Towards Explainable Abnormal Infant Movements Identification: A Body-part Based Prediction and Visualisation Framework

Kevin D. McCay[1], Edmond S. L. Ho[1], Dimitrios Sakkos[1], Wai Lok Woo[1],
Claire Marcroft[2], Patricia Dulson[2], and Nicholas D. Embleton[2]

[1]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK
[2]Newcastle Neonatal Service, NUTH NHS Foundation Trust, Newcastle upon Tyne, UK

***Abstract*—Providing early diagnosis of cerebral palsy (CP) is key to enhancing the developmental outcomes for those affected. Diagnostic tools such as the General Movements Assessment (GMA), have produced promising results in early diagnosis, however these manual methods can be laborious.**

In this paper, we propose a new framework for the automated classification of infant body movements, based upon the GMA, which unlike previous methods, also incorporates a visualization framework to aid with interpretability. Our proposed framework segments extracted features to detect the presence of Fidgety Movements (FMs) associated with the GMA spatiotemporally. These features are then used to identify the body-parts with the greatest contribution towards a classification decision and highlight the related body-part segment providing visual feedback to the user.

We quantitatively compare the proposed framework's classification performance with several other methods from the literature and qualitatively evaluate the visualization's veracity. Our experimental results show that the proposed method performs more robustly than comparable techniques in this setting whilst simultaneously providing relevant visual interpretability.

***Index Terms*—infants, cerebral palsy, general movements assessment, machine learning, explainable AI, visualization**

I. Introduction

Cerebral palsy (CP) is the term for a group of lifelong neurological conditions which can cause difficulties with mobility, posture and coordination. CP can also cause problems with swallowing, speech articulation, vision, and has been associated with a diminished ability to learn new skills. There is significant variance in the severity of CP, with some individuals showing very minor symptoms whilst others may be severely disabled [6]. CP is attributed to non-progressive damage to the brain in early infancy [4], [16] and is one of the most common physical disabilities in childhood. However, early diagnosis of CP can be difficult, with a confirmed diagnosis rarely made before 18 months of age [10]. The difficulty in providing an early diagnosis is problematic, as early intervention care is considered particularly important for those with emerging and diagnosed CP.

This project is supported in part by the Royal Society (Ref: IES/R1/191147).
Corresponding author: Edmond S. L. Ho (e.ho@northumbria.ac.uk)

Currently, the General Movements Assessment (GMA) is used to evaluate infant movement by manually observing spontaneous infant movements at a specific stage in development. In a typically developing infant "Fidgety Movements" (FMs) are detectable from 3 to 5 months post term [14] and consistently have a similar appearance. The absence of these movement characteristics consequently allows for abnormal FM patterns to be identified and classified [5]. However, the challenges associated with applying these assessments in practice depends upon the availability of appropriately trained clinicians. In order to address the issues surrounding manual clinical assessment, several studies have been carried out which attempt to assess the viability of automating assessments to predict motor impairment based upon observed motion quality using computer vision-based approaches [10]. Examples such as [2] [1] [3] explore a per frame background subtraction method for analysis, whereas more recent methods [17] [13] [15], propose the use of Optical Flow-based methods to track and assess infant movements. Whilst reasonable results are obtained, these methods typically struggle to deal with intra-class variation, as well as anomalies within the recorded video footage such as illumination changes, camera movement, subject-scaling, and resolution inconsistencies. This makes it difficult for these approaches to be adopted in a real-world clinical setting.

On the other hand, with the advancement of pose estimation techniques, high-quality skeletal poses can be extracted from video automatically. Recent work such as [11], [12] proposed using histogram-based pose features to automate GMA by classifying infant movements into FM+ (normal) and FM- (abnormal). The pose-based features, namely Histograms of Joint Orientation 2D (HOJO2D) [11] and Histograms of Joint Displacement 2D (HOJD2D) [11] are computed from the orientation of the body segment and the displacement of the joints, respectively. Encouraging classification performance on traditional classifiers [11] and deep learning frameworks [12] were demonstrated. Wu et al. [18] proposed *Movement Complexity Index* which determines the complexity of the body movements of the infant by computing the correlations between the movements of the joints using the Spearman

Correlation Coefficient Matrix (SCCM). Although the method focuses on analyzing the features to predict the risk level of CP of the infant without the need of the training process as in machine learning based approaches, the features are computed from 3D skeletal data which requires specialized image sensing devices to capture those data. Furthermore, the method requires the user to specify a threshold level of the computed index to separate normal/abnormal, and it is unclear if this can be generalized to other datasets.

The aforementioned studies suggest that an automated system could potentially help to reduce the time and cost associated with current manual clinical assessments, and also assist clinicians in making earlier and more confident diagnoses by providing additional information about the assessed infant movements. However, these methods are also not without their setbacks. One of the main issues with using machine-learning approaches in the medical domain is the problem of 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 [8]. As such, we propose a new motion classification and visualization framework, which takes an RGB video as the input and analyzes the movement of individual body parts to determine if FMs are present (FM+) or absent (FM-), subsequently identifying normal or abnormal general movements from segments of the sequence. To make our proposed framework fully interpretable, an important aspect is the integration of an automatically generated visualization capable of relaying pertinent information to the assessor. The visualization highlights body-parts which are showing movement abnormalities, and are subsequently providing the most significant contribution towards the classification result. As such, our proposed contributions are summarized as:

- A new body-part based classification framework for the automated prediction of CP based upon body movement extracted from videos.
- A visualization feature to highlight pertinent body-parts in the video to improve the model interpretability.

Experimental results showed that out proposed fidgety movements prediction framework achieved 100% accuracy and outperforms the existing work on the benchmark MINI-RGBD [7] dataset. The details of our proposed classification and visualisation framework are discussed in Section II. Our evaluation is discussed in Section III. Our hope is that this contribution will aid in the adoption of such technologies in this domain, through accurate, quantifiable and explainable results. A demo video is available on https://youtu.be/6CZZmWnT4mo

II. Methodology

In this section, we discuss the proposed classification and visualization framework illustrated in Figure 1.

A. Pose-based motion features

The first step of our proposed framework is extracting features from input video data. McCay et al. [11] demonstrated the effectiveness of using histogram-based motion features, namely Histograms of Joint Orientation 2D (HOJO2D) and Histograms of Joint Displacement 2D (HOJD2D), extracted from 2D skeletal poses in detecting FMs from videos. In this paper, an early fusion (i.e. concatenation) of the HOJO2D and HOJD2D is used as the input motion features, since better performance has been demonstrated [11], [12].

B. Spatiotemporal Fidgety Movement Detection

In order to detect the presence of FMs spatiotemporally, the motion features have to be extracted from 1) different body-parts and 2) different temporal segments individually. Inspired by this, we propose motion feature extraction from 5 different body-parts in the spatial domain, namely *left arm, right arm, left leg, right leg, and head-torso*. For the temporal domain, we compute HOJO2D and HOJD2D features (8 bins) for the 5 body parts from every 100-frame segment. In doing so, each video is represented by multiple histogram-based motion features accordingly. For example, a 1000-frame video will be represented by 50 fused features of HOJO2D and HOJD2D.

In this work, we formulate the FMs detection problem as a binary classification. Since each video is annotated with FM+ or FM-, we label all the fused features extracted according to the holistic annotation of the video. When training the classifier all features are used, while the temporal location information is not used. In other words, no matter whether the features are extracted from the beginning or near the end of the video, they will be used to train a single classifier. This proposed approach provides distinct advantages over previous methods, i.e. 1) the classifier will be trained by more data samples rather than using only one histogram representation for the whole video as proposed in [11], [12], and 2) a focus on the presence/absence of FMs while ignoring the temporal information when training the classifier.

We follow McCay et al. [11] on using an ensemble classifier on MATLAB R2020a that consists of a wide range of classifiers to boost the performance of the classification results. Given the multiple fused features extracted from a video, all the features will be classified as FM+ or FM-, this information is then used in visualizing the results (Section II-D). As the features were extracted in sequential order in the temporal domain, the classification result on each histogram-based motion feature is essentially detecting FMs spatiotemporally.

C. Late Fusion for Cerebral Palsy Prediction

While the method presented in Section II-B provides precise information on the presence/absence of FMs spatiotemporally, directly using all motion features as a cerebral palsy prediction for the whole video will result in sub-optimal performance since the temporal ordering is less important in the GMA than the presence/absence of sustained FMs at any point in the sequence. To tackle this problem, we propose representing each of the 5 body parts using a single scalar score s, with this being the average score of the classification result (FM+

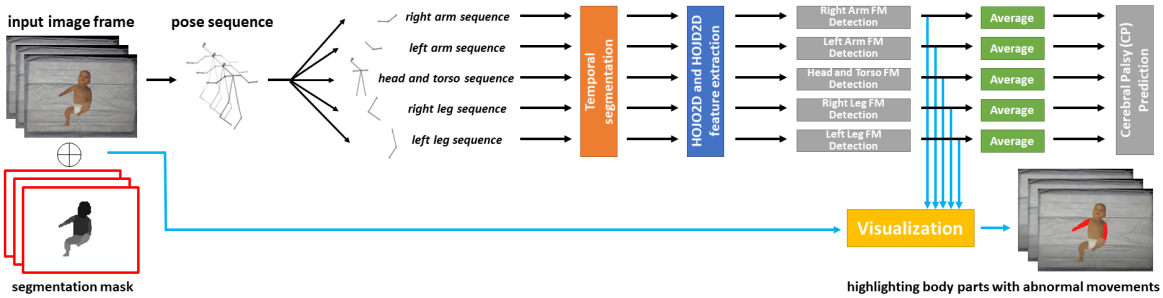

Fig. 1: The overview of the proposed prediction and visualization framework.

as 0 and FM- as 1) across all temporal segments for each body-part. Therefore, the range of s will be between 0 and 1.

Here, we propose to use a late fusion approach to train an ensemble classifier for cerebral palsy prediction. Specifically, each video is represented by using the 5 scores obtained from the body parts. The binary classifier will predict whether the motion in the video is considered *normal* or *abnormal*.

D. Visualization

While machine learning-based frameworks have obtained excellent performance in a wide range of visual understanding tasks, most of the existing frameworks can be considered black-box approaches since most of the classification frameworks only output the predicted label without specifying exactly what influences the classification decision. Whilst this is acceptable in typical computer vision tasks, it is less preferable in healthcare applications, since it is essential for the clinicians to verify the prediction as well.

To extract body part information from an input image, the CDCL [9] pre-trained body segmentation model is used in this work. The body is segmented into 6 parts; head, torso, upper arms, lower arms, pelvis and upper legs, and lower legs. An example of the segmentation result is illustrated in Figure 1 (bottom left-hand corner). Specifically, given an input infant image, CDCL [9] returns an image mask for segmentation. To align with those 5 body parts to be used in this work, we separate the segmentation masks for the arms and legs into the left and right masks. Here, k-means clustering is used to divide the pixels on each segment mask into two groups.

In order to make our proposed framework more interpretable, we include a visualization module that highlights the body-parts that are contributing to the classification decision. Our proposed method highlights the body-parts in *red* to indicate the *absence of fidgety movements* based on the scores computed in the body part abnormality detection explained in Section II-B, providing clinicians with an intuitive visualization such as the examples illustrated in Figure 2.

III. Evaluation

In this section, we evaluate the effectiveness of our proposed method using the public dataset MINI-RGBD [7] with Fidgety movement annotation by an experienced GMs assessor in [11]. We first compare the performance of our method on fidgety movement detection with baselines methods in Section III-A. Next, we present the visualization results as qualitative analysis in Section III-B. We follow the standard protocol as in [11], [12], [18] to conduct a leave-one-subject out cross-validation to ensure the results presented in this section are obtained base on *unseen data* during the training process.

Method	Accuracy	Sensitivity	Specificity
[11] w/ LDA	66.67%	50.00%	75.00%
[11] w/ SVM	83.33%	50.00%	**100.00%**
[11] w/ Decision Tree	75.00%	50.00%	87.50%
[11] w/ kNN (k=1)	75.00%	25.00%	**100.00%**
[11] w/ kNN (k=3)	50.00%	00.00%	75.00%
[11] w/ Ensemble	66.67%	50.00%	75.00%
FCNet [12]	83.33%	75%	87.5%
Conv1D-1 [12]	83.33%	75%	87.5%
Conv1D-2 [12]	91.67%	75%	**100.00%**
Conv2D-1 [12]	83.33%	75%	87.5%
Conv2D-2 [12]	83.33%	75%	87.5%
Movement Complexity Index [18]	91.67%	**100.00%**	87.5%
Our method	**100.00%**	**100.00%**	**100.00%**

TABLE I: Classification accuracy comparison between our proposed framework and baseline methods.

A. Quantitative Evaluation on the Fidgety Movement Detection Results

To demonstrate the overall performance of our proposed framework, we first evaluate the cerebral palsy prediction of the whole input video as explained in Section II-C. We compared with the existing methods and the results are presented in Table I. Using our framework, we achieved a perfect prediction with 100% accuracy. This highlights the effectiveness of our proposed framework over the previous work ([11], [12], [18]).

B. Qualitative Evaluation on the Visualization Results

We further provide qualitative results to demonstrate the effectiveness of our proposed framework. As presented in Section II-D, we detect the absence (FM-) or presence (FM+) of fidgety movements of each body part in each temporal segment (see Section II-B). The body parts with a prediction of FM- will be highlighted in red. An example is illustrated in Figure 2. Readers are referred to https://youtu.be/6CZZmWnT4mo to evaluate the visual quality of the results. From the results, it can be seen that the highlighted body-parts generally show less complex or more repetitive movements in the videos annotated