Body Pose Estimation in Depth Images for Infant Motion Analysis

Nikolas Hesse[1], A. Sebastian Schröder[2], Wolfgang Müller-Felber[2],
Christoph Bodensteiner[1], Michael Arens[1], and Ulrich G. Hofmann[3,4,5]

***Abstract*— Motion analysis of infants is used for early detection of movement disorders like cerebral palsy. For the development of automated methods, capturing the infant's pose accurately is crucial. Our system for predicting 3D joint positions is based on a recently introduced pixelwise body part classifier using random ferns, to which we propose multiple enhancements. We apply a feature selection step before training random ferns to avoid the inclusion of redundant features. We introduce a kinematic chain reweighting scheme to identify and to correct misclassified pixels, and we achieve rotation invariance by performing PCA on the input depth image. The proposed methods improve pose estimation accuracy by a large margin on multiple recordings of infants. We demonstrate the suitability of the approach for motion analysis by comparing predicted knee angles to ground truth angles.**

I. Introduction

Movement disorders like cerebral palsy (CP) can be detected at an early age. The current medical gold standard method for early detection of CP is the *General Movements Assessment* (GMA) [1], which requires a trained expert, often a doctor, to manually examine video recordings of infants to evaluate their movements. Multiple drawbacks exist for this method: it is time-consuming, it requires an expert who is repeatedly trained on the GMA, and the outcome is based on a subjective opinion.

Our goal is to automate the task of motion analysis to identify infantile motor disorders. This paper focuses on a fundamental step for motion analysis: capturing the body pose accurately and reliably. In addition, the system is required to be cheap, easy to set up, usable by non-experts and non-intrusive for the infants (see Fig. 1). We build upon an approach for infant body pose estimation in single depth images using random ferns [2]. A fern is a special kind of decision tree, where at each level of the tree, the same binary feature is evaluated. We analyze the different stages of that approach and present multiple contributions to enhance the quality of pose estimation. We integrate a feature selection step in the training procedure to avoid redundant features and show how the choice of training data improves results. We incorporate a kinematic chain reweighting scheme to identify and to correct misclassified pixels. We perform PCA on the input data to find the main body axis, and compensate for rotations. We evaluate pose estimation accuracy on multiple recordings of infants using three challenging metrics.

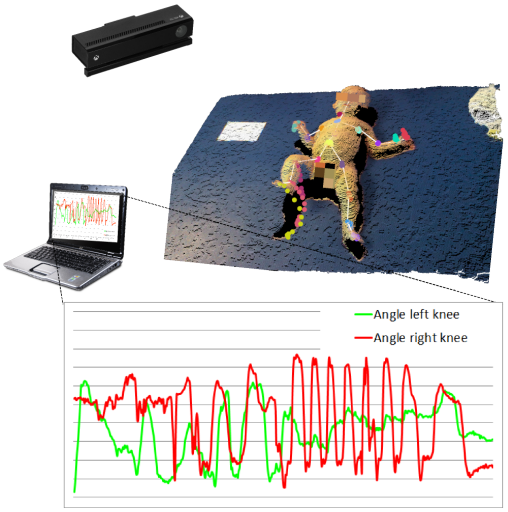


Fig. 1. System setup. A depth camera is mounted above the examination table. A connected laptop running the proposed system estimates the body pose in every frame and outputs information about movements, e.g. knee angles. Best viewed in color.

II. Related work

With the introduction of commodity depth sensors like the Microsoft Kinect and its body tracking capabilities, many applications for analyzing humans and their movements were developed. However, the Kinect body tracking is limited to persons taller than 1 m. Multiple systems aim to automate *infant* motion analysis, in order to produce objective and repeatable measures for assessing movements. Most systems need either markers or sensors that are attached to the infant [3], [4], [5] or they lack 3D information [6], [7]. Other approaches overcome these limitations by fitting a simplified body model to the whole body [8] or lower limbs [9] of infants captured by RGB-D devices. We focus on a recently introduced approach for infant body pose estimation in single depth images using random ferns [2] in the following section.

III. Methods

This section presents methods for improving infant body pose estimation using random ferns [2]. The ferns, which are a variant of decision trees, are trained to assign a body part class to each input depth pixel. When traversing the fern, input depth is compared to depth at an offset pixel in each split node (feature). If the difference is larger than a given

[1]Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany, nikolas.hesse@iosb.fraunhofer.de
[2]Department of Paediatric Neurology and Developmental Medicine, Dr. von Hauner Children's Hospital, Ludwig-Maximilians-Universität (LMU), Munich, Germany
[3]Section for Neuroelectronic Systems, Neurosurgery, Medical Center - University of Freiburg, Germany
[4]Faculty of Medicine, University of Freiburg, Germany
[5]Freiburg Institute for Advanced Studies (FRIAS), University of Freiburg

threshold, the feature output is 1, else 0. Leaf nodes contain a probability distribution over body part classes that is learned from synthetic training data. 3D joint positions are calculated based on estimated body part regions (see [2] for details).

A. Feature selection

When analyzing trained ferns from [2], we found that binary depth comparison features are included for which pixel offsets ϕ or depth thresholds τ take values that lead to an evaluation of all training pixels to exclusively one side (all 0 or all 1). Leaf nodes which are on the wrong side of that particular feature will never be traversed during training, and therefore do not contribute to classification. Simply removing these features would reduce memory requirements, but to improve classification, replacement by new (better) features is desirable. Replacing features in an existing fern enforces recalculation of probability distributions in all leaf nodes, since, due to the properties of ferns, each feature influences all leaf nodes. Therefore, we filter out redundant features *before* training to keep training times low and to avoid re-processing. We randomly generate a large set of features and evaluate them on the training data once, using information gain measure (Fig. 2). Only features with information gain above a user-specified threshold are added to a set from which features are drawn randomly during training. The information gain measure is often used for evaluating candidate features in training random decision trees (e.g. [10]). It is defined by

$$ig(\theta) = \sum_{d \in \{L,R\}} \frac{|S^d(\theta)|}{|S|} I(S^d(\theta)), \tag{1}$$

where θ represents the feature parameters, consisting of pixel offset $\phi = (\phi_x, \phi_y)$ and depth threshold τ. S is the set of all training pixels, S^L and S^R are the left and right subsets according to the evaluation using θ. I is the Shannon entropy of the distribution of body part classes corresponding to pixels in S:

$$I(S) = -\sum_c p(c|S) \log p(c|S), \tag{2}$$

where $p(c|S)$ is the normalized histogram of the set of body part classes $c(u)$ for all $u \in S$.

B. Training data generation

Evaluations in [2] showed failure cases for wrongly classified body parts (e.g. pixels on knee classified as hand). The poses used in their training are taken from a data set that contains adults performing everyday activities. These poses are mapped to a synthetic body model of an infant. However, these poses are not typical for infants and their specific motions. This bias in training data will severly decrease the classification accuracy of the ferns. Due to the infeasibility of recording a large amount of motion capture data from infants performing the complete range of possible poses, we have synthetically generated a wide range of what we consider baby-like poses. We determine angle ranges for the extremities which represent infant poses realistically according to visual inspection. We generate 30,000 poses by randomly combining angles within these ranges. We pick small random angles between 10 and -10 degrees for joints that do not belong to extremities. We generate labeled depth images from three different frontal viewpoints for each of the poses. To ensure that the training data is not biased towards either side, we mirror the generated data, resulting in an overall number of 180k training images.

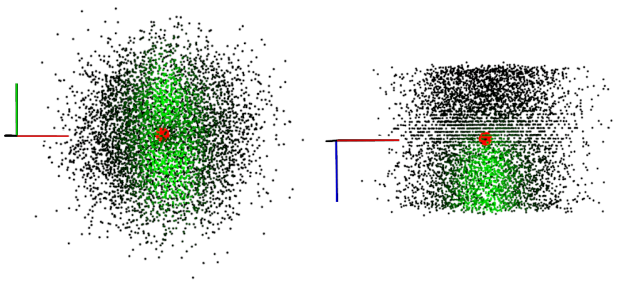


Fig. 2. 5000 randomly generated features, evaluated on infant training data. Green depicts large information gain, black small gain. Pixel offsets are plotted on x (red) and y (green) axis, depth threshold on z (blue) axis. Left: view on XY-axes. Right: view on XZ-axes. Red dot is input pixel which the offsets are related to. Best viewed in color.

C. Kinematic chain reweighting and filtering

In certain scenarios body part regions are misclassified, e.g. estimated hand regions on the knees, which are easily identifiable for humans using prior knowledge about the human body. However, random ferns independently classify each pixel w.r.t the classified outcome of neighboring pixel classes. Therefore, we add a post-processing step incorporating prior information about valid connections between human body parts. After predicting body part labels for all pixels, we form clusters of neighboring pixels that share the same label. We construct a graph, adding one node for each cluster and connecting neighboring clusters by edges. Edge weights correspond to the number of bones between body parts in the underlying skeleton. Pixels of a cluster are considered misclassifications and filtered out if the shortest path from the cluster to the root node (body center) violates kinematic chain constraints. Body part probabilites for misclassified pixels are reweighted depending on surrounding (correct) body parts. We further enforce the existence of only one cluster per body part class which leads to body part labels that conform to the kinematic chain of the body.

D. Rotation invariance

The binary depth features used in the ferns are not invariant to rotations. Instead of trying to capture all possible variations of rotation in the training data, we train on upright positions with limited range of rotation. In our setting, the babies are always filmed from above, so that the main body axis is displayed vertically in the camera image. If the main axis strongly diverges from being vertical, the prediction will be distorted. We use the 2D positions of pixels from estimated body parts that belong to the torso from the previous frame as input for principal component analysis

Fig. 3. Effect of rotation correction using PCA. Left: without correction. Right: with correction. Colors represent different body parts. Lines display connections between predicted joints (colored dots). Best viewed in color.

(PCA). We rotate the feature offsets to conform to the axes we get from the first two eigenvectors of PCA. This way, the classification is performed as if an upright body was given as input (see Fig. 3).

IV. Evaluation

We quantitatively compare our methods to [2] on three sequences of different infants moving freely in supine position without external stimuli. Recordings were only made if parents gave their written consent. From two of the sequences, we select sections of 500 frames where the infants are most active, whereas the third sequence consists of 4500 frames. A Microsoft Kinect was mounted to the wall facing downward about 1 meter above an examination table, so that the recorded infant is facing the camera frontally.

Manual annotation of ground truth 3D joint positions is a cumbersome, yet inaccurate process. We fit a 3D body model to the recorded sequences and visually verify the plausibility and accuracy of results and consider them ground truth for our evaluation. The background is removed from the depth images prior to evaluation so that only pixels remain that are part of the infant.

A. Error metrics

Pose estimation approaches are often evaluated by indicating the *average joint position error* (AJPE) [2]. If used as a foundation for motion analysis, though, we need more strict evaluation measures. Therefore, we use the *worst-case accuracy* (WCA) as proposed by [11], which is the percentage of frames in which *all* joints lie within a certain distance from the ground truth. Additionally, we introduce a measure that we call the *jitter accuracy* (JA). We define the difference of predicted joint position deviation relative to ground truth in consecutive frames (*jitter error*) by

$$je_{i,j} = \|(x_{i,j} - gt_{i,j}) - (x_{i-1,j} - gt_{i-1,j})\|, \qquad (3)$$

where $i = 2, \ldots, N$ is the frame number, N the total number of frames, j the joint index, $x_{i,j}$ the predicted position of joint j in frame i, $gt_{i,j}$ the ground truth position of joint j in frame i and $\|\cdot\|$ the euclidean distance. The jitter accuracy is the percentage of frames in which the jitter error of *all* joints is smaller than a certain threshold.

	Without RWF	With RWF
[2]	1.782 (3.103)	1.382 (1.459)
Ours	1.212 (0.931)	1.224 (0.897)
Ours (FS)	1.236 (1.121)	1.222 (0.852)

TABLE I

Average joint position error (and standard deviation) in cm over all sequences.

B. Results

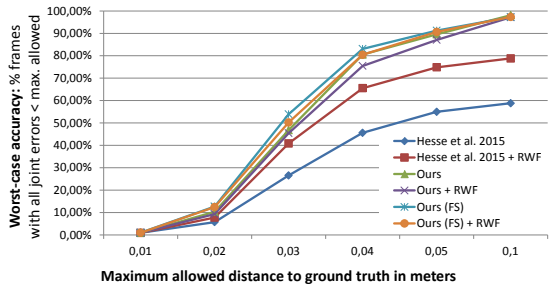
We compare the proposed methods to the baseline approach [2]. All classifiers are trained using the same number of ferns (15), fern depth (13) and pixel offset neighborhood radius (20 cm). PCA rotation correction is applied with all methods. *FS* indicates that the feature selection step is used prior to training, *RWF* means that kinematic chain reweighting and filtering is applied. For results without RWF, the filtering procedure from [2] is applied.

Average joint position error (Table I). AJPE of [2] is reduced by 0.4 cm (22%) by applying RWF. When using better training data (*Ours*) the improvement is even bigger. Yet, no further gain is obtained by combining *Ours* and RWF. We find the explanation that *Ours* provides much cleaner estimates, especially on frames where [2] shows huge errors, e.g. classifying knee as hand. RWF fixes that same kind of error - hence there is no improvement when combined with *Ours*. Error even increases slightly for RWF with *Ours*, which we believe is due to the fact that when class labels of body part patches are changed by RWF, predicted positions seem to shift too far sometimes.

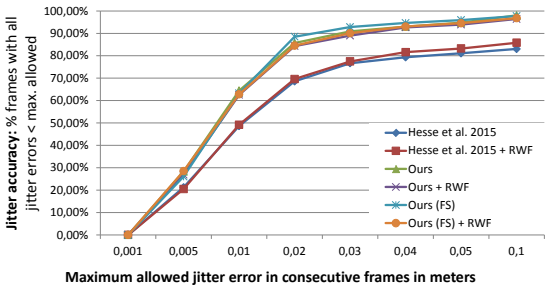
Worst-case accuracy (Fig. 4a). A similar trend is visible for the worst-case accuracy. Our proposed methods clearly outperform [2] by e.g. reaching 90% of all frames with all joint errors smaller than 5 cm, whereas [2] reaches 55%. A joint position error of 5 cm exceeds tolerable error limits for infant motion analysis, but we want the reader to keep in mind that the worst-case accuracy does not distinguish between *all* joint errors being larger than the threshold and just one. If we allowed one joint error larger than the threshold, our best WCA for a maximum distance of *3 cm* is 85% ([2]: 55%).

Jitter accuracy (Fig. 4b). There is no big gain in jitter accuracy by applying RWF, but *Ours* significantly reduces jitter error in comparison with [2]. At a maximum jitter error of 2 cm, there is a gain of nearly 20% in accuracy. The jitter accuracy is an important measure when using pose estimation for motion analysis, since jumps of the joint positions in consecutive frames will be erroneously regarded as movements.

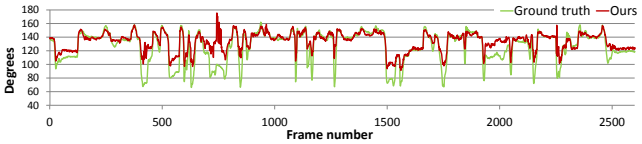
In the conducted experiments, the feature selection step does not lead to a significant improvement in accuracy. If there are no redundant features included in training without feature selection, there will be no benefit from this step. We still believe feature selection to be a valid enhancement, because it prevents the inclusion of redundant features independent of the amount of randomness used during training.



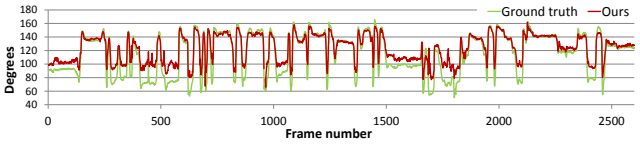
(a) Worst-case accuracy. Average over all sequences.



(b) Jitter accuracy. Average over all sequences.



(c) Left knee angles for 2600 frames of one sequence.



(d) Right knee angles for 2600 frames of one sequence..

Fig. 4. Evaluation results. Best viewed in color.

Altogether, we show that the proposed methods outperform [2] by large margin on multiple challenging error metrics.

V. Infant motion analysis

In Fig. 4c and 4d, we illustrate how our system captures movement information like joint angles accurately by comparison with angles calculated from ground truth joint positions. Results are presented for left and right knee for 2600 frames of the longest sequence. Although the angle values do not match the peak levels exactly, we observe that the estimated angles reflect the ground truth very well. Doctors will be able to get an impression of the movement quality by one glance on the plotted angles. It will, e.g., be clearly visible if there is an absence of motion on one side of the body. Motion features based on angles have been shown to successfully detect and predict cerebral palsy [3]. Others have employed features based on trajectories, velocities and accelerations [4], [5], [6] which our system can measure as well. In a subsequent step, we will study early detection of CP using different features.

VI. Conclusion

We presented multiple enhancements to infant body pose estimation using random ferns. We propose a feature selection step before training to filter out irrelevant features. We introduce a kinematic chain reweighting scheme in a post-processing stage that identifies and corrects misclassified pixels that do not obey human skeleton constraints. Furthermore, we show the importance of training data representing test data correctly. Our methods outperform a recently introduced approach by a large margin in terms of average joint position error, worst-case accuracy and newly introduced metric jitter accuracy. Finally, we illustrate the accuracy of extracted angles that can serve as a basis for motion features for the automatic detection of motor disorders.

References

[1] C. Einspieler, H. F. Prechtl, F. Ferrari, G. Cioni, and A. F. Bos, "The qualitative assessment of general movements in preterm, term and young infants – review of the methodology," *Early human development*, vol. 50, no. 1, pp. 47–60, 1997.

[2] N. Hesse, G. Stachowiak, T. Breuer, and M. Arens, "Estimating body pose of infants in depth images using random ferns," in *IEEE International Conference on Computer Vision Workshops*, 2015, pp. 35–43.

[3] D. Karch, K.-S. Kang, K. Wochner, H. Philippi, M. Hadders-Algra, J. Pietz, and H. Dickhaus, "Kinematic assessment of stereotypy in spontaneous movements in infants," *Gait & posture*, vol. 36, no. 2, pp. 307–311, 2012.

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

[5] F. Heinze, K. Hesels, N. Breitbach-Faller, T. Schmitz-Rode, and C. Disselhorst-Klug, "Movement analysis by accelerometry of newborns and infants for the early detection of movement disorders due to infantile cerebral palsy," *Medical & biological engineering & computing*, vol. 48, no. 8, pp. 765–772, 2010.

[6] H. Rahmati, O. M. Aamo, O. Stavdahl, R. Dragon, and L. Adde, "Video-based early cerebral palsy prediction using motion segmentation," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2014, pp. 3779–3783.

[7] A. Stahl, C. Schellewald, Ø. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerød, "An optical flow-based method to predict infantile cerebral palsy," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 4, pp. 605–614, 2012.

[8] M. D. Olsen, A. Herskind, J. B. Nielsen, and R. R. Paulsen, "Model-based motion tracking of infants," in *European Conference on Computer Vision Workshops*, 2014, pp. 673–685.

[9] M. M. Serrano, Y.-P. Chen, A. Howard, and P. A. Vela, "Lower limb pose estimation for monitoring the kicking patterns of infants," in *IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 2157–2160.

[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1297–1304.

[11] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 103–110.