general movements: A diagnostic tool for the functional assessment of the young nervous system. *Ment. Retard. Dev. Disabil. Res. Rev.* **2005**, *11*, 61–67. [CrossRef]
8. Teitelbaum, P.; Teitelbaum, O.; Nye, J.; Fryman, J.; Maurer, R.G. Movement analysis in infancy may be useful for early diagnosis of autism. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 13982–13987. [CrossRef]
9. Gurevitz, M.; Geva, R.; Varon, M.; Leitner, Y. Early markers in infants and toddlers for development of ADHD. *J. Atten. Disord.* **2014**, *18*, 14–22. [CrossRef]
10. Jaspers, M.; de Winter, A.F.; Buitelaar, J.K.; Verhulst, F.C.; Reijneveld, S.A.; Hartman, C.A. Early childhood assessments of community pediatric professionals predict autism spectrum and attention deficit hyperactivity problems. *J. Abnorm. Child Psychol.* **2013**, *41*, 71–80. [CrossRef]
11. Athanasiadou, A.; Buitelaar, J.; Brovedani, P.; Chorna, O.; Fulceri, F.; Guzzetta, A.; Scattoni, M.L. Early motor signs of attention-deficit hyperactivity disorder: A systematic review. *Eur. Child Adolesc. Psychiatry* **2020**, *29*, 903–916. [CrossRef] [PubMed]
12. Balter, L.J.; Wiwe Lipsker, C.; Wicksell, R.K.; Lekander, M. Neuropsychiatric Symptoms in Pediatric Chronic Pain and Outcome of Acceptance and Commitment Therapy. *Front. Psychol.* **2021**, *12*, 836. [CrossRef] [PubMed]
13. Micai, M.; Fulceri, F.; Caruso, A.; Guzzetta, A.; Gila, L.; Scattoni, M.L. Early behavioral markers for neurodevelopmental disorders in the first 3 years of life: An overview of systematic reviews. *Neurosci. Biobehav. Rev.* **2020**, *116*, 183–201. [CrossRef] [PubMed]
14. Peyton, C.; Pascal, A.; Boswell, L.; DeRegnier, R.; Fjørtoft, T.; Støen, R.; Adde, L. Inter-observer reliability using the General Movement Assessment is influenced by rater experience. *Early Hum. Dev.* **2021**, *161*, 105436. [CrossRef]
15. Irshad, M.T.; Nisar, M.A.; Gouverneur, P.; Rapp, M.; Grzegorzek, M. Ai approaches towards Prechtl's assessment of general movements: A systematic literature review. *Sensors* **2020**, *20*, 5321. [CrossRef]
16. Wilson, R.B.; Vangala, S.; Elashoff, D.; Safari, T.; Smith, B.A. Using Wearable Sensor Technology to Measure Motion Complexity in Infants at High Familial Risk for Autism Spectrum Disorder. *Sensors* **2021**, *21*, 616. [CrossRef]
17. Ghazi, M.A.; Ding, L.; Fagg, A.H.; Kolobe, T.H.; Miller, D.P. Vision-based motion capture system for tracking crawling motions of infants. In Proceedings of the 2017 IEEE International Conference on Mechatronics and Automation (ICMA), Takamatsu, Japan, 6–9 August 2017; pp. 1549–1555.
18. Marcroft, C.; Khan, A.; Embleton, N.D.; Trenell, M.; Plötz, T. Movement recognition technology as a method of assessing spontaneous general movements in high risk infants. *Front. Neurol.* **2015**, *5*, 284. [CrossRef]
19. Cabon, S.; Porée, F.; Simon, A.; Rosec, O.; Pladys, P.; Carrault, G. Video and audio processing in paediatrics: A review. *Physiol. Meas.* **2019**, *40*, 02TR02. [CrossRef]
20. Cattani, L.; Alinovi, D.; Ferrari, G.; Raheli, R.; Pavlidis, E.; Spagnoli, C.; Pisani, F. Monitoring infants by automatic video processing: A unified approach to motion analysis. *Comput. Biol. Med.* **2017**, *80*, 158–165. [CrossRef]
21. Sun, Y.; Kommers, D.; Wang, W.; Joshi, R.; Shan, C.; Tan, T.; Aarts, R.M.; van Pul, C.; Andriessen, P.; de With, P.H. Automatic and continuous discomfort detection for premature infants in a NICU using video-based motion analysis. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 5995–5999.
22. Jorge, J.; Villarroel, M.; Chaichulee, S.; Guazzi, A.; Davis, S.; Green, G.; McCormick, K.; Tarassenko, L. Non-contact monitoring of respiration in the neonatal intensive care unit. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 286–293.
23. Lorato, I.; Stuijk, S.; Meftah, M.; Kommers, D.; Andriessen, P.; van Pul, C.; de Haan, G. Towards Continuous Camera-Based Respiration Monitoring in Infants. *Sensors* **2021**, *21*, 2268. [CrossRef]
24. Nagy, Á.; Földesy, P.; Jánoki, I.; Terbe, D.; Siket, M.; Szabó, M.; Varga, J.; Zarándy, Á. Continuous Camera-Based Premature-Infant Monitoring Algorithms for NICU. *Appl. Sci.* **2021**, *11*, 7215. [CrossRef]
25. Chaurasia, S.K.; Reddy, S. State-of-the-art survey on activity recognition and classification using smartphones and wearable sensors. *Multimed. Tools Appl.* **2021**, *81*, 1–32. [CrossRef]
26. Leo, M.; Farinella, G.M. *Computer Vision for Assistive Healthcare*; Academic Press: Cambridge, MA, USA, 2018.
27. Bouchabou, D.; Nguyen, S.M.; Lohr, C.; LeDuc, B.; Kanellos, I. A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. *Sensors* **2021**, *21*, 6037. [CrossRef] [PubMed]
28. Silva, N.; Zhang, D.; Kulvicius, T.; Gail, A.; Barreiros, C.; Lindstaedt, S.; Kraft, M.; Bölte, S.; Poustka, L.; Nielsen-Saines, K.; et al. The future of General Movement Assessment: The role of computer vision and machine learning—A scoping review. *Res. Dev. Disabil.* **2021**, *110*, 103854. [CrossRef] [PubMed]

29. Redd, C.B.; Karunanithi, M.; Boyd, R.N.; Barber, L.A. Technology-assisted quantification of movement to predict infants at high risk of motor disability: A systematic review. *Res. Dev. Disabil.* **2021**, *118*, 104071. [CrossRef]
30. Raghuram, K.; Orlandi, S.; Church, P.; Chau, T.; Uleryk, E.; Pechlivanoglou, P.; Shah, V. Automated movement recognition to predict motor impairment in high-risk infants: A systematic review of diagnostic test accuracy and meta-analysis. *Dev. Med. Child Neurol.* **2021**, *63*, 637–648. [CrossRef]
31. Rahman, M.; Usman, O.L.; Muniyandi, R.C.; Sahran, S.; Mohamed, S.; Razak, R.A.; others. A Review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain Sci.* **2020**, *10*, 949. [CrossRef]
32. Orlandi, S.; Guzzetta, A.; Bandini, A.; Belmonti, V.; Barbagallo, S.D.; Tealdi, G.; Mazzotti, S.; Scattoni, M.L.; Manfredi, C. AVIM—A contactless system for infant data acquisition and analysis: Software architecture and first results. *Biomed. Signal Process. Control* **2015**, *20*, 85–99. [CrossRef]
33. Kanemaru, N.; Watanabe, H.; Kihara, H.; Nakano, H.; Takaya, R.; Nakamura, T.; Nakano, J.; Taga, G.; Konishi, Y. Specific characteristics of spontaneous movements in preterm infants at term age are associated with developmental delays at age 3 years. *Dev. Med. Child Neurol.* **2013**, *55*, 713–721. [CrossRef]
34. Baccinelli, W.; Bulgheroni, M.; Simonetti, V.; Fulceri, F.; Caruso, A.; Gila, L.; Scattoni, M.L. Movidea: A Software Package for Automatic Video Analysis of Movements in Infants at Risk for Neurodevelopmental Disorders. *Brain Sci.* **2020**, *10*, 203. [CrossRef]
35. Tomasi, C.; Kanade T. Detection and Tracking of Point Features. *Int. J. Comput. Vis.* **1991**, *9*, 137–154. [CrossRef]
36. Caruso, A.; Gila, L.; Fulceri, F.; Salvitti, T.; Micai, M.; Baccinelli, W.; Bulgheroni, M.; Scattoni, M.L. Early Motor Development Predicts Clinical Outcomes of Siblings at High-Risk for Autism: Insight from an Innovative Motion-Tracking Technology. *Brain Sci.* **2020**, *10*, 379. [CrossRef] [PubMed]
37. Migliorelli, L.; Moccia, S.; Pietrini, R.; Carnielli, V.P.; Frontoni, E. The babyPose dataset. *Data Brief* **2020**, *33*, 106329. [CrossRef]
38. Hesse, N.; Bodensteiner, C.; Arens, M.; Hofmann, U.G.; Weinberger, R.; Sebastian Schroeder, A. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
39. Hesse, N.; Pujades, S.; Black, M.J.; Arens, M.; Hofmann, U.G.; Schroeder, A.S. Learning and tracking the 3D body shape of freely moving infants from RGB-D sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2540–2551. [CrossRef] [PubMed]
40. Schroeder, A.S.; Hesse, N.; Weinberger, R.; Tacke, U.; Gerstl, L.; Hilgendorff, A.; Heinen, F.; Arens, M.; Dijkstra, L.J.; Rocamora, S.P.; et al. General Movement Assessment from videos of computed 3D infant body models is equally effective compared to conventional RGB video rating. *Early Hum. Dev.* **2020**, *144*, 104967. [CrossRef]
41. Huang, X.; Fu, N.; Liu, S.; Vyas, K.; Farnoosh, A.; Ostadabbas, S. Invariant representation learning for infant pose estimation with small data. *arXiv* **2020**, arXiv:2010.06100.
42. Chambers, C.; Seethapathi, N.; Saluja, R.; Loeb, H.; Pierce, S.R.; Bogen, D.K.; Prosser, L.; Johnson, M.J.; Kording, K.P. Computer vision to automatically assess infant neuromotor risk. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 2431–2442. [CrossRef] [PubMed]
43. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef]
44. Rajagopalan, S.; Dhall, A.; Goecke, R. Self-stimulatory behaviours in the wild for autism diagnosis. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 755–761.
45. Rehg, J.; Abowd, G.; Rozga, A.; Romero, M.; Clements, M.; Sclaroff, S.; Essa, I.; Ousley, O.; Li, Y.; Kim, C.; et al. Decoding children's social behavior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3414–3421.
46. Tariq, Q.; Daniels, J.; Schwartz, J.N.; Washington, P.; Kalantarian, H.; Wall, D.P. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS Med.* **2018**, *15*, e1002705. [CrossRef]
47. Billing, A.E. DREAM: Development of Robot-Enhanced Therapy for Children with Autism Spectrum Disorders. EU-FP7 Grant 611391. 2019. Available online: https://github.com/dream2020/data (accessed on 20 January 2022).
48. Rihawi, O.; Merad, D.; Damoiseaux, J.L. 3D-AD: 3D-autism dataset for repetitive behaviours with kinect sensor. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
49. Prechtl, H.F. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early Hum. Dev.* **1990**; doi: 10.1016/0378-3782(90)90011-7. [CrossRef]
50. Beccaria, E.; Martino, M.; Briatore, E.; Podestà, B.; Pomero, G.; Micciolo, R.; Espa, G.; Calzolari, S. Poor repertoire General Movements predict some aspects of development outcome at 2 years in very preterm infants. *Early Hum. Dev.* **2012**, *88*, 393–396. [CrossRef]
51. Einspieler, C.; Marschik, P.B.; Bos, A.F.; Ferrari, F.; Cioni, G.; Prechtl, H.F. Early markers for cerebral palsy: insights from the assessment of general movements. *Future Neurol.* **2012**, *7*, 709–717. [CrossRef]
52. Einspieler, C.; Marschik, P.B.; Pansy, J.; Scheuchenegger, A.; Krieber, M.; Yang, H.; Kornacka, M.K.; Rowinska, E.; Soloveichick, M.; Bos, A.F. The general movement optimality score: A detailed assessment of general movements during preterm and term age. *Dev. Med. Child Neurol.* **2016**, *58*, 361–368. [CrossRef] [PubMed]

53. Hesse, N.; Stachowiak, G.; Breuer, T.; Arens, M. Estimating body pose of infants in depth images using random ferns. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 35–43.
54. Hesse, N.; Schröder, A.S.; Müller-Felber, W.; Bodensteiner, C.; Arens, M.; Hofmann, U.G. Body pose estimation in depth images for infant motion analysis. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 1909–1912.
55. Khan, M.H.; Schneider, M.; Farid, M.S.; Grzegorzek, M. Detection of infantile movement disorders in video data using deformable part-based model. *Sensors* **2018**, *18*, 3202. [CrossRef] [PubMed]
56. Barry, M.J. Physical therapy interventions for patients with movement disorders due to cerebral palsy. *J. Child Neurol.* **1996**, *11*, S51–S60. [CrossRef] [PubMed]
57. Marschik, P.B.; Pokorny, F.B.; Peharz, R.; Zhang, D.; O'Muircheartaigh, J.; Roeyers, H.; Bölte, S.; Spittle, A.J.; Urlesberger, B.; Schuller, B.; et al. A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders. *Curr. Neurol. Neurosci. Rep.* **2017**, *17*, 43. [CrossRef] [PubMed]
58. Doroniewicz, I.; Ledwoń, D.J.; Affanasowicz, A.; Kieszczyńska, K.; Latos, D.; Matyja, M.; Mitas, A.W.; Myśliwiec, A. Writhing movement detection in newborns on the second and third day of life using pose-based feature machine learning classification. *Sensors* **2020**, *20*, 5986. [CrossRef]
59. Moccia, S.; Migliorelli, L.; Carnielli, V.; Frontoni, E. Preterm infants' pose estimation with spatio-temporal features. *IEEE Trans. Biomed. Eng.* **2019**, *67*, 2370–2380. [CrossRef]
60. McCay, K.D.; Ho, E.S.L.; Shum, H.P.H.; Fehringer, G.; Marcroft, C.; Embleton, N.D. Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access* **2020**, *8*, 51582–51592. [CrossRef]
61. Øberg, G.K.; Jacobsen, B.K.; Jørgensen, L. Predictive value of general movement assessment for cerebral palsy in routine clinical practice. *Phys. Ther.* **2015**, *95*, 1489–1495.
62. Tsuji, T.; Nakashima, S.; Hayashi, H.; Soh, Z.; Furui, A.; Shibanoki, T.; Shima, K.; Shimatani, K. Markerless Measurement and evaluation of General Movements in infants. *Sci. Rep.* **2020**, *10*, 1422. [CrossRef] [PubMed]
63. Reich, S.; Zhang, D.; Kulvicius, T.; Bölte, S.; Nielsen-Saines, K.; Pokorny, F.B.; Peharz, R.; Poustka, L.; Wörgötter, F.; Einspieler, C.; others. Novel AI driven approach to classify infant motor functions. *Sci. Rep.* **2021**, *11*,9888. [CrossRef] [PubMed]
64. Ihlen, E.A.; Støen, R.; Boswell, L.; de Regnier, R.A.; Fjørtoft, T.; Gaebler-Spira, D.; Labori, C.; Loennecken, M.C.; Msall, M.E.; Möinichen, U.I.; et al. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: A multi-site cohort study. *J. Clin. Med.* **2020**, *9*, 5. [CrossRef] [PubMed]
65. Leo, M.; Looney, D.; D'Orazio, T.; Mandic, D.P. Identification of defective areas in composite materials by bivariate EMD analysis of ultrasound. *IEEE Trans. Instrum. Meas.* **2011**, *61*, 221–232. [CrossRef]
66. Schmidt, W.; Regan, M.; Fahey, M.; Paplinski, A. General movement assessment by machine learning: Why is it so difficult. *J. Med. Artif. Intell* **2019**, *2*. [CrossRef]
67. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
68. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
69. Adde, L.; Brown, A.; Van Den Broeck, C.; DeCoen, K.; Eriksen, B.H.; Fjørtoft, T.; Groos, D.; Ihlen, E.A.F.; Osland, S.; Pascal, A.; et al. In-Motion-App for remote General Movement Assessment: A multi-site observational study. *BMJ Open* **2021**, *11*, e042147. [CrossRef]
70. Sakkos, D.; Mccay, K.D.; Marcroft, C.; Embleton, N.D.; Chattopadhyay, S.; Ho, E.S. Identification of Abnormal Movements in Infants: A Deep Neural Network for Body Part-Based Prediction of Cerebral Palsy. *IEEE Access* **2021**, *9*, 94281–94292. [CrossRef]
71. Zamzmi, G.; Kasturi, R.; Goldgof, D.; Zhi, R.; Ashmeade, T.; Sun, Y. A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Rev. Biomed. Eng.* **2017**, *11*, 77–96. [CrossRef]
72. Zamzmi, G.; Pai, C.Y.; Goldgof, D.; Kasturi, R.; Sun, Y.; Ashmeade, T. Automated pain assessment in neonates. In *Scandinavian Conference on Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 350–361.
73. Pacheco, C.; Mavroudi, E.; Kokkoni, E.; Tanner, H.G.; Vidal, R. A Detection-based Approach to Multiview Action Classification in Infants. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 6112–6119.
74. Tariq, Q.; Fleming, S.L.; Schwartz, J.N.; Dunlap, K.; Corbin, C.; Washington, P.; Kalantarian, H.; Khan, N.Z.; Darmstadt, G.L.; Wall, D.P. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study. *J. Med. Internet Res.* **2019**, *21*, e13822. [CrossRef] [PubMed]
75. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
76. Vyas, K.; Ma, R.; Rezaei, B.; Liu, S.; Neubauer, M.; Ploetz, T.; Oberleitner, R.; Ostadabbas, S. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In Proceedings of the 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Pittsburgh, PA, USA, 13–16 October 2019; pp. 1–6.
77. Washington, P.; Kline, A.; Mutlu, O.C.; Leblanc, E.; Hou, C.; Stockham, N.; Paskov, K.; Chrisman, B.; Wall, D. Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, United States, Yokohama, Japan, 8–13 May 2021; pp. 1–7.

78. Negin, F.; Ozyer, B.; Agahian, S.; Kacdioglu, S.; Ozyer, G.T. Vision-assisted recognition of stereotype behaviors for early diagnosis of Autism Spectrum Disorders. *Neurocomputing* **2021**, *446*, 145–155. [CrossRef]
79. Nabil, M.A.; Akram, A.; Fathalla, K.M. Applying machine learning on home videos for remote autism diagnosis: Further study and analysis. *Health Inform. J.* **2021**, *27*, 1460458221991882. [CrossRef]
80. Gamra, M.B.; Akhloufi, M.A. A review of deep learning techniques for 2D and 3D human pose estimation. *Image Vis. Comput.* **2021**, *114*, 104282. . [CrossRef]
81. Kwon, H.; Kim, M.; Kwak, S.; Cho, M. Learning self-similarity in space and time as generalized motion for video action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13065–13075.
82. Kwon, H.; Kim, M.; Kwak, S.; Cho, M. Motionsqueeze: Neural motion feature learning for video understanding. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 345–362.
83. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14676–14686.
84. Li, K.; Wang, S.; Zhang, X.; Xu, Y.; Xu, W.; Tu, Z. Pose Recognition with Cascade Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1944–1953.
85. Wang, J.; Jin, S.; Liu, W.; Liu, W.; Qian, C.; Luo, P. When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11855–11864.
86. Zhao, Z.; Liu, W.; Xu, Y.; Chen, X.; Luo, W.; Jin, L.; Zhu, B.; Liu, T.; Zhao, B.; Gao, S. Prior Based Human Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7951–7961.
87. Li, C.; Xie, C.; Zhang, B.; Han, J.; Zhen, X.; Chen, J. Memory attention networks for skeleton-based action recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef] [PubMed]
88. Chen, C.F.R.; Panda, R.; Ramakrishnan, K.; Feris, R.; Cohn, J.; Oliva, A.; Fan, Q. Deep analysis of cnn-based spatio-temporal representations for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6165–6175.
89. Sciortino, G.; Farinella, G.M.; Battiato, S.; Leo, M.; Distante, C. On the estimation of children's poses. In *International Conference on Image Analysis and Processing*; Springer: Catania, Italy, 11–15 September 2017; pp. 410–421.
90. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Zhang, R.; Lv, X. Spatial–temporal pooling for action recognition in videos. *Neurocomputing* **2021**, *451*, 265–278. [CrossRef]
91. Bilal, M.; Maqsood, M.; Yasmin, S.; Hasan, N.U.; Rho, S. A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. *J. Supercomput.* **2021**, 1–36. [CrossRef]
92. Hong, C.; Yu, J.; Wan, J.; Tao, D.; Wang, M. Multimodal deep autoencoder for human pose recovery. *IEEE Trans. Image Process.* **2015**, *24*, 5659–5670. [CrossRef]
93. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
94. Jaderberg, M.; Simonyan, K.; Zisserman, A.; others. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2017–2025.
95. Wang, X.; Zhang, S.; Qing, Z.; Shao, Y.; Zuo, Z.; Gao, C.; Sang, N. OadTR: Online Action Detection with Transformers. *arXiv* **2021**, arXiv:2106.11149.
96. Zhang, Y.; Li, X.; Liu, C.; Shuai, B.; Zhu, Y.; Brattoli, B.; Chen, H.; Marsic, I.; Tighe, J. Vidtr: Video transformer without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13577–13587.
97. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video transformer network. *arXiv* **2021**, arXiv:2102.00719.
98. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. *arXiv* **2021**, arXiv:2104.11227.
99. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. Vivit: A video vision transformer. *arXiv* **2021**, arXiv:2103.15691.
100. Shi, B.; Dai, Q.; Hoffman, J.; Saenko, K.; Darrell, T.; Xu, H. Temporal Action Detection with Multi-level Supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 8022–8032.
101. Bao, W.; Yu, Q.; Kong, Y. Evidential Deep Learning for Open Set Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13349–13358.
102. Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G. Big self-supervised models are strong semi-supervised learners. *arXiv* **2020**, arXiv:2006.10029.
103. Yu, X.; Rao, Y.; Zhao, W.; Lu, J.; Zhou, J. Group-aware Contrastive Regression for Action Quality Assessment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 7919–7928.

104. Li, D.; Qiu, Z.; Pan, Y.; Yao, T.; Li, H.; Mei, T. Representing Videos As Discriminative Sub-Graphs for Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3310–3319.
105. Hu, K.; Shao, J.; Liu, Y.; Raj, B.; Savvides, M.; Shen, Z. Contrast and Order Representations for Video Self-Supervised Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 7939–7949.
106. Huang, D.; Wu, W.; Hu, W.; Liu, X.; He, D.; Wu, Z.; Wu, X.; Tan, M.; Ding, E. ASCNet: Self-supervised Video Representation Learning with Appearance-Speed Consistency. *arXiv* **2021**, arXiv:2106.02342.
107. Wang, L.; Tong, Z.; Ji, B.; Wu, G. TDN: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1895–1904.
108. Kulal, S.; Mao, J.; Aiken, A.; Wu, J. Hierarchical Motion Understanding via Motion Programs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6568–6576.
109. Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; Fu, Y. Learning Salient Boundary Feature for Anchor-free Temporal Action Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3320–3329.
110. Li, J.; Todorovic, S. Anchor-Constrained Viterbi for Set-Supervised Action Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9806–9815.
111. Luo, W.; Zhang, T.; Yang, W.; Liu, J.; Mei, T.; Wu, F.; Zhang, Y. Action Unit Memory Network for Weakly Supervised Temporal Action Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9969–9979.
112. Tirupattur, P.; Duarte, K.; Rawat, Y.S.; Shah, M. Modeling Multi-Label Action Dependencies for Temporal Action Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1460–1470.
113. Wang, S.; Yap, K.H.; Ding, H.; Wu, J.; Yuan, J.; Tan, Y.P. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13475–13484.
114. Morais, R.; Le, V.; Venkatesh, S.; Tran, T. Learning Asynchronous and Sparse Human-Object Interaction in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16041–16050.
115. Bai, X.; Wang, X.; Liu, X.; Liu, Q.; Song, J.; Sebe, N.; Kim, B. Explainable Deep Learning for Efficient and Robust Pattern Recognition: A Survey of Recent Developments. *Pattern Recognit.* **2021**, 120, 108102. [CrossRef]
116. Li, J.; Todorovic, S. Action Shuffle Alternating Learning for Unsupervised Action Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12628–12636.
117. Ozuysal, M.; Fua, P.; Lepetit, V. Fast keypoint recognition in ten lines of code. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
118. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.