Motor Function Assessment of Children with Cerebral Palsy using Monocular Video

Motor Function Assessment of Children with Cerebral Palsy using Monocular Video

Peijun Zhao†, Moises Alencastre-Miranda†, Zhan Shen‡, Ciaran O'Neill†,
David Whiteman⋆, Javier Gervas-Arruga⋆, and Hermano Igo Krebs†

***Abstract*—The assessment of movement abilities in individuals with neurological disorders is a critical task in clinical practice. Currently, clinical assessments are time-consuming and rely on qualitative scales typically conducted by trained clinicians. Moreover, these assessments offer only coarse snapshots of a person's abilities, failing to track the minutiae of recovery over time. To overcome these limitations, we propose a machine learning approach based on spatial-temporal graph convolutional network (STGCN) to extract movement features from pose data obtained from monocular videos collected with mobile devices (e.g., smartphones, tablets). Our proposed method achieves an accuracy of around 76.6% in evaluating children with Cerebral Palsy (CP) using the Gross Motor Function Classification System (GMFCS), a 5% improvement in accuracy compared to current state-of-the-art methods, and shows substantial agreement with professional assessments based on the weighted Cohen's Kappa ($\kappa_{lw} = 0.733$). Furthermore, the proposed method can be efficiently implemented on a wide range of mobile devices in real-time or near real-time.**

***Index Terms*—Cerebral Palsy, Gross Motor Function, Machine Learning, Graph Neural Networks, Mobile Phone**

I. Introduction

Various neurological disorders, including Cerebral Palsy (CP), Metachromatic Leukodystrophy, Stroke, and Parkinson's, may impede an individual's motor control and coordination capabilities. Clinicians employ qualitative assessments to gauge patients' conditions and devise intervention strategies. The Gross Motor Function Classification System (GMFCS) is an assessment used to assess children with CP, which encompasses five levels, from those who can independently walk or run on all surfaces (level I) to those with severely restricted mobility requiring assistive devices (level V) [1]. The GMFCS is a qualitative and rudimentary nominal scale that non-professionals struggle to use accurately [2] [3]. As a result, typically the GMFCS assessment requires visits to a clinic, where the clinical evaluator asks the child to perform a variety of physical exercises so that they can observe and classify the child's movement abilities. A typical assessment session lasts around an hour, and must be conducted regularly to assess any intervention, which can place a large time burden on the family or caregiver over the course of a treatment.

†Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA ({zhaoymn, moisesam, ciaranon, hikrebs}@mit.edu). Dr. Hermano Igo Krebs is an IEEE Fellow.

‡Zhan Shen was with Robotics Institute, University of Michigan (zhanshen@umich.edu).

⋆Takeda Development Center Americas, Inc., Lexington, MA, USA ({david.whiteman, javier.gervas}@takeda.com)

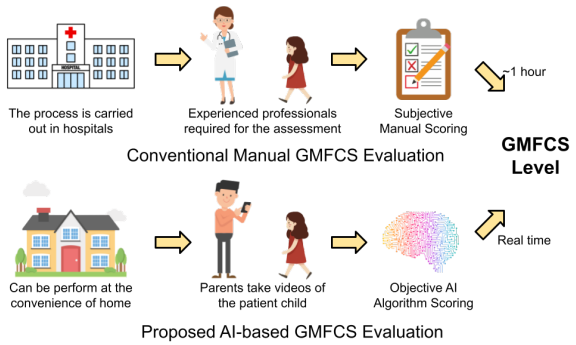


Fig. 1. The proposed AI-based GMFCS Assessment is much more convenient, faster, and cheaper than conventional evaluation.

Recent progress in the field of machine learning and computer vision has opened up a plethora of opportunities to assess a person's movement abilities in general and a CP youngster's in particular. Researchers have used videos showing the child performing different tasks to determine their GMFCS levels [4]. AI-enabled GMFCS evaluation would be much more convenient, particularly when employing a single camera. It can also be administered at home, affording long-term, almost continuous tracking of the children's motor function. Our approach might be particularly suited to monitor the child developmental status, to evaluate the clinical efficacy of therapeutic interventions and adjust treatment plans. The need to use a standardized method to classify gross motor function in CP led to the development of the GMFCS scale [5], which is subject to a high degree of subjectivity with significant inter-rater variability (sometimes κ_{lw} as low as 0.64 [6]). Computer vision approaches minimize such variability, making them ideally suited for clinical trials, registries, and for telemedicine.

In previous computer vision based motion assessment works, the skeleton/pose information was extracted from the video using mature off-the-shelf tools, e.g., OpenPose [7]. To further extract insights from these skeleton data, prior art typically rely on hand-picked/crafted features. As a result, the accuracy is upper-bounded by the qualities of the features.

In this work, we adopted an agnostic end-to-end data-driven approach. Considering that the motion can be seen as the variation of human skeleton graph across time, we employed a Spatial-Temporal Graph Neural Network (STGCN) to get the overall movement features used for GMFCS classification. We conducted a comprehensive evaluations of our proposed method and provide an ablation study to gain a deeper understanding of the proposed approach. Results show that our method has close agreement with ground truth professional la-

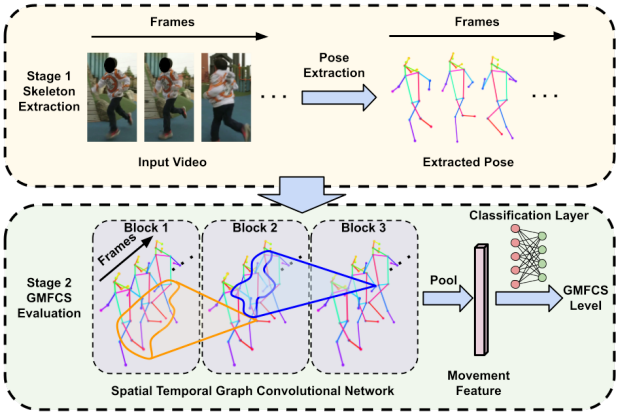


Fig. 2. The workflow of our proposed method.

belling. Finally, we evaluate the running time of our proposed method on different of mobile platforms. We release the code[1].

II. Related Works

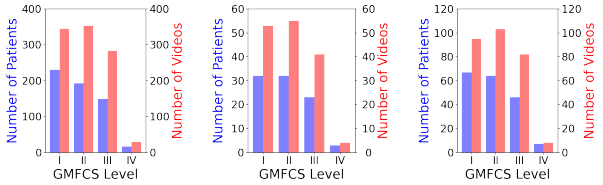
With the recent advancements in computer vision, there has been a growing interest in utilizing it for movement analysis in CP. Previous studies have mainly focused on predicting CP from infant movements using machine learning techniques with hand-crafted features [8]–[11]. The current state-of-the-art employing computer vision was proposed by Kidzinski et al. [4], which uses a skeleton-based approach and applies a 1-D convolutional neural network on the time-series data of a few expert-selected keypoints and several hand-crafted features, with experiments conducted on a large dataset with thousands of videos. In this paper, we propose a novel method for this task using the same dataset, and compare both methods.

III. Proposed Method

Our proposed, two-step workflow is shown in Fig. 2. For each input video, we first run human detection and tracking algorithms, and then apply pose estimation algorithms on the segments of each detected human. The first step can be done with different off-the-shelf methods, including bottom-up approaches like OpenPose [7] as in previous work.

The core contribution in this paper lies in step 2, where we utilize Spatial-Temporal Graph Convolutional Networks [12] to extract the movement features, and a classification module to make the final assessment based on the latent movement features. The STGCN consists of multiple blocks, and in each block, the information of each keypoint and its neighbors in spatial and temporal dimensions are aggregated with convolutions, which are used as the feature for the next block. After the final block, the information is gathered with a pooling operation to get an overall feature vector, which is further classified. STGCN and its variants have shown very good performance on human action recognition. Human action recognition are quite similar to GMFCS assessment tasks, as both attempt to extract and classify movement features from a sequence of human poses, from both spatial and temporal perspectives. Due to the scarcity of medical data,

[1]https://github.com/zhaoymn/gmfcs_stgcn



(a) Training (b) Validation (c) Testing

Fig. 3. Number of patients and videos in the dataset.

we adopt transfer learning, with the STGCN pre-trained on an action recognition dataset named "NTU RGB+D 120" [13]. The original classification layer from the pre-trained model is removed and replaced by our own classification module, which contains 2 linear layers with 4 final output neurons for GMFCS levels I to IV.

IV. Evaluation

A. Dataset

We used a publicly available dataset from Kidzinski et. al. [4]. This dataset contains videos from CP youngsters collected in a clinical setting with their GMFCS level assessed by a health care professional (ground truth). Average age of the youngsters is $11y.o.$ $(s.d. = 5.9)$, average height of $133cm$ $(s.d. = 22)$, and weight of $34kg$ $(s.d. = 17)$. The original paper lacks some details on how to reproduce the exact training, validation and testing split, and we cannot get the same dataset split running their provided code, therefore we used our own protocol for pre-processing.

We use data with GMFCS levels I to IV, because children at level V cannot move by themselves. We checked all the skeleton videos and manually removed 85 videos that contained more than one person, after which we have 1,450 videos from 861 patients. We split the dataset into training, validation and testing. We used stratified sampling and sample each GMFCS level separately. For each GMFCS score, we split the dataset using the patient's ID with ratio of 7:1:2, as we did not want any patient to appear in any two of the training, validation, and testing datasets. The detailed ground truth GMFCS level distribution of the dataset is shown in Fig. 3.

We further sampled the videos using a sliding window with length of 124 frames per sample as in [4] with an overlap of 90% between samples. We kept samples with an average over 80% keypoint availability.

B. Implementation

We implemented our method using PyTorch. Pieces of our code and pre-trained STGCN model were borrowed from Pyskl [14]. We used trainable adjacent matrix in graph aggregation and used MSTCN for temporal convolution. For our training policy, we first froze the pre-trained backbone STGCN for 3 epochs to train the classification layers, and then the last 2 STGCN blocks were unfrozen for further task specific training. We used Adam optimizer with initial learning rate of 1e-4, which gradually decreases, and weight decay of 5e-5. The model was trained for 10 epochs with batch size of 128 and the best weights were selected according to validation accuracy.

C. Performance Comparison

We evaluated our proposed approach on the testing set, where each sample was classified and the final result of the video was determined via a majority voting approach. Our method performance was then compared against the prior state-of-the-art approach [4]. Previous method put the displacement of 8 joints and another 8 hand-crafted features into separate channels of time-series data, and use 1D Convolutional Neural Network for further temporal feature extraction. The key difference between our method and prior art is that we do not hand-pick or select features, and we put spatial constraints (graph topology) during training to fuse information from different joints.

To have a fair comparison, we directly run their official code on our dataset split, with their provided pre-processing, training and testing pipelines. We were able to produce better results than the numbers reported in their original paper (accuracy 66%). Since network training involves randomness, we run each method 5 times. The prior state of the art method is able to achieve an accuracy of 71.61% $(s.d.\ 0.76\%)$, and our method can get 76.60% $(s.d.\ 0.35\%)$. To provide further analysis, we pick one model from each method and compare them in Fig. 4.

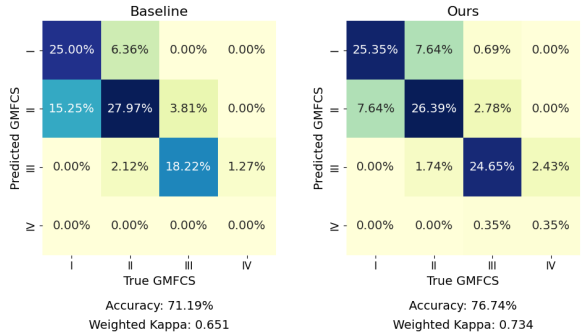


Fig. 4. Our proposed method outperformed the previous method in terms of accuracy and linear weighted Cohen's Kappa, which is used to measure the agreement of two voters, i.e., clinician and AI algorithm in our case.

As can be seen in the results, our proposed approach has an accuracy of around 5% higher than the previous approach. According to the confusion matrix, the error mostly happens between Level I and Level II. This is because GMFCS Level I and Level II are inherently similar, and it's very challenging for machine learning methods to learn the subtle difference. Furthermore, the ground truth labels provided by healthcare professionals could be sub-optimal, as these two levels could be confusing to human raters as well. As for the two methods compared here, our proposed approach correctly classified 75.3% of the Level I and Level II samples, while the baseline approach has an accuracy of 69.1%. We believe that this is due to the much stronger representation ability of our proposed model, which captured more subtle features in these two levels. As a result, our method could possibly perform even better if the quality and quantity of the training dataset further improve.

Also, we can see that both models struggle to correctly classify Level IV samples, which is because we do not have sufficient Level IV training data.

Our method has a linear weighted Kappa of 0.734, much higher than the previous method 0.651, which represents a substantial agreement with ground truth. Since the linear weighted Kappa of two professionals can sometimes be as low as 0.64 [6], thus our proposed method can be considered to work quite well compared to clinicians.

D. Ablation Study

To get a better understanding of the proposed method, we performed an ablation study. We compared the following 3 variations of our model: (1) Fixed: The weights of the backbone STGCN were kept fixed after loading the pre-trained model; (2) All: All the blocks of the backbone STGCN were fully unfreezed to be fine-tuned; (3) No-Pre: The weights of the backbone STGCN were trained from scratch with the dataset. The results are summarized in Fig. 5. When we fixed the STGCN weights to the pre-trained weights, the accuracy is significantly worse. This may be due to the domain difference between action recognition and the GMFCS scoring. On the other hand, when we allow all the STGCN blocks to be trainable, the accuracy is just a bit worse than our proposed training method, showing that our approach is generally robust to how many STGCN blocks are involved in fine-tuning. However, if we don't load the pre-trained weights, the performance is much worse, because the model significantly overfits to the training data due to limited dataset size.

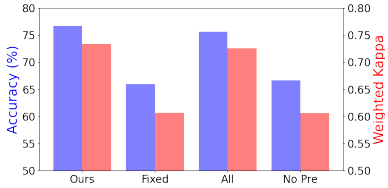


Fig. 5. Ablation Study.

E. Running Speed on Mobile Devices

As the videos of the patients are considered sensitive data, ideally, we would want to do all the computing on end user devices, so that no visual data is transferred online. We evaluate the runtime of the proposed method on mobile devices across a variety of platforms using a web APP, which allows cross-platform adoption of our system. Note that although it's a web APP, the computation happens on client end, and no visual data is transferred to the server. We use PoseNet [15] from Tensorflow.js as pose extractor, which runs with WebGL backend on GPU. The STGCN PyTorch model is converted to ONNX model and runs with ONNX Runtime Web. Due to some unsupported operators with ONNX Runtime Web's WebGL backend, we run it with WASM backend instead, which uses CPU. The experiment is carried out within a React application running with Chrome browser, and the results are summarized in Table. I.

As shown in the results, the pose extraction can reach around 30 FPS on the latest mobile phones (e.g., Samsung S23U) by utilizing the onboard GPU. Though the STGCN runs slower due to it running on the CPU, it is only called once in a while when there are enough frames (124) to process, and

TABLE I
RUNNING TIME ON MOBILE DEVICES

Device	Platform	CPU	GPU	PoseNet (ms)	STGCN (ms)
Samsung S23U	Android	Snapdragon 8.2	Adreno 740	35.1	116.3
Samsung Tablet S8+	Android	Snapdragon 8.1	Adreno 730	47.7	121.4
Google Pixel 4a	Android	Snapdragon 730G	Adreno 618	66.7	475.2
Apple iPhone 7 plus	iOS	A10 Fusion	PowerVR 7XT+	87.6	873.2
ASUS ROG Strix	Windows	Core i9-12900H	RTX 3080Ti	8.2	76.7
Alienware x14	Linux	Core i7-12700H	RTX 3060m	12.4	77.5

by using the CPU it can actually run in parallel to the pose extraction. As a result, the proposed method can be considered to run close to real time at mobile devices.

V. DISCUSSION

It is still an ongoing discussion on whether to partially use expert knowledge or completely rely on neural networks for feature extraction from the original data. While expert knowledge can be useful in tasks with clear physical laws or mathematical operations, it may be difficult to derive concise mathematical formulations for other problems. Using neural networks for feature extraction enables the building of complex models in a data-driven manner, which may be a better representation than expert-driven manual models.

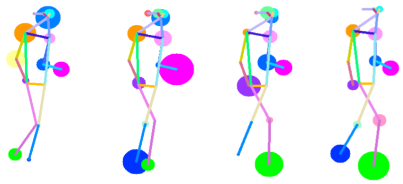


Fig. 6. The change of saliency map during one step.

In the case of GMFCS assessment, we believe that a data-driven approach to extract features could possibly outperform expert-defined features, as evidenced by Section IV. Furthermore, it is worth stating that with the advent of deep learning, neural networks are to a lesser degree a "black box": there are interpretability opportunities of deep neural networks. For example, we can calculate saliency maps [16] which helps us to identify the most important input features. Fig. 6 shows an example of the change in saliency map of the input human pose graph during one step. As we can see, the focus of the neural network is mostly on the upper body when the step begins, and the attention shifts to the lower body as the human subject moves. As a result, features from the upper body might also be useful for GMFCS assessment, which were otherwise neglected in the previous method [4]. Designing better features could also be possible, which can be especially useful when the training set is small. We leave this to future work.

VI. CONCLUSION

In this study, we utilized computer vision AI techniques for GMFCS estimation and compared it to therapist assessments. We used STGCN based networks to learn spatial and temporal features of human pose information from single-view videos. The results show the proposed method to be around 5% more accurate than the prior art (76.60%, up from 71.61%), and is in fundamental agreement with the ground truth professional labeling (average $\kappa_{lw} = 0.733$). Additionally, the proposed approach runs in near real-time on various mobile platforms. Our study suggests AI-based GMFCS assessment has great potential for smart and personalized healthcare. Future work will focus on further improving the accuracy, decreasing runtime, and estimating the uncertainty.

ACKNOWLEDGEMENTS

This research was supported by Takeda Development Center Americas, INC. (successor in interest to Millennium Pharmaceuticals, INC.) MIT Grant #6947514.

REFERENCES

[1] A. Paulson and J. Vargus-Adams, "Overview of four functional classification systems commonly used in cerebral palsy," in *Children (Basel)*, vol. 4(4), no. 30, 2017.

[2] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, M. Bax, D. Damiano, B. Dan, and B. Jacobsson, "A report: The definition and classification of cerebral palsy," in *Developmental Medicine & Child Neurology*, vol. 109, 2007.

[3] G. Rackauskaite, P. Thorsen, P. V. Uldall, and J. R. Østergaard, "Reliability of gmfcs family report questionnaire," *Disability and rehabilitation*, vol. 34, no. 9, pp. 721–724, 2012.

[4] Ł. Kidziński, B. Yang, J. L. Hicks, A. Rajagopal, S. L. Delp, and M. H. Schwartz, "Deep neural networks enable quantitative movement analysis using single-camera videos," *Nature communications*, vol. 11, no. 1, p. 4054, 2020.

[5] R. Palisano, P. Rosenbaum, S. Walter, D. Russell, E. Wood, and B. Galuppi, "Development and reliability of a system to classify gross motor function in children with cerebral palsy," *Developmental medicine & child neurology*, vol. 39, no. 4, pp. 214–223, 1997.

[6] B. C. McDowell, C. Kerr, and J. Parkes, "Interobserver agreement of the gross motor function classification system in an ambulant population of children with cerebral palsy," *Developmental Medicine & Child Neurology*, vol. 49, no. 7, pp. 528–533, 2007.

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] K. D. McCay, P. Hu, H. P. H. Shum, W. L. Woo, C. Marcroft, N. D. Embleton, A. Munteanu, and E. S. L. Ho, "A pose-based feature fusion and classification framework for the early prediction of cerebral palsy in infants," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 8–19, 2021.

[9] N. Silva, D. Zhang, T. Kulvicius, A. Gail, C. Barreiros, S. Lindstaedt, M. Kraft, S. Bölte, L. Poustka, K. Nielsen-Saines *et al.*, "The future of general movement assessment: The role of computer vision and machine learning–a scoping review," *Research in developmental disabilities*, vol. 110, p. 103854, 2021.

[10] D. Sakkos, K. D. Mccay, C. Marcroft, N. D. Embleton, S. Chattopadhyay, and E. S. Ho, "Identification of abnormal movements in infants: A deep neural network for body part-based prediction of cerebral palsy," *IEEE Access*, vol. 9, pp. 94 281–94 292, 2021.

[11] H. Zhang, E. S. Ho, and H. P. Shum, "Cp-agcn: Pytorch-based attention informed graph convolutional network for identifying infants at risk of cerebral palsy," *Software Impacts*, vol. 14, p. 100419, 2022.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[13] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[14] H. Duan, J. Wang, K. Chen, and D. Lin, "Pyskl: Towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7351–7354.

[15] D. Oved, I. Alvarado, and A. Gallo, "Real-time human pose estimation in the browser with tensorflow.js," Available at: https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html, 2018, accessed on: Apr. 29, 2023.

[16] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.