Open video data sharing in developmental science and clinical practice

Peter B. Marschik, Tomas Kulvicius, Sarah Flügge, ..., Florentin Wörgötter, Christa Einspieler, Dajie Zhang

peter.marschik@med.uni-goettingen.de

Highlights

Face-blurring is an adequate and efficient solution for sharing movement video data

General movement assessment (GMA) can be reliably done on pseudonymized videos with blurred infant faces

Automated GMA delivers comparable classifications with or without head keypoints

Innovative solutions are needed for safe, fair, and efficient video data sharing

Marschik et al., iScience 26, 106348
April 21, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.106348

OPEN ACCESS
Open video data sharing in developmental science and clinical practice

Peter B. Marschik, Tomas Kulvicius, Sarah Flügge, Claudius Widmann, Karin Nielsen-Saines, Martin Schulte-Rüther, Britta Hünig, Sven Bölte, Luise Poustka, Jeff Sigafoos, Florentin Wörgötter, Christa Einspieler, and Dajie Zhang

SUMMARY

In behavioral research and clinical practice video data has rarely been shared or pooled across sites due to ethical concerns of confidentiality, although the need of shared large-scaled datasets remains increasing. This demand is even more imperative when data-heavy computer-based approaches are involved. To share data while abiding by privacy protection rules, a critical question arises whether efforts at data de-identification reduce data utility? We addressed this question by showcasing an established and video-based diagnostic tool for detecting neurological deficits. We demonstrated for the first time that, for analyzing infant neuromotor functions, pseudonymization by face-blurring video recordings is a viable approach. The redaction did not affect classification accuracy for either human assessors or artificial intelligence methods, suggesting an adequate and easy-to-apply solution for sharing behavioral video data. Our work shall encourage more innovative solutions to share and merge stand-alone video datasets into large data pools to advance science and public health.

INTRODUCTION

Data is the origin of knowledge gain and scientific progress. In recent decades, our private and professional lives have undergone rapid and significant changes in the way we receive, generate, and disseminate ‘data’. Among the ever-expanding nearly uncontrollable sources and quantity of information in our daily life, rigorous data of scientific value are arguably rare. Obtaining high-quality data in behavioral science needs commitment of participants and the endorsement of funders and institutions. Even more, elegant data owe a great debt to the ingenuity and unfailing efforts of the scientists who acquired them. When studying infant and child development, for example, data acquisition often requires longitudinal designs that can involve years or even decades of data collection to capture development across groups and settings. Video recordings are an important and widely used method of data collection in such studies. These data document dynamic human behaviors in real time and space. Sharing these valuable assets in the field, just like sharing data in other scientific fields, can maximize the benefits of resources, avoid redundant investment, improve study visibility, transparency and reproducibility, provide training resources, promote novel knowledge, and catalyze new cooperation.

Data sharing, however, faces a thicket of thorny issues, which have triggered contentious discourses by researchers, leading academic journals, funders and international organizations in the past years. Although researchers are increasingly expected by funders, the scientific community, and tax payers to share data, a lack of actual incentives for primary researchers, aggravated by the extra operational costs and efforts necessary to curate data, impede such practice. Especially, when working with video data with identifiable individuals, ensuring the protection of participants’ confidentiality is critical. Although major legal regulations such as the Health Insurance Portability and Accountability Act of the United States and the General Data Protection Act (GDPR) of the European Union all oblige protection of data containing full-face images, which are regarded as sensitive personal identifiers, practical guidance on how such data can, if at all, be shared while protecting participants’ confidentiality is scant. For data collected in the past, for example, before GDPR was in force in 2018, clinicians and researchers could not foresee all the future needs and legal updates for data sharing beyond the scope of the original plan, hence may have missed the chance to obtain participants’ consent for data sharing with other third parties (i.e., beyond

---

**iScience**

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
what was consented to at the project start). How then, can we embrace societal interests and privacy frameworks and legitimate sharing of scientific video data? In this study, we addressed this issue by empirically testing a simple and widely used pseudonymization approach in a specific scientific setting. Essentially, we would like to invite, through our work, fellow scientists and the community to revisit the discussion on data sharing, especially the sharing of video data, and seek practical approaches to make this happen for the good of sustainable science and public health.

When investigating behavioral and neurofunctional development with infants, non-intrusive methods are especially desirable. Among such approaches, the Prechtl’s GMA\(^\text{19}\) has become a worldwide established clinical tool, applicable during the very first months of human life, for identifying heightened risk for neurological impairment such as cerebral palsy.\(^\text{19–22}\) GMA is renowned for its non-intrusiveness, excellent predictive validity, and peerless efficiency concerning required diagnostic time and resources\(^\text{23}\) (for methodological details please see\(^\text{19,24}\)). The presence or absence of the fidgety movements (FM), for example, an age-specific motor pattern observable from the third to the fifth month in typically developing infants, has proven to be a highly sensitive and specific predictor for neurological deficits.\(^\text{21,22}\) As a standard GMA requires only a 3-min guided video recording of an infant’s spontaneous whole-body movements, which will be clinically evaluated by trained experts. It can be easily implemented in daily clinical routines or at the infant home, thus being especially flexible and suitable for either high- or low-resource settings. With GMA, high risk for neurological impairments can be excluded or detected (e.g.,\(^\text{20,22}\)), and potential intervention may be introduced early, which will mitigate long-term cost and burden for the health system.\(^\text{25}\) GMA relies on human visual gestalt perception to classify typical vs. atypical infant motor patterns. As such, the excellence of the assessors does require specific high-quality training on the one hand and continuous practice and recalibration on the other hand.

Despite the fact that GMA is accredited globally and indicated for application for the youngest population at risk for adverse neurological outcomes, the extent to which it has been scaled up in practice is still limited. One may presume that as artificial intelligence (AI) approaches can avoid unfavorable human and environmental factors affecting clinical reasoning, they are likely to have the potential to bolster GMA and outspend its application. Indeed, we have seen a boom of computer-based approaches to complement the classic man-powered GMA during the past decade.\(^\text{26}\) Unfortunately, shared expert-annotated large GMA datasets are still absent. While large datasets are generally required to train machine learning algorithms, shared expert-annotated and approved large datasets are indispensable for evaluating and comparing performances of different AI approaches from different groups.\(^\text{27–29}\) Moreover, as mentioned above, if such shared data repositories would be available within the scientific community, they would contribute enormously to train and recalibrate human GMA assessors. It is impossible for any single research or clinical site to accumulate sufficient amount of valid scientific data, for example through performing GMA, to cover diverse conditions of various etiologies. To train human assessors, as well as computational models, to achieve reliable performance with high sensitivity and specificity, adequate data representing different classes (e.g., typical versus atypical GMs; movements from children with normal versus adverse neurological outcomes) are imperative. Sharing data across centers, hence seems to be the ultimate way out.

Besides obtaining participants’ informed consent to data sharing,\(^\text{30}\) algorithmic face blurring is a widely used approach for pseudonymization. It protects the privacy of individuals when sharing visual data sources across industries from street mapping and social media to pictorial journalism. This technique has also been applied in the scientific fields, such as neuroimaging and dentistry.\(^\text{31–33}\) Face blurring commonly covers the eye region, which retains part of the facial expressions and could enhance data utility despite redaction. Face blurring is a straightforward method and is far easier to apply than, for example, generating avatars or synthetic surrogate faces.\(^\text{34,35}\) Is it viable to apply face blurring approaches for video data sharing in research and practice in infant

### Table 1. Intra-rater reliability

| Subset 1 | Subset 2 | Subset 3 | Combined (840 snippets) |
|----------|----------|----------|-------------------------|
| Assessor 1 | 0.90 [0.83, 0.97] | 0.97 [0.93, 1] | 0.97 [0.94, 1] | 0.95 [0.92, 0.97] |
| Assessor 2 | 0.87 [0.79, 0.94] | 0.93 [0.88, 0.99] | 0.94 [0.89, 0.99] | 0.91 [0.88, 0.95] |

Intra-rater agreement (Cohen’s kappa $k$ and [0.95 CI]) for classes FM + vs. FM between the original face-visible condition and the face-blurred condition for the three non-overlapping randomly chosen data subsets each of 280 snippets. 

*Correspondence: pETER.marSchik@med.uni-goettingen.de
https://doi.org/10.1016/j.isci.2023.106348

OPEN ACCESS
and child development? With the current study, we aim to examine the viability of using face blurring in this field by showcasing GMA, given its aforementioned scientific and clinical significance in child health.

In particular, we ask for the first time whether human assessors are able to perform comparable GMA when the infant faces are visible or blurred. Focusing on the classification of FM (presence vs. absence), we hypothesize that the performances of well-trained human GMA assessors do not differ in the two conditions: face-visible vs. face-blurred. Simultaneously, we ask whether AI methods could deliver comparable movement classifications using features with or without head key points (analogs of face-visible vs. face-blurred conditions). We hypothesize that the performances of the AI method are also comparable in the two different conditions.

RESULTS

Movement classification by human assessors

Classification performance by human assessors in the face-visible condition

As reported in Reich et al.,27 out of the 2,800 snippets, 990 were labeled by at least one assessor as “not assessable”, mirroring infants’ frequently fluctuating behavioral status. Of the remaining 1,810 snippets, 1,784 snippets (98.6%) were labeled identically by both assessors: either FM+ (N = 956) or FM- (N = 828). For classes FM+ and FM-, the inter-rater agreement was excellent (Cohen’s kappa $\kappa = 0.97$, with 0.95 CI [0.96, 0.98]). The intra-rater reliability by rerating 280 randomly chosen snippets (i.e., 10% of the sample) for the two classes was Cohen’s kappa $\kappa = 0.95$ with 0.95 CI [0.91, 1] for assessor 1 and $\kappa = 0.85$ with 0.95 CI [0.78, 0.93] for assessor 2.27

Comparing assessors’ performance in the face-blurred condition to that in the face-visible condition

The performances of the two assessors in the current face-blurred condition are compared with their own ratings in the original face-visible condition. Classification results for the classes FM+ vs. FM- are presented in Table 1 for the two conditions for each assessor. For both assessors and across all the three subsets, the

| Number of neurons | Without head key points | With head key points |
|-------------------|-------------------------|----------------------|
|                   | One fully connected layer (64 filters of size 7x1) |                      |
|                   | Mean                    | CI (95%)             | Mean                    | CI (95%)             |
| 50                | 83.5219                 | $\pm 2.1751$         | 86.8297                 | $\pm 2.5079$         |
| 100               | 83.4665                 | $\pm 2.6224$         | 86.9402                 | $\pm 1.4941$         |
| 150               | 84.0259                 | $\pm 1.5520$         | 86.8838                 | $\pm 0.4613$         |
| 200               | 84.2511                 | $\pm 2.2011$         | 85.8183                 | $\pm 1.8428$         |
| 300               | 85.0349                 | $\pm 1.2096$         | 86.1557                 | $\pm 1.4715$         |
| 500               | 83.8589                 | $\pm 3.0408$         | 86.7150                 | $\pm 1.2738$         |
|                   | Two fully connected layers (64 filters of size 7x1) |                      |
|                   | Mean                    | CI (95%)             | Mean                    | CI (95%)             |
| 50, 25            | 84.3616                 | $\pm 2.7607$         | 86.4363                 | $\pm 1.9307$         |
| 100, 50           | 85.1465                 | $\pm 1.3817$         | 86.6028                 | $\pm 0.5126$         |
| 150, 100          | 86.7712                 | $\pm 2.0580$         | 87.1074                 | $\pm 0.8705$         |
| 200, 100          | 86.0432                 | $\pm 2.7556$         | 86.4914                 | $\pm 1.6482$         |
| 300, 150          | 86.0435                 | $\pm 2.3402$         | 84.5860                 | $\pm 1.5265$         |
| 300, 200          | 87.1636                 | $\pm 3.1892$         | 87.1628                 | $\pm 1.3906$         |
| 500, 250          | 84.4717                 | $\pm 2.4624$         | 87.2206                 | $\pm 1.6098$         |
| 500, 300          | 85.4268                 | $\pm 1.2196$         | 85.9311                 | $\pm 0.9800$         |

Classification accuracy of fidgety movements without and with head key points when using network architectures with one and two fully connected layers, respectively. Bold Numbers represent highest average classification accuracy within each group. Error bars denote confidence intervals of mean (95%).
classifications (FM+ vs. FM-) in the face-blurred condition presented excellent to perfect agreements with each assessor’s original classification in the face-visible condition. Note that the assessors did not receive any feedback nor communicated with each other throughout the rating procedure. Still, for both assessors, the performances, i.e., the agreement between the face-blurred vs. face-visible conditions improved from the first to the third subset.

**Inter-rater agreement between assessors in the face-blurred condition**

In the original face-visible condition, the two assessors’ agreement with each other for classes FM+ and FM-with the entire sample (i.e., 2800 snippets) was Cohen’s kappa $\kappa = 0.97$ with 0.95 CI [0.96, 0.98]. In the current experiment with face-blurred snippets, the agreement on classes FM+ and FM-between the two assessors was 0.78 with 0.95 CI [0.69, 0.87] for the first, 0.89 with 0.95 CI [0.82, 0.96] for the second, and 0.99 with 0.95 CI [0.99, 1] for the third subset of snippets. Again, although there was no feedback to the assessors nor communication between the assessors all along, the inter-rater agreements between the assessors also increased from the first to the third subset.

**Class “not assessable”**

Out of the 840 snippets, more were labeled “not assessable” in the face-visible condition (277 by assessor 1 and 249 by assessor 2) than were in the face-blurred condition (266 by assessor 1 and 188 by assessor 2). For assessor 1, 184 out of the 840 snippets were labeled in both conditions as “not assessable” and for assessor 2, 123 were.

**Movement classification with machine learning approach**

Results of classification performance when using networks with and without head key points are presented in Tables 2 and 3. Classification accuracies in most of the cases are above 86% and below 88% for both models with or without head features. In Figure 1 we compare best classification accuracy scores obtained for the networks with one and two fully connected (FC) layers. When comparing network architectures, results show that, although on average networks with two FC layers lead to better classification performance than networks with one FC layer (85.03% and 86.94% vs. 87.16% and 88.17%, see also caption of Figure 1).

| Filter size | Without head key points | With head key points |
|-------------|-------------------------|----------------------|
|             | Mean CI (95%)           | Mean CI (95%)         |
| 5x1         | 72.6364 ± 10.0515       | 76.6857 ± 6.3535     |
| 7x1         | 86.0432 ± 2.7556        | 86.4914 ± 1.6482     |
| 9x1         | 86.2668 ± 1.8780        | 86.9394 ± 1.2111     |
| 15x1        | 84.0270 ± 3.5293        | 87.5558 ± 1.4402     |
| 21x1        | 84.7555 ± 1.8379        | 87.8927 ± 0.5532     |
| 31x1        | 85.8750 ± 2.0184        | 88.1733 ± 2.4673     |

Classification accuracy of fidgety movements without and with head key points when using different filter sizes and different numbers of filters in the convolutional layer, respectively. The network with two fully connected layers was used in this case with 200 and 100 neurons per layer. Bold numbers represent highest average classification accuracy within each group. Error bars denote confidence intervals of mean (95%).

| Number of filters | Filter size 7x1, two fully connected layers (200, 100) |  |
|-------------------|-----------------------------------------------------|------|
|                   | Mean CI (95%)                                       | Mean CI (95%) |
| 16                | 85.5948 ± 2.0099                                    | 84.6407 ± 1.1507 |
| 32                | 84.6442 ± 3.9150                                    | 85.2024 ± 1.9710 |
| 64                | 86.0432 ± 2.7556                                    | 86.4914 ± 1.6482 |
| 128               | 86.2682 ± 2.1049                                    | 85.9864 ± 1.5832 |
| 256               | 86.5480 ± 3.0186                                    | 86.7161 ± 2.0170 |
| 512               | 84.8097 ± 1.9802                                    | 85.9873 ± 1.4872 |

Table 3. Classification accuracy
Most importantly, when comparing models with and without head features, results demonstrate that there is no statistically significant difference when comparing classification accuracy with and without head key points (see caption of Figure 1; see also STAR Methods, Figures 2–4). Results show that the absence of head key points does not have a significant effect on the specific classification of FM (i.e., FM + vs. FM -).

**DISCUSSION**

This study explored the feasibility of one simple and widely used approach, i.e., face-blurring, for pseudonymization of video data in a scientific setting assessing infant neuromotor functions. We demonstrated for the first time that movements classification performances of both human GMA assessors and an AI approach remain intact after the videos are pseudonymized by face-blurring.

Despite increasing calls for data-sharing from leading funding organizations and journals, sharing data appears to be more desired than practiced.36–39 Even many authors declare their willingness to share data, most do not respond or decline data access requests when asked.37 The low compliance rate (i.e., less than 7%) was the same as for authors who did not provide a willing-to-share-data statement.37 Barriers seem to persist to stop scientists actually sharing their data.39,41 Many researchers decline data sharing by claiming to protect the data subjects’ confidentiality. This concern becomes even stronger when it comes to video data sharing, where participants’ identities are likely to be more difficult to conceal, and little practical guidance is available to support video data sharing.

While deidentification is certainly a key issue in video data sharing to protect participants’ confidentiality, a crucial question naturally arises whether efforts at data redaction reduce data utility. Our data suggest that well-trained experienced human GMA assessors’ performances are not negatively affected by face-blurring, although the assessors may benefit from a brief adaptation to the altered video presentation in practice (i.e., rating face-obscured infants). This was suggested by the slightly lower, although still excellent accuracy of classification for the very first subset of the experiment by both assessors, followed by even higher, nearly perfect performance in the later subsets fully comparable to that in the standard face-visible condition (Table 1). Note that the assessors in this study rated the face-blurred snippets without getting familiarized to the video presentation (face-blurred) beforehand, thus the result exemplifies the true
performance variations, if any, that human GMA assessors may experience with this specific pseudonymization approach. Future studies need to sample more assessors with diverse GMA experiences to examine whether different raters are able to work with face-blurred GMA video data and deliver unaffected movement classification.

When comparing classification models with and without head features, results demonstrate that there is a slight drop in classification performance (see Figure 1). In general, it is expected that less features may lead to worse classification accuracy. However, in this case there is no statistically significant difference when comparing classification accuracy with and without head key points (see caption of Figure 1). These results suggest that for detecting the presence or absence of FM, the performance of the AI approach remains comparable regardless whether the facial-features of the infants are excluded or not. The current classification accuracy is comparable to our previous study in FM classification 27 and to a more sophisticated neural architecture of Nguyen-Thai and colleagues. 28

Our results suggest the viability, for both human assessors and computer-based methods, of using face-blurring techniques to pseudonymize and share video data for movement analyses. These are promising news for scientists worldwide who study infant neuromotor development. Movements video data can be deidentified without reducing utility, enabling multicentred sharing and pooling data. In light of the GDPR, the challenge of proper application of pseudonymization to personal data remains. Indeed, there is no single easy solution that works for all research purposes in all possible scenarios (European Union Agency for Cybersecurity17). Moreover, wherever there is a pseudonymization technology, there might be an opposing approach (i.e., reidentification) to it. 42

With our study, we intend to promote exploration and discussion on efficient and innovative solutions to share different types of valuable video data documenting human behaviors in clinical and scientific settings. This includes solutions to existing data for which participants’ consent for data sharing may not

**Figure 2. Body key points**

We used key points 1–5 for face blurring, and key points 1–21 (with head key points) or key points 1–16 (without head key points) for movement classification. Key points 22–25 were not used in this study due to poor position estimation of these key points.
be available or achievable any more. There are several approaches which allow protecting participants’ privacy while sharing video data, such as directly providing skeletal key-points, \(^1\) skinned multi-infant linear model (SMIL \(^2\)), or depth images \(^3\). When sharing only the skeletal key-points data, for example, a wealth of information of the videos will be removed. Access to any features beyond the skeleton points would no longer be possible. It is yet unknown whether human assessors would be able to work with skeletal data to assign different behaviors to the designated classes. Moreover, further analyses all rely on the specific method used to extract the skeleton key-points. The applicability and utility of this data redaction method is limited compared to the face-blurring approach, which contains rich information of the original videos that could be processed further to answer various research questions by both human assessors and computer-based approaches. The SMIL model utilizes depth images (RGB-D data). As the majority of existing video data for GMA were recorded with a single RGB camera instead of with multiple cameras or RGB-D cameras, these existing data cannot be pseudonymized by the SMIL models.

It has to be pointed out that to apply the face-blurring approach, data-cleaning before sharing is necessary. In case of GMA, infants’ active-wakiness is required for assessment. \(^4\) After the faces are blurred, the user cannot reliably detect the infants’ behavioral states, as reported under “class ‘not assessable’” in our results section. As the assessors in the face-blurred condition can only see obscured faces, judging whether or not a 5-s snippet was “not assessable” (i.e., the infant presents one or more of the following states: fussy/crying, drowsy, yawning, refluxing, over-excited, self-soothing, or distracted by the environment) was challenging. Obviously, if the assessors could see the entire faces of the infants, they could more easily identify the non-assessable snippets. Some other technical attempts to deidentify individuals on video footages while preserving their dynamic facial attributes in real-time have been made (e.g., “face-swap”). \(^5\) Such approaches need to be empirically tested for their utility, reliability, and efficiency for easy implementation in research and clinical practices.

Data sharing is a systemic ambition requiring diverse skillsets encompassing scientific, technological, financial, administrative, political, ethical, and legal issues. \(^6\) Scientists who welcome data sharing calls and regulations urgently need systematic in-practice support from policy makers, from professionals and experts in data curation and data protection, so that their capacity and mindset may focus mainly on the subject matter of science. Individual video data may belong to the most sensitive and resource-consuming type of data to collect, curate, and process, within and beyond infant and child development research. If researchers do not have to worry that their data might be misinterpreted or misused by other beneficiaries, or their original ideas behind the data might be scooped, more scientists might embrace and practice data sharing and exchange. Scientists who share video data behind their studies ought to be fairly and adequately accredited and rewarded for their investment. This way, data sharing, especially video data sharing will become incentivized and fruitful, connecting disciplines in scientific and clinical communities, ultimately benefiting science and public health.

**Limitations of the study**

The face-blurring solution discussed in this study may not be applicable for another research setting, such as studies on social interaction, as face-blurring would eliminate data utility in studies relying on facial data. The authors acknowledge that there exists no single easy solution that works for all research purposes in all

---

**Figure 3. Pseudonymization - Face blurring**

Flow diagram of face blurring procedure.

(A) original image.

(B) extracted body key points using OpenPose. \(^5\)

(C) obscured face with a blurring mask.
possible scenarios. Moreover, wherever there is a pseudonymization technology, there might be an opposing approach (i.e., reidentification) to it. Our work shall prime more innovative research tackling the problem of safely and responsibly sharing of scientific video data.

Another limitation of the study is that only two human GMA assessors were involved in the current proof-of-concept exploration. To examine whether a data deidentification approach retains data utility for human users, future studies are required. Larger sample of human assessors with different training and practicing backgrounds to test the user experience is necessary. A data redaction approach is only viable if the users are able to work with the redacted data. This needs to be empirically tested for any proposed de-identification method.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Dataset and participants
- METHOD DETAILS
  - Movements classification by human assessors 1
  - Face blurring procedure
  - Movement classification by human assessors 2
  - Automated movement classification
- QUANTIFICATION AND STATISTICAL ANALYSIS

ACKNOWLEDGMENTS

First and foremost, we would like to thank all families for their participation, their support for basic science, their patience with scientists, and commitment to our studies. The authors would like to thank team members involved in recruitment, data acquisition, and curation: Dr. Florian Pokorny, Magdalena Krieger-Tomantschger, Dr. Robert Peharz, Iris Tomantschger, Laura Langmann, Claudia Zitta, and Gunter Vogrinec. We thank Lennart Jahn and Dr. Simon Reich for data pre-processing. We were/are supported by BioTechMed-Graz and the Deutsche Forschungsgemeinschaft (DFG – stand-alone grant 456967546, SFB1528 – project C03), the Laerdal Foundation, the Bill and Melinda Gates Foundation (OPP1128871), the Volkswagen Foundation (project IDENTIFIED), the LeibnizScience Campus, the BMBF Germany, the Austrian Science Fund (KLI811), and the Fondation Paralysie Cérébrale (ENSEMBLE-II) for data acquisition.
preparation and analyses. Special thanks also to our interdisciplinary international network of collaborators for discussing this study with us and for refining our ideas and analytical processes in a global consortia approach.

AUTHOR CONTRIBUTIONS
Conceptualization, P.B.M, T.K. C.E., and D.Z.; Methodology, P.B.M, T.K., L.P., S.B., K.N., and D.Z.; Data Curation S.F., C.W., and C.E.; Software, T.K., S.F., C.W., and F.W.; Formal Analysis, P.B.M., T.K., and D.Z.; Writing – Original Draft, P.B.M, T.K. and D.Z.; Writing – Review & Editing, S.F., C.W., K.N., M.S., B.H., S.B., L.P., J.S., F.W., and C.E.; Visualization, T.K.; Supervision, P.B.M., T.K., F.W., C.E., and D.Z.

DECLARATION OF INTERESTS
The authors declare no competing interests.

INCLUSION AND DIVERSITY
We support inclusive, diverse, and equitable conduct of research.

Received: September 12, 2022
Revised: December 19, 2022
Accepted: March 2, 2023
Published: March 7, 2023

REFERENCES

1. Bauchner, H., Golub, R.M., and Fontanarosa, P.B. (2016). Data sharing: an ethical and scientific imperative. JAMA 315, 1237–1239. https://doi.org/10.1001/jama.2016.2420.

2. Drazen, J.M., Morrissey, S., Malina, D., Hamel, M.B., and Campbell, E.W. (2016). The importance — and the complexities — of data sharing. N. Engl. J. Med. 375, 1182–1183. https://doi.org/10.1056/NEJMe1611027.

3. Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., du Sert, N.P., Simonsen, U., Wagenmakers, E.-J., Ware, J.J., and Ioannidis, J.P.A. (2017). A manifesto for reproducible science. Nat.HumBehav. 1, 0021. https://doi.org/10.1038/s41593-016-0021.

4. Chevrier, R., Foulfi, V., Gaudet-Blavignac, C., Robert, A., and Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J. Med. Internet Res. 21, e13484. https://doi.org/10.2196/13484.

5. Gudi, N., Kamath, P., Chakraborty, T., Jacob, A.G., Parsenkar, S.S., Sarbadhikari, S.N., and John, O. (2022). Regulatory frameworks for clinical trial data sharing: scoping review. J. Med. Internet Res. 24, e33591. https://doi.org/10.2196/33591.

6. Kiley, R., Peatfield, T., Hansen, J., and Reddington, F. (2017). Data sharing from clinical trials - a research funder’s perspective. N. Engl. J. Med. 377, 1990–1992. https://doi.org/10.1056/NEJMe1708278.

7. Ledford, H. (2017). Open-data contest unearths scientific gems — and controversy. Nature 543, 299. https://doi.org/10.1038/nature.2017.21372.

8. Malin, B., and Goodman, K., Section Editors for the IMIA Yearbook Special Section. Y.S.S. (2018). Between access and privacy: challenges in sharing health data. Yearb. Med. Inform. 27, 55–59. https://doi.org/10.1055/s-0038-1641216.

9. Pisani, E., Aaby, P., Breugelmans, J.G., Carr, D., Groves, T., Helsini, M., Kamuya, D., Kern, S., Littler, K., Marsh, V., et al. (2016). Beyond open data: realising the health benefits of sharing data. BMJ 355, i299. https://doi.org/10.1136/bmj.i2995.

10. PLOS Medicine Editors (2016). Can data sharing become the path of least resistance? PLoS Med. 13, e1001949. https://doi.org/10.1371/journal.pmed.1001949.

11. Rosenbaum, L. (2017). Bridging the data-sharing divide — seeing the devil in the details, not the other camp. N. Engl. J. Med. 376, 1902. https://doi.org/10.1056/NEJMp1704482.

12. Ursin, G., Malila, N., Chang-Claude, J., Gunter, M., Kaaks, R., Kampman, E., Lambe, M., van Leeuwen, F., Magnusson, P., Nilbert, M.C., et al. (2019). Sharing data safely while preserving privacy. Lancet 394, 1902. https://doi.org/10.1016/s0140-6736(19)32603-0.

13. Peterson, E.D., and Rockhold, F.W. (2018). Finding means to fulfill the societal and academic imperative for open data access and sharing. JAMA Cardiol. 3, 793–794. https://doi.org/10.1001/jamacardio.2018.0129.

14. International Consortium of Investigators for Fairness in Trial Data Sharing, Devereaux, P.J., Guyatt, G., Gerstein, H., Connolly, S., and Yusuf, S. (2016). toward fairness in data sharing. N. Engl. J. Med. 375, 405–407. https://doi.org/10.1056/NEJMsb1616595.

15. Lo, B., and DeMets, D.L. (2016). Incentives for clinical trialists to share data. N. Engl. J. Med. 375, 1112–1115. https://doi.org/10.1056/NEJMsa1608351.

16. Bierie, B.E., Crossas, M., and Pierce, H.H. (2017). Data authorship as an incentive to data sharing. N. Engl. J. Med. 376, 1684–1687. https://doi.org/10.1056/NEJMsb1616595.

17. European Union Agency for Cybersecurity (2019). Pseudonymisation Techniques and Best Practices: Recommendations on Shaping Technology According to Data Protection and Privacy Provisions (Publications Office).

18. Office for Civil Rights (2012). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (HHS.gov).

19. Prechtl, H.F., Einspieler, C., Cioni, G., Bos, A.F., Ferrari, F., and Sontheimer, D. (1997). An early marker for neurological deficits after perinatal brain lesions. Lancet 349, 1361–1363. https://doi.org/10.1016/s0140-6736(97)60182-3.

20. Einspieler, C., Bos, A.F., Kriebel-Tomantschger, M., Alvarado, E., Barbosa, V.M., Bertoncelli, N., Burger, M., Chorna, O., Del Secco, S., Deltignier, R.A., et al. (2019). Cerebral palsy: early markers of clinical phenotype and functional outcome. J. Clin. Med. 8, 1616. https://doi.org/10.3390/jcm81101616.

21. Morgan, C., Romeo, D.M., Chorna, O., Novák, I., Galea, C., Del Secco, S., and Guzzetta, A. (2019). The pooled diagnostic accuracy of neuroimaging, general movements, and neurological examination for diagnosing cerebral palsy early in high-risk infants: a case control study. J. Clin. Med. 8, 1879. https://doi.org/10.3390/jcm8111879.
22. Novak, I., Morgan, C., Adde, L., Blackman, J., Boyd, R.N., Brunstrom-Hernandez, J., Cioni, G., Damiano, D., Darrah, J., Eliasson, A.C., et al. (2017). Early, accurate diagnosis and early intervention in cerebral palsy: advances in diagnosis and treatment. JAMA Pediatr. 171, 897–907. https://doi.org/10.1001/jamapediatrics.2017.1689.

23. Bosanquet, M., Copeland, L., Ware, R., and Boyd, R. (2013). A systematic review of tests to predict cerebral palsy in young children. Dev. Med. Child Neurol. 55, 418–426. https://doi.org/10.1111/dmcn.12140.

24. Einspieler, C., Prechtl, H., Bos, A., Ferrari, F., and Cioni, G. (2004). Prechtl’s method on the role of computer vision and machine learning - a future of general movement assessment: the Connectome Project informatics: quality visualization. Neuroimage 20, 202–219. https://doi.org/10.1016/j.neuroimage.2015.05.077.

25. Mikulcn, E., Russo, S., Zauli, F., d’Orio, P., Parmigiani, S., Favaro, J., Knight, W., Squezra, S., Perri, P., Cardinale, F., et al. (2021). A comparative study between state-of-the-art MRI deidentification and AnonyMi: a new method combining re-identification risk reduction and geometrical preservation. Hum. Brain Mapp. 42, 5523–5534. https://doi.org/10.1002/hbm.25639.

26. Silva, N., Zhang, D., Kulvicius, T., Gail, A., Reich, S., Zhang, D., Kulvicius, T., Bolte, S., Bruno, K., and Siemons, J. (2020). Enhancing early detection of neurological and developmental disorders and provision of intervention in low-resource settings in Uttar Pradesh, India: study protocol of the G.A.N.E.S.H. programme. BMJ Open 10, e037335. https://doi.org/10.1136/bmjopen-2020-037335.

27. Toldo, M., Varishthananda, S., Einspieler, C., Tripathi, N., Singh, A., Verma, S.K., Vishwakarma, K., Zhang, D., Diwedi, A., Gupta, R., et al. (2020). Enhancing early detection of neurological and developmental disorders and provision of intervention in low-resource settings in Uttar Pradesh, India: study protocol of the G.A.N.E.S.H. programme. BMJ Open 10, e037335. https://doi.org/10.1136/bmjopen-2020-037335.

28. Silva, N., Zhang, D., Kulvicius, T., Gail, A., Barreiros, C., Lindstaedt, S., Kraft, M., Bolte, S., Pousta, L., Nielsen-Sanes, K., et al. (2021). The future of general movement assessment: the role of computer vision and machine learning - a scoping review. Res. Dev. Disabil. 110, 103854. https://doi.org/10.1016/j.rsd.2021.103854.

29. Reich, S., Zhang, D., Kulvicius, T., Bolte, S., Nielsen-Sanes, K., Pokorny, F.B., Pezar, R., Pousta, L., Wörgötter, F., Einspieler, C., and Marschik, P.B. (2021). Novel AI-driven approach to classify infant motor functions. Sci. Rep. 11, 9888. https://doi.org/10.1038/s41598-021-8947-5.

30. Nguyen-Thai, B., Le, V., Morgan, C., Badawi, N., Tran, T., and Venkatesh, S. (2021). A spatio-temporal attention-based model for infant movement assessment from videos. IEEE J. Biomed. Health Inform. 25, 3911–3920. https://doi.org/10.1109/jbhi.2021.3077957.

31. Groos, D., Adde, L., Aubert, S., Boswell, L., de Regnier, R.A., Fjortoft, T., Gaebler-Spira, D., Haukeland, A., Loennecken, M., Møll, A., et al. (2022). Development and validation of a deep learning method to predict cerebral palsy from spontaneous movements in infants at high risk. JAMA Netw. Open 5, e2221325. https://doi.org/10.1001/jamanetworkopen.2022.21325.

32. Marcus, D.S., Harms, M.P., Snyder, A.Z., Jenkinson, M., Wilson, J.A., Glasser, M.F., Barch, D.M., Arche, K.A., Burgess, G.C., Ramaratnam, M., et al. (2013). Human Connectome Project informatics: quality control, database services, and data visualization. Neuroimage 80, 202–219. https://doi.org/10.1016/j.neuroimage.2015.05.077.

33. Milchenko, M., and Marcus, D. (2013). Obscuring surface anatomy in volumetric imaging data. Neuroinformatics 11, 65–75. https://doi.org/10.1007/s12021-012-9160-3.

34. Grasshof, S., Ackermann, H., Brandt, S.S., and Ostermann, J. (2021). Multilinear modelling of faces and expressions. IEEE Trans. Pattern Anal. Mach. Intell. 43, 3540–3554. https://doi.org/10.1109/TPAMI.2020.2986496.

35. Meden, B., Enerlici, Z., Struc, V., and Peer, P. (2018). k-Same-Net: k-anonymity with generative deep neural networks for face deidentification. Entropy 20, 60. https://doi.org/10.3390/e20010060.

36. Danchev, V., Min, Y., Borghj, J., Baozich, M., and Ioannidis, J.P.A. (2021). Evaluation of data sharing after implementation of the International Committee of Medical Journal Editors data sharing statement requirement. JAMA Netw. Open. 4, e2033972. https://doi.org/10.1001/jamanetworkopen.2020.33972.

37. Gabelica, M., Bojić, R., and Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: mixed-methods study. J. Clin. Epidemiol. 150, 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019.

38. Ohmann, C., Moher, D., Siebert, M., Motschall, E., and Naudet, F. (2021). Status, use and impact of sharing individual participant data from clinical trials: a scoping review. BMJ Open 11, e049228. https://doi.org/10.1136/bmjopen-2021-049228.

39. Rowhani-Farid, A., and Barnett, A.G. (2016). Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ 352, h619. https://doi.org/10.1136/bmj.h619.

40. Houtkoop, B.L., Chambers, C., Nichols, T.E., and Bishop, D.V.M., et al. (2016). Many researchers were not compliant with their published data sharing statement: mixed-methods study. J. Clin. Epidemiol. 150, 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019.

41. Thambawita, V., Isaksen, J.L., Hicks, S.A., Ghouse, J., Alhberg, G., Linneberg, A., Grarup, N., Ellervik, C., Olesen, M.S., Hansen, T., et al. (2021). DeepFake electrocardiograms using generative adversarial networks and the meaning of the end for privacy issues in medicine. Sci. Rep. 11, 21896. https://doi.org/10.1038/s41598-021-02195-2.

42. Yang, H.C., Rahmanti, A.R., Huang, C.W., and Li, Y.C.J. (2022). How can research on artificial empathy be enhanced by applying Deepfakes? J. Med. Internet Res. 24, e29506. https://doi.org/10.2196/29506.

43. Rydzewska, L.H.M., Stewart, L.A., and Tierney, J.F. (2017). Sharing individual participant data: through a systematic reviewer lens. Trials 23, 167. https://doi.org/10.1186/s13063-021-05787-4.

44. Warren, E. (2016). Strengthening research through data sharing. Nature. 535, 401–403. https://doi.org/10.1038/NEJMp1607282.

45. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: real-time multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. 43, 172–186. https://doi.org/10.1109/TPAMI.2019.2929257.
53. Marschik, P.B., Pokorny, F.B., Peharz, R., Zhang, D., O’Muircheartaigh, J., Roeyers, H., Bolte, S., Spittle, A.J., Urlesberger, B., Schuller, B., et al. (2017). A novel way to measure and predict development: a heuristic approach to facilitate the early detection of neurodevelopmental disorders. Curr. Neurol. Neurosci. Rep. 17, 43. https://doi.org/10.1007/s11910-017-0748-8.

54. Krieber-Tomantschger, M., Pokorny, F.B., Krieber-Tomantschger, I., Langmann, L., Poustka, L., Zhang, D., Treue, S., Tanzer, N.K., Einspieler, C., Marschik, P.B., and Körner, C. (2022). The development of visual attention in early infancy: insights from a free-viewing paradigm. Infancy. 27, 433–458. https://doi.org/10.1111/ina.12449.

55. Cao, X., Li, X., Ma, L., Huang, Y., Feng, X., Chen, Z., Zeng, H., and Cao, J. (2022). AggPose: Deep Aggregation Vision Transformer for Infant Pose Estimation. Preprint at arXiv. arXiv:2205.05277. https://doi.org/10.24963/ijcai.2022/700.

56. Groos, D., Adde, L., Staen, R., Ramampiaro, H., and Ihlen, E.A.F. (2022). Towards human-level performance on automatic pose estimation of infant spontaneous movements. Comput. Med. Imaging Graph. 95, 102012. https://doi.org/10.1016/j.compmedimag.2021.102012.

57. Chambers, C., Seethapathi, N., Saluja, R., Loeb, H., Pierce, S.R., Bogen, D.K., Prosser, L., Johnson, M.J., and Kording, K.P. (2020). Computer vision to automatically assess infant neuromotor risk. IEEE Trans. Neural Syst. Rehabil. Eng. 28, 2431–2442. https://doi.org/10.1109/TNSRE.2020.3029121.

58. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.
STAR METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Software and algorithms | | |
| OpenPose | Cao et al., 2021 | https://doi.org/10.1109/TPAMI.2019.2929257, https://github.com/CMU-Perceptual-Computing-Lab/openpose |
| Face Blurring Algorithm | This paper | https://doi.org/10.5281/zenodo.7624553 |
| Network architectures for movement classification | This paper | https://doi.org/10.5281/zenodo.7624553 |
| Other | | |
| Skeleton key-points used for movement classification | This paper | https://doi.org/10.5281/zenodo.7624553 |

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Peter B Marschik (peter.marschik@med.uni-goettingen.de or peter.marschik@ki.se).

Materials availability
This study did not generate new unique reagents or materials.

Data and code availability
- Original video data cannot be shared due to legal restrictions (privacy protection). All other data is available from the lead contact upon request.
- All code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Dataset and participants
To address our research questions, the baseline performance of human GMA assessors in the natural, face-visible condition is needed. These data are available and can be adopted from our previous study. 27 In that study, data from a prospective longitudinal cohort of 51 typically developing infants were analyzed (f/m: 26/25). 27 Data acquisition was conducted at iDN’s BRAINtegrity lab at the Medical University of Graz, Austria, within an umbrella study profiling typical cross-domain development during the first months of life. 34 The movement data in form of RGB video stream were collected in a standard laboratory setting following the Prechtl’s general movements Assessment guidelines. 24 Details on data recording and participants information have been described before. 27,53,54 Note that all the 51 participants were typically developing infants and presented normal age-adequate general movements at all evaluated age points (i.e., seven measurements per infant, biweekly from 4 to 16 weeks of post-term age). For the machine learning algorithm presented in that study, 2800 5-s video-snippets were randomly selected from a total of 19,451 available snippets. We demonstrated that 5s is sufficient for the well-trained and experienced GMA raters to detect the presence vs. absence of FM in the specific research setting. 27 The dataset with the 2800 snippets was used for the current study. Data segmentation, annotation and analyses were performed at the Systemic Ethology and Developmental Science Unit - SEE, Department of Child and Adolescent Psychiatry and Psychotherapy at the University Medical Center Göttingen, Germany. The study was approved by the Institutional Review Board of the Medical University of Graz, Austria (27–476ex14/15) and the University Medical Center Göttingen, Germany (20/9/19). Parents were informed of all experimental procedures.
and study purpose, and provided their written informed consent for participation and publication of results.

**METHOD DETAILS**

**Movements classification by human assessors**

In the following we describe the movements classification by human assessors in the face-visible condition. In our recent study, the 2800 snippets were annotated by two well-trained and experienced human GMA assessors. The infant faces were visible to the assessors, as in a standard general movements assessment. The assessors, independent from each other, classified each snippet as “fidgety movements present” (FM+), “fidgety movements absent” (FM-), or “not assessable” (i.e., the infant presents one or more of the following states during the specific 5 s: fussy/crying, drowsy, yawning, refluxing, over-excited, self-soothing, or distracted by the environment, all of which could distort infants’ movement pattern and shall not be assessed for GMA). These existing classification data are adopted by the current study for the face-visible condition of human assessors. To conduct new experiments in the face-blurred condition, we carried out a face blurring procedure with the original 2800 snippets.

**Face blurring procedure**

Our proposed face blurring procedure consists of two steps: (1) extraction of body key points (see Figure 2) in order to determine position of eyes and nose (Figure 3B), and (2) face masking by applying blurring filter in the area around eyes and nose (Figure 3C).

**Extraction of body key points**

To determine the position of the mask, i.e., the anatomical area around the eyes and the nose, we used a state-of-the-art pose estimation method OpenPose. The authors are aware that OpenPose is not specifically trained on infants and that models specifically trained on infants may lead to a better body pose key estimation and movement recognition accuracy. However, the main goal of the study is not to develop a highly accurate model for body key point detection and/or to achieve high movement recognition accuracy. Rather, we aimed to analyze whether information obtained from face key points significantly affect classification performances. We therefore chose for the current study also OpenPose which has been successfully used in recent studies to classify infant movements.

In this study, we used OpenPose body key points not only to determine the position of the mask but also as features for movement classification and analysis about the importance of head key points.

OpenPose is based on deep learning and extracts 25 body key points from 2D images including five head key points, i.e., eyes, nose, and ears (see Figure 2). Thus, for each frame in the video (5s x 50 frames per second [fps] = 250 frames with resolution of 1920 x 1080 pixels) we extracted 25 body key points. For each body key point, OpenPose returns three values: $x_i$ and $y_i$ coordinates in the image and a reliability score $r_i$ between 0 and 1 of each key point ($i = 1 \ldots 25$). The reliability score defines how confident the algorithm is in predicting coordinates of the key point. Five head key points 1–5 (eyes, nose, and ears) were used to determine the position of the mask, whereas key points 1–21 (excluding toes and heels) or key points 6–21 (excluding toes, heels and five head key points) were used for movement classification.

**Face blurring**

As stated above, we used five head key points 1–5 (nose, eyes and ears; see Figure 2) to determine the position of the mask for face blurring. A center point coordinates $c_x$ and $c_y$ of the elliptic mask was defined as average values of head key points:

$$c_x(f) = \text{mean}([eyL_x, eyR_x, ns_x, erL_x, erR_x]),$$

$$c_y(f) = \text{mean}([eyL_y, eyR_y, ns_y]),$$

if average reliability score $r_{avg} = \text{mean}([eyL_r, eyR_r, ns_r]) > 0.35$, otherwise

$$c_x(f) = c_x(f - 1),$$
\[ c_y(f) = c_y(f - 1), \]

where \( f = 1 \ldots 250 \) denotes the frame number, \( \text{eyL} \) – left eye, \( \text{eyR} \) – right eye, \( \text{ns} \) – nose, \( \text{erL} \) – left ear, \( \text{erR} \) – right ear. Note that for \( c_y \) we only used eye and nose key points since we wanted to mask the area around eyes and nose but leave the mouth area still visible (please see discussion).

We have also applied an exponential moving average filter to reduce jerk of the mask movement in video:

\[
\begin{align*}
    c_x(f) &= a * c_x(f - 1) + (1 - a) * c_x(f), \\
    c_y(f) &= a * c_y(f - 1) + (1 - a) * c_y(f),
\end{align*}
\]

with \( a = 0.5 \).

Finally, we applied an elliptic blurring mask with the center point \( c_x, c_y \), width \( w = 150 \) pixels and height \( h = 68 \) pixels. To generate the blurring mask, we used the normalized box filter using standard OpenCV (https://opencv.org) blurring function with kernel size \( k = 25 \). In addition, to prevent reconstruction we also applied random noise to the blurred pixels from uniform distribution between 0 and 25 (which on average corresponds to about 5% of relative noise).

**Movement classification by human assessors**

In the following we describe the movements classification by human assessors in the face-blurred condition. From the 2800 snippets, three subsets of 280 snippets each were randomly selected. Snippets in the subsets did not overlap with each other. The two experienced GMA assessors from the previous study\(^27\) independently rated the three subsets of face-blurred snippets, making the FM+, FM-, or not assessable classifications. The current rating took place 8 months after the original assessments, so that the memory effect is barely conceivable. The presenting order of the three subsets were the same for both assessors. The assessors did not receive feedback, nor did they communicate with each other throughout the assessment. The rating procedure remained identical as in the previous study, except that the assessors only saw infants with blurred faces. The classification results of the three subsets will be compared to the assessors’ original classifications (i.e., face-visible condition\(^27\)).

**Automated movement classification**

In the following we describe a machine learning approach for movements classification in the face-blurred and face-visible conditions. After applying the face blurring mask, the head key points could not be detected reliably. Thus, the question arises whether head key points are indispensable features for the specific fidgety movement classification; i.e., whether classification accuracy would decrease significantly when head key points are excluded. To answer this question, we performed classification of FM into two classes (FM+ and FM-) when using features with and without head key points. Note that the goal here is not to arrive at the network architecture with the highest classification accuracy but to compare classification performances in the two conditions.

**Features**

We used OpenPose body key points (x and y coordinates) extracted from all video frames (250 in total) as features for a neural classifier (see Figure 4). As shown in Figure 2, we used 21 key points (without toes and heels) or 16 key points (without toes and heels and without five head key points). Note that for classification experiments we extracted body key points using the original videos. We discarded toes and heels because these key points were not detected reliably using OpenPose. Thus, we constructed a feature matrix of size 250x42 or 250x32 depending on whether we used head features or not.

**Pre-processing**

In some cases, e.g., due to occlusions, OpenPose cannot detect coordinates of key points in images and these values by default are set to 0. We pre-processed not detected key points in those frames by setting 0 values to the values obtained from linear interpolation between the frames with non 0 values.

In addition, we applied min-max normalization (values between 0 and 1) to remove the influence of body size. Note that min-max normalization was performed after an interpolation step and that the remaining body key points with 0 values (e.g., at the beginning or at the end of the video) were excluded for min-max normalization to avoid effect of outliers (0 values).
Network architectures and training procedure

We used simple shallow multi-layer network architectures with one convolutional layer and one or two FC layers (see Figure 4). Again, our goal here is not to arrive at the network architecture with the highest classification accuracy but to compare performances when using features with and without head key points. For other more sophisticated network architectures see e.g.,27–29 We also performed an ablation study where we investigated network architectures with different number of neurons in the FC layers and different number of filters and different filter sizes in the convolutional layer (see Table 2 and Table 3).

For network training we used the Adam optimizer58 with the binary cross-entropy as a loss function and the batch size of 32 samples. To prevent the network from overfitting we used validation stop (1/8 of training data) where we stopped training if classification accuracy on the validation set was not improving in ten consecutive epochs. For each parameter set we performed training of the network ten times and then selected the model with the classification accuracy on the validation set which then was evaluated on the test set. Neural networks were implemented using TensorFlow (https://www.tensorflow.org/) and Keras API (https://keras.io/).

Evaluation procedure

As presented above, our dataset consisted of 1784 samples (956 FM+ and 828 FM-). For comparison of movement classification performance when using models with and without head features, we performed a 5-fold cross-validation, where each time 1/5 of the data was used for testing and the rest 4/5 of the data was used for training. Note that training data was split into training set (7/8 of the data) for parameter update and validation set (1/8 of the data) for training stop (see above). We used mean classification accuracy obtained from five test sets to analyze importance of head features for classification performance.

QUANTIFICATION AND STATISTICAL ANALYSIS

Cohen’s kappa was calculated for intra- and inter-rater reliabilities of the human assessors. To compare classification accuracies of network architectures with and without head key points, we calculated average classification accuracies across five test sets, confidence intervals of mean (CI 95%), and p values for comparison of means using two-sample t-test. Statistical significance was set at p < 0.05.