3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism

Elisabeta Marinoiu[2*] Mihai Zanfir[2*] Vlad Olaru[2] Cristian Sminchisescu[1,2]
{elisabeta.marinoiu, mihai.zanfir, vlad.olaru}@imar.ro cristian.sminchisescu@math.lth.se
[1]Department of Mathematics, Faculty of Engineering, Lund University
[2]Institute of Mathematics of the Romanian Academy

Abstract

We introduce new, fine-grained action and emotion recognition tasks defined on non-staged videos, recorded during robot-assisted therapy sessions of children with autism. The tasks present several challenges: a large dataset with long videos, a large number of highly variable actions, children that are only partially visible, have different ages and may show unpredictable behaviour, as well as non-standard camera viewpoints. We investigate how state-of-the-art 3d human pose reconstruction methods perform on the newly introduced tasks and propose extensions to adapt them to deal with these challenges. We also analyze multiple approaches in action and emotion recognition from 3d human pose data, establish several baselines, and discuss results and their implications in the context of child-robot interaction.

1. Introduction

Autism affects the lives of millions of people around the world. It is estimated that 1 out 100 people in Europe suffers from autism [1], whereas the Centers for Disease Control and Prevention estimates that 1 in 68 children in the US has autism, with a prevalence of male cases over female that amounts to a factor of 4.5 times higher [2]. The challenges the people with autism face when interacting with others revolve around confusion, fear or basic misunderstanding of emotions and affects. They have difficulties using and understanding verbal and non-verbal communication, recognizing and properly reacting to other people's feelings, and fail to respond, either verbally or non-verbally, to social and emotional signs coming from others.

In contrast, persons with autism cope well with rule-based, predictable systems such as computers [12, 29, 24]. Recent developments have shown the advantages of using humanoid robots for psycho-educational therapy, as children with autism feel more comfortable around such robots than in the presence of humans, who may be perceived as hard to understand and sometimes even frightening. While humanoid robots capable of facial expressions could help improve the ability of children with autism to recognize other people's emotions, most studies are based on remote controlled human-robot interaction (HRI). Less work has been done to automatically track and detect children' facial expressions, body pose and gestures, or vocal behavior in order to properly assess and react to the their behavior, as recorded by robot cameras in unconstrained scenes. Thus, robot-assisted therapy cannot yet be used for emotion recognition and, subsequently, to enable appropriate responses to such emotions.

In this paper, we introduce fine-grained action classification and emotion prediction tasks defined on non-staged videos, recorded during robot-assisted therapy sessions of children with autism. The data is designed to support robust, context-sensitive, multi-modal and naturalistic HRI solutions for enhancing the social imagination skills of such children. Our contributions can be summarized as follows:

- We analyze a large scale video dataset containing child-therapist interactions and subtle behavioral annotations. The dataset is challenging for its long videos, large number of action and emotion (valence-arousal) annotations, difficult viewpoints, partial views, and occlusions between child and therapist.

- We adapt state-of-the-art 3d human pose estimation models to this setting, making it possible to reliably track and reconstruct both the child and the therapist, from RGB data, at comparable performance levels with an industrial-grade Kinect system. This is desirable as our proposed models offer not just 3d human pose reconstructions, but additionally detailed human body part segmentation information which can be effective, in the long run, in precisely capturing complex interactions or subtle behavior.

*Authors contributed equally

- We establish several action and emotion recognition baselines, including systems based on child representations, and models that jointly capture the child and the therapist. The data, annotations and recognition models are made available online at `http://vision.imar.ro/de-enigma`.

2. Related Work

Despite their social challenges, people with autism have rather normal – sometimes above normal – capabilities of interacting with predictable systems such as computers [12, 29, 24]. In recent years, the interaction approaches based on humanoid robots such as Nao [5] or Zeno R25 [4] multiplied significantly. Such robots are sometimes preferred to humans because they are more comfortable to interact with in terms of predictability, behavior complexity and perceived threat. Not surprisingly, the human-like look of these robots is a further incentive for their use over screen-based computing technology [11, 27, 10].

Humanoid robots have been used beyond enhancing learning methodologies for children with autism. For instance, in order to explore the capacity of such children to develop the ability to recognize the emotions of other people, a robot called Milo portrayed various emotions – e.g. happiness, sadness, anger, fear – through facial expressions, while the child selected the appropriate emotion using a tablet-based multiple choice interface [3]. Another study aimed at comparing the emotion expression recognition abilities of children with autism with those of typically developing children [31] has shown that, by using gestures to convey emotional expressions by a humanoid robot (Zeno) in a social skill therapy setting, it can significantly impact the prediction accuracy of expressing emotion. Other studies [34, 35] evaluated the benefits of using a humanoid robot (KASPAR) to engage children with autism into imitative, collaborative game playing.

Although a person's facial expressions is the main focus in emotion understanding [25, 17], the body language expressed through pose offers complementary information. [18] investigated the role of body movement and posture in expressing emotion as complementary to facial expressions and discussed their importance in the context of embodied conversational agents. Here, we present an automated approach for continuous emotion recognition in the valence-arousal space, using only 3d skeleton data. Although considerable steps have been taken in automatically detecting, classifying and interpreting human action from body pose features [20, 16, 13, 9, 37], many approaches rely on RGB-D sensors such as Kinect [33] to estimate 3d human pose, with datasets recorded in a controlled setup, where actions are a-priori defined. Recent advances in 2d and 3d pose estimation [7, 28, 21, 6, 22, 26, 38] can potentially offer an alternative to depth sensors by providing reliable pose estimates from only RGB data. Still, such methods have not yet been tested in the context of a highly challenging, real world, action classification problem.

Many complementary human sensing datasets are available for both pose and action recognition. Here we focus on a very different problem domain – autism therapy –, with unique challenges, from permission to data release to the therapeutic setup, fine-grained action and emotion annotations, the complexity of viewing angles and interactions, as well as data large-scale. Rehg et al. [30] also proposed a dataset of children interacting with parents and therapists, but focused on understanding the behaviour of infants in order to potentially help with early diagnosis. Their different approach is to analyze the engagement level by detecting smile, gaze and a fixed set of objects, relying on finding specific phrases mentioned by the therapist to help segment a video into predefined stages. Our approach is complementary: we deal with older children that have already been diagnosed and undertake robot assisted therapy in a less constrained environment, and focus on understanding body gestures aiming at technological development personalized for children needs.

3. DE-ENIGMA Action Annotation Setup

The DE-ENIGMA [32] dataset[1] contains multi-modal recordings of therapy sessions of children with autism. The sessions are either therapist-only or robot-assisted; the former are captured for control purposes, while the latter are those of interest for this paper. In robot-assisted sessions a child and a therapist sit in front of a table on which a robot is placed. The therapist remotely controls the robot and uses it to engage the child in the process of learning emotions. The sessions consist of a 'free-play' part (where the child plays with toys of his choice), and an actual therapy part. The therapy is based on scenarios in which the therapist shows cards depicting various emotions (happy, sad, angry, etc.) which are also reproduced by the robot, and the child must match the emotions to those performed. The cards are either in the therapist's hand or lie on the same table as the robot, then the child has to pick up the one of his choice.

In this paper, we consider only the RGB + depth modalities recorded using a Kinect v2 camera (at 30 FPS) placed right above the robot head, towards the child (see fig. 1). The child is facing the camera frontally, but due to the constraints in robot positioning and recording cameras, most of the time only the upper body is visible. The therapist is also placed in front of the table, but she usually faces the child and the camera observes a side view of her. Most of the time the therapist is severely occluded, with only half of the upper body and arms visible. An illustration of the setup is given in fig. 1.

[1] Available online at : `http://de-enigma.eu/resources/the-de-enigma-database/`

Figure 1. Experimental setup, constraints and recording challenges. The leftmost picture shows the constraints imposed by the table and robot placement as well as the position and tilt of the camera. First, the camera must be placed behind the robot to avoid interfering with the therapy. Second, the field of view must avoid the robot, so the camera has to be lifted up to a certain height above the robot, or placed laterally. Finally, the positions of the stools of the therapist and child, who need to sit close to the table to use the cards (see second picture), together with the robot height contribute to the final adjustment of the camera's height and orientation angle. The second picture shows standard recording conditions, in which, inevitably because of the table, only partial views of the therapist and child are available (their legs are occluded by the table). The other two pictures show various challenging situations that appeared during the recordings: the therapist and child interaction results in occlusion (third picture); the child and the therapist get out of the field of view (fourth picture).

Recordings. A selection of recordings from multiple therapy sessions of 7 children was annotated with 37 action classes. 19 classes describing the therapist's actions were also annotated, but have not been used in the analysis of this paper. The children selection covers a variety of gestures and interactions for typical therapy sessions.

Annotation procedure. The therapy scenarios cover a wide variety of body gestures and actions performed by children (see table 1). We have annotated a total of 3757 sequences, with an average duration of 2.1 seconds. The annotation of therapy videos relies on an extensive web-based tool developed by us that can (i) select temporal extents and (ii) assign them a class label. Features that improve the annotation experience such as shortcuts for precise temporal adjustments, current selection replays, previous annotations filtering and visualization, or user session management, are also included.

Details	All annotations	Working subset
No. of sequences	3757	2031
No. of classes (child)	37	24
No. of classes (therapist)	19	0
No. of subjects	7	7
No. of therapy sessions	24	24
Coverage	38%	23%
Total length of annotations	132.1min	74.4 min
Average sequence length	2.1s	2.1s
No. of interacting sequences	1861	749

Table 1. Details of the annotated dataset. The experiments use 2,031 annotated videos describing children body movements and behaviour. A large part of these sequences (749) describes actions performed by children in response/collaboration to the therapist. The annotated sequences in our working subset cover, on average, 23% of the therapy sessions.

The dataset was annotated by 4 people, each receiving videos from the therapy sessions of at least 3 children. To eliminate possible mistakes, each annotator's work was verified by the other annotators. An initial set of originally proposed actions has been extended by the annotators with repetitive actions of a particular child, e.g., one of the children repetitively touched his chest with his hands. The experiments presented in this paper use a subset of 2031 annotated sequences spanning over 24 classes common to all children. Even if the selected classes refer to children behavior, some of them relate to the therapist, e.g., *Pointing to therapist*, *Turning towards therapist*. We refer to those as interacting sequences. Among the annotated sequences, around a third (749 out of $2,031$) are interacting sequences. Table 1 contains statistics of the annotations, while examples of the annotated classes are shown in fig. 2.

The annotated action classes are heavily imbalanced, as shown in fig. 3. The children behave quite differently in the number of annotated sequences, some being considerably less active than others, see fig. 4. Significant differences are observed between the sequences with the same class label, as shown in fig. 5. These variations arise naturally in non staged videos and are part of the dataset challenge.

4. Skeleton Reconstruction from RGB data

Our long term goal is to automatically interpret and react to a child's actions in the challenging setting of a therapy session. In order to understand the child, we rely on high-level features associated to her/his 3d pose and shape. In this section we review several state-of-the-art 3d pose estimation methods, discuss their shortcomings and show how to adapt them to our particular setup.

The task of 3d pose estimation is defined as a function from an input image, $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$, to body joint coordinates $\mathbf{J} \in \mathbb{R}^{N \times 3}$. Different systems may consider slightly different kinematic tree configurations, but a common set includes the head, neck, shoulders, elbow, wrists, hips, knees and ankles.

DMHS [28] is a multitask deep neural network that estimates both the 2d and 3d joint positions and the semantic

Figure 2. Examples of annotations we provide. Some of the actions are defined in relation to the therapist (*High-five*, *Grab card from therapist*) or the robot (*Point to Robot*), while other describe the child independently (*Clap hands*, *Wave*, etc).

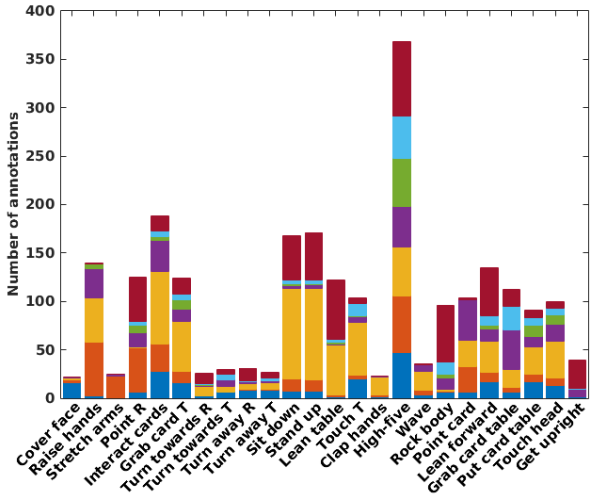

Figure 3. Action distribution in the dataset. Each bar-color corresponds to one child. For brevity, on the x axis R stands for robot and T for therapist, e.g. the label *Point R* refers to the action *Point to the Robot*. Note highly imbalanced distributions of annotated sequences per class and uneven action distribution across children. Some classes, e.g. *Touch therapist*, *Point to Robot*, exhibit considerably different number of annotations across children.

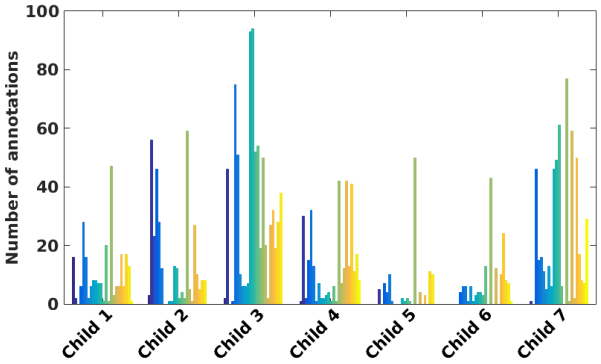

Figure 4. Annotation distribution per child (classes are color-coded). Both the total number of annotations and the class distributions are heavily imbalanced (e.g. children 3 and 7 have considerably more annotations than 5 and 6). This indicates how much they gesticulate and how responsive/engaged they were during therapy.

human body part labeling of the person. DMHS is trained on fully visible humans from Human80K [14], a subset of [15], which contains data for 11 adult actors performing 15 different actions in a laboratory setup. This makes it non-straightforward to use for partially visible children.

DMHS Adaptation to Partially Visible People. To improve the DMHS-based 3d pose estimation of partially visible people, we collect statistics of those human keypoint configurations that are frequently visible in natural images. We use images from the COCO Training [19] (Keypoints Challenge) that are annotated with 2d joints and select the 50 most frequent configurations to create a new dataset (H80KPartial) based on Human80k with a similar distribution of partial configurations as collected from COCO. For each image in Human80K, we sample a configuration following the COCO distribution and crop the image to show only the joints visible in the selected configuration.

Next, we fine-tune the semantic segmentation on H80KPartial and use it as an initialization in refining the 3d pose estimation task of the network. We adapt the semantic segmentation to partially visible people as we follow the original training procedure, in which the 3d task uses feedback from the semantic task. We test the original DMHS method and our fine-tuned adaptation for partial views (DMHSPV) on both Human80k and H80KPartial datasets. Table 2 shows results for the semantic human body part labeling task. For Human80K, the accuracy of both DMHS and DMHSPV methods is similar (with slightly better results for DMHSPV). However, the tests on H80KPartial reveal considerably improved accuracy of our fined-tuned variant of DMHS over the original (from 59.6%

Figure 5. Gesture variations between children. In the first row we show different children doing a high-five: they can use either hand and their posture varies significantly. The same can be observed in the second row, where we show different children pointing to the robot. They vary in how they perform the gesture (with one or two hands), how close they get to the robot, and in how their body is oriented.

to 78.0%). This shows success in extending the network's capabilities for partially visible humans, while preserving its accuracy for fully visible ones. The same increase in accuracy is perceptually visible when testing the two networks on images from the DE-ENIGMA dataset, as illustrated in fig. 6. Note that even for severely occluded children, DMHSPV provides plausible 3d pose estimates.

Method	H80KPartial	Human80K
DMHS	59.6%	79.0%
DMHSPV	78.0%	79.9%

Table 2. Accuracy of semantic human body part labeling for the original DMHS and our fine tuned version, DMHSPV, for both full and partially visible human poses from Human80K.

Method	H80KPartial	Human80K
DMHS	79.6 mm	63.3 mm
DMHSPV	57.6 mm	63.9 mm

Table 3. Mean per joint position error (mm) for the original DMHS and the fine tuned version, DMHSPV, for both fully and partially visible human poses from Human80K. The H80KPartial error is computed over visible joints only.

Parametric Human Model Inference. We rely on a feedforward-feedback model presented in our accompanying paper [36] to combine human detection, 2d and 3d pose prediction from DMHSPV with a shape-based volumetric refinement based on a SMPL body representation [21].

Following [36], we first transfer the pose appearance from DMHSPV to SMPL, then use this configuration as initialization for semantic image fitting. We experiment with both single and multiple frame inference where additional temporal smoothing constraints (constant velocity assumptions for 3d joints in camera space) are considered, as in [36]. The temporal inference runs in windows of 15 frames for both the therapist and the child – see fig. 7 for results.

5. Skeleton-based Action Classification

We experiment with several skeleton-based action recognition models and perform ablation studies with different types of 2d and 3d human body reconstructions. We use a cross-validation setting on children where we consider only the upper-body joints of the human skeleton.

2d Pose Features. Recent methods for 2d pose estimation [7, 28] have both good accuracy and speed. However, using just the 2d body joints locations for interpreting a child's actions might be insufficient, as the depth information could be crucial in the disambiguation of different actions. Nonetheless, we also test the output of a state-of-the-art 2d pose estimator in the context of action recognition.

3d Pose Features. We consider the 3d human skeletons obtained from DMHSPV, the single frame SMPL model inference, DMHS-SMPL-F, and the temporally smoothed inference, DMHS-SMPL-T.

Interaction Modeling. Since almost a third of the annotations involve forms of child-therapist interaction, either the child's response to a therapist initiated action (*High-five*, *Point to card*) or defined in relation to the therapist (*Turn away from therapist*, *Point to therapist*, etc.), considering the therapist representation could help discriminating between different labels. Each proposed pose feature and model pair has its own integration method of the therapist pose reconstruction.

Moving Pose. One of our baselines is based on the framework [37] which uses a KNN classifier on frame-level 3d pose descriptors. The descriptors are built by concatenating the 3d skeleton positions with the velocity and acceleration of the joints computed over a 5-frame window. In our work, we consider only the static pose and the velocity component, as in our tests the use of acceleration components did not improve action classification significantly. For 3d pose

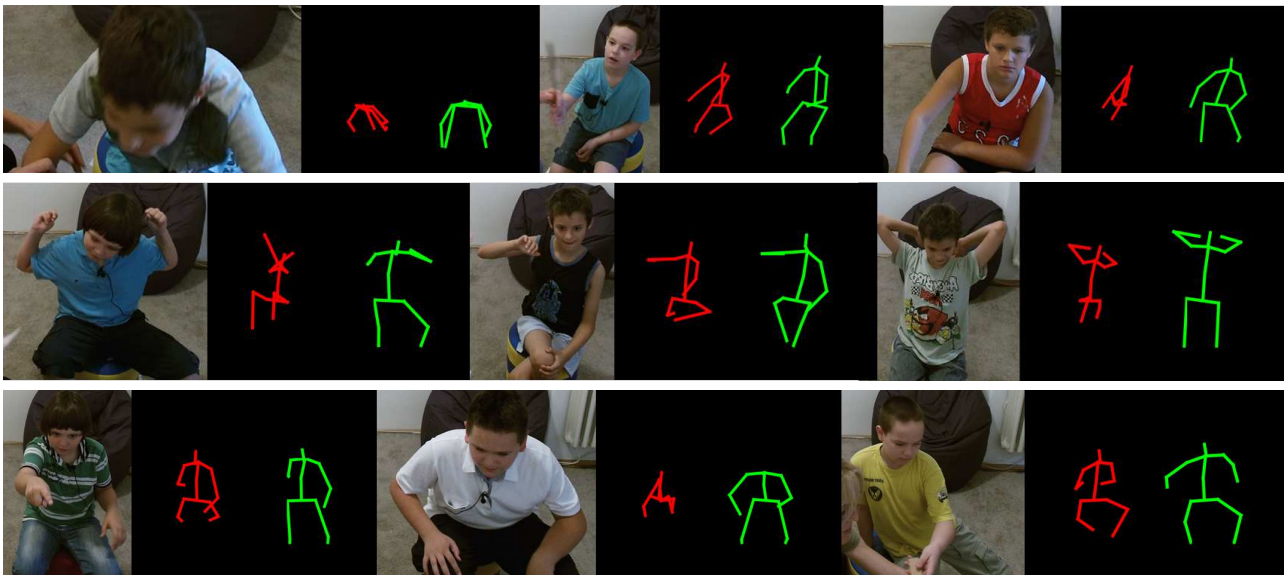

Figure 6. Importance of fine tuning for partially visible poses. First column shows the child's bounding box. With red, we show the 3d pose obtained with the original DMHS method, and with green the one obtained with our method adapted to partially visible persons, DMHSPV.

features, we follow the procedure described in [37]. The 3d human skeleton is transformed to a fixed body size and the pose is centered in the root joint (*i.e*. the hip center joint) to ensure translation invariance. We also adapt the procedure to 2d pose features. In this case, since the 2d joint locations are defined in image space, we unit-normalize and center the pose at the root defined as the center of mass of all visible 2d joints in a frame. The final per-frame descriptor is $X_c = [P_c, \alpha\delta P_c]$, where P_c is the pose of the child and δP_c is the velocity computed over a 3-frame window.

Including the pose features of the therapist in this model is straightforward. We first apply the same body normalization to the therapist and center at the root joint of the child. One extension to represent both the child and the therapist can be: $X_{ct} = [P_c, \alpha\delta P_c, \alpha' P_t, \alpha''\delta P_t]$, where P_t is the therapist pose and $\alpha, \alpha', \alpha''$ are cross-validated weights.

Table 4 shows detailed results for different 2d and 3d human pose estimations methods. Comparing the action classification accuracy, when using only the child's pose versus when we also consider the therapist, shows in all tests, that modeling interactions increases accuracy. Also note that results for 3d pose features (*i.e*. DMHSPV, DMHS-SMPL-F and DMHS-SMPL-T), estimated using only RGB data, are at par to those obtained using depth (*i.e*. Kinect).

Convolutional Neural Networks. Our convolutional neural network baseline for action classification [8, 16] uses a lightweight network architecture Conv(3x3)-ReLU-Conv(3x3)-ReLU-Pool(2x2)-Conv(3x3)-ReLU-Conv(3x3)-Pool(2x2)-Dropout-FC-FC, that takes as input a time sequence of raw 3d skeleton configurations resized to a fixed temporal length. For this study, we consider the 3d pose features of both the child and the therapist. To avoid overfitting, we add random rotations ($\pm 15^\circ$ around the Y axis) to each training sequence.

Pose Feature	MP - Child	MP - Child + Therapist
Kinect [33]	46.96%	**47.49%**
DMHSPV	32.92%	34.95%
2D [7]	40.83%	44.14
DMHS-SMPL-F	43.53%	45.07%
DMHS-SMPL-T	44.20%	45.68%

Table 4. Comparative results for different pose estimation methods for action classification when using the Moving Pose framework. We also investigate the impact of modeling the therapist in the classification accuracy.

We feed the CNN with the 3d skeleton features obtained with Kinect, as well as DMHS-SMPL-T, which was the best performing RGB model in the Moving Pose framework. We obtain improved performance compared to the Moving Pose, 53.1% accuracy using 3d pose features from Kinect and 47.9% accuracy using similar estimates from DMHS-SMPL-T.

Recurrent Neural Networks. We also establish a hierarchical bidirectional recurrent network baseline, HBRNN [9], previously shown to perform well in standard skeleton-based action-classification datasets. This model consists of a hierarchy of 5 bidirectional recurrent networks, each receiving as input the joints of 5 skeleton sub-components: torso, left arm, right arm, left leg and right leg. In subsequent layers, the representations learned by sub-component

Figure 7. Examples of 2d and 3d pose reconstructions on the annotated dataset. From left to right: 2d joint position estimates, 3d pose estimation obtained using DMHSPV, projection of the inferred shape model overlaid on the original image and inferred 3d shape model. Note that the 3d estimates from DMHPV are centered in the hip joint and we only show them with a different translation for visualization purposes. On the other hand, the models inferred in the fourth column are shown with their inferred translation.

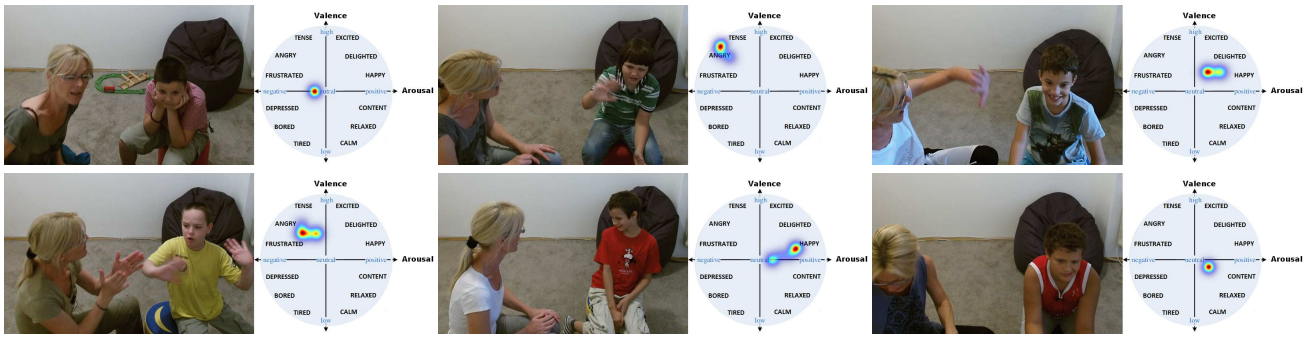

Figure 8. Examples of continuous emotion annotations made by one of the five specialized therapists. We show one frame (left) and the associated distribution over a 40 frames window (right). The children in our database exhibit a wide range of body poses and emotions.

nets are hierarchically fused and fed as inputs to upper layers. Since in our case the legs of the therapist and child are occluded, we do not use them, and use only the joint subset corresponding to the torso, left and right arm. We also add the joints from the left and right arms of the therapist as the 4th and 5th components, respectively. We test the network using Kinect 3d pose estimates and the ones inferred by DMHS-SMPL-T. The action classification accuracy is 37.8% for Kinect and 36.2% for DMHS-SMPL-T.

6. Continuous Emotion Prediction

A video selection from [32], including those 7 children used for action classification experiments, was also annotated with continuous emotions in a valence-arousal space by 5 specialized therapists. The valence axis specifies whether the emotion is positive or negative, whereas arousal controls its intensity. Fig. 8 shows examples of emotions in the valence-arousal space for our children. Representing emotions in a continuous space allows to capture more subtle affect states than using a few categorical emotion classes. Previous work [17, 23] on automatic valence-arousal prediction focused on using facial features to capture emotions. Here we propose a complementary approach centered on 3d body features to automatically predict continuous emotional states. We pre-process the data as in [25] to obtain per frame values for each annotator and align them to obtain a reliable ground-truth valence/arousal signal.

We use a personalized evaluation protocol for the 7 children. Each child's individual sessions are split into train/test in a a leave-one-session-out procedure and we report mean results for all children. The evaluation metrics are the standard ones in the literature [25, 17]: root-mean-square error ($RMSE$), Pearson product-moment correlation coefficient (PCC) and sign-agreement score ($SAGR$). Results for our CNN model, jointly trained to regress both valence and arousal, are shown in table 5. Notice the similar performance of models based on Kinect and DMHS-SMPL-T 3d pose features.

Emotion Axis	Pose Feature	RMSE ↓	PCC ↑	SAGR ↑
Valence	Kinect	0.116	**0.184**	0.787
	DMHS-SMPL-T	**0.099**	0.169	**0.844**
Arousal	Kinect	0.111	0.345	0.973
	DMHS-SMPL-T	**0.107**	**0.388**	**0.977**

Table 5. Continuous emotion prediction. Using 3d skeleton estimates of DMHS-SMPL-T, we obtain better or similar results compared to the 3d skeleton produced by Kinect.

7. Conclusions

We have introduced large-scale fine-grained action and emotion recognition tasks defined on non-staged videos recorded during robot-assisted therapy sessions of children with autism. The tasks are challenging due to the large numbers of sequences (over 3,700), long videos (10-15 minutes each), large number of highly variable actions (37 child action classes, 19 therapist actions), and because children are only partially visible and observed under non-standard camera viewpoints. Age variance and unpredictable behavior add to the challenges. We investigated how state-of-the-art RGB 3d human pose reconstruction methods combining feedforward and feedback components can be adapted to the problem, and evaluated multiple action and emotion recognition baselines based on 2d and 3d representations of the child and therapist. Our results indicate that properly adapted, the current 2d and 3d reconstruction methods from RGB data are competitive with industrial grade RGB-D Kinect systems. With action recognition baselines in the 40-50% performance range, the large-scale data we introduce represents a challenge in modeling behavior, with impact in both computer vision, and child-robot interaction with applications to autism.

Acknowledgments: This work was supported in part by the EU Horizon 2020 Grant No. 688835 DE-ENIGMA, European Research Council Consolidator grant SEED, CNCS-UEFISCDI PN-III-P4-ID-PCE-2016-0535, and SSF. Corresponding authors: Vlad Olaru and Cristian Sminchisescu.

References

[1] Autism Europe. http://www.autismeurope.org/about-autism/prevalence-rate-of-autism/.

[2] Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/data.html.

[3] Employing Milo for Autism. http://thegarlandmessenger.com/employing-milo-for-autism.

[4] Robokind. Advanced Social Robots. http://robokind.com/.

[5] Softbank Robotics. https://www.ald.softbankrobotics.com/en/robots/nao.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[7] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[8] Y. Du, Y. Fu, and L. Wang. Skeleton based action recognition with convolutional neural network. In *ACPR*, 2015.

[9] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[10] P. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H.-L. Cao, M. Coeckelbergh, C. Costescu, D. David, A. Beir, Y. Fang, Z. Ju, J. Kennedy, H. Liu, A. Mazel, A. Pandey, K. Richardson, E. Senft, S. Thill, G. Perre, B. Vanderborght, D. Vernon, H. Yu, and T. Ziemke. How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn: journal of behavioral robotics*, pages 18–38, 2017.

[11] W. Farr, N. Yuill, and H. Raffle. Social benefits of a tangible user interface for children with autistic spectrum conditions. *Autism*, 14(3):237–252, 2010.

[12] V. Gizzonio, P. Avanzini, M. Fabbri-Destro, C. Campi, and G. Rizzolatti. Cognitive abilities in siblings of children with autism spectrum disorders. *Experimental Brain Research*, 232(7):2381–2390, 2014.

[13] Z. Huang, C. Wan, T. Probst, and L. V. Gool. Deep learning on lie groups for skeleton-based action recognition. *CoRR*, abs/1612.05877, 2016.

[14] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014.

[15] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[16] Q. Ke, M. Bennamoun, S. An, F. A. Sohel, and F. Boussaïd. A new representation of skeleton sequences for 3d action recognition. *CoRR*, abs/1703.03492, 2017.

[17] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[18] M. Lhommet and S. C. Marsella. Expressing emotion through posture and gesture. *The Oxford Handbook of Affective Computing*, 2015.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[20] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.

[21] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[22] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[23] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *Transactions on Affective Computing*, 2017.

[24] D. Moore, P. McGrath, and J. Thorpe. Computer-aided learning for people with autism - a framework for research and development. *Innovations in Education and Training International*, 37:218–227, 2000.

[25] M. A. Nicolaou, H. Gunes, and M. Pantic. Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. German Research Center for AI (DFKI), 2010.

[26] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017.

[27] C. A. Pop, R. E. Simut, S. Pintea, J. Saldien, A. S. Rusu, J. Vanderfaeillie, and B. Vanderborght. Social robots vs. computer display: Does the way social stories are delivered make a difference for their effectiveness on asd children? *Journal of Educational Computing Research*, 49(3):381–401, 2013.

[28] A. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[29] S. Ramdoss, R. Lang, A. Mulloy, J. Franco, J. O'Reilly, R. Didden, and G. Lancioni. Use of computer-based interventions to teach communication skills to children with autism spectrum disorders. a systematic review. *Journal of Behavioral Education*, 20:55–76, 2011.

[30] J. M. Rehg, G. D. Abowd, A. Rozga, M. Romero, M. A. Clements, S. Sclaroff, I. Essa, O. Y. Ousley, Y. Li, C. Kim, et al. Decoding children's social behavior. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3414–3421. IEEE, 2013.

[31] M. J. Salvador, S. Silver, and M. H. Mahoor. An emotion recognition comparative study of autistic and typically-developing children using the zeno robot. In *ICRA*, pages 6128–6133. IEEE, 2015.

[32] J. Shen, E. Ainger, A. M. Alcorn, S. B. Dimitrijevic, A. Baird, P. Chevalier, N. Cummins, J. J. Li, E. Marchi, E. Marinoiu, V. Olaru, M. Pantic, E. Pellicano, S. Petrovic, V. Petrovic, B. R. Schadenberg, B. Schuller, S. Skendi, C. Sminchisescu, T. T. Tavassoli, L. Tran, B. Vlasenko,

M. Zanfir, V. Evers, and C. De-Enigma. Autism data goes big: A publicly-accessible multi-modal database of child interactions for behavioural and machine learning research. *International Society for Autism Research Annual Meeting*, 2018.

[33] 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 *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304. IEEE Computer Society, 2011.

[34] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian. A pilot study with a novel setup for collaborative play of the humanoid robot kaspar with children with autism. *International Journal of Social Robotics*, 6(1):45–65, 2014.

[35] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn. Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism. *IEEE Transactions on Autonomous Mental Development*, 6(3):183–199, 2014.

[36] A. Zanfir, E. Marinoiu, and C. Sminchisescu. Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes – The Importance of Multiple Scene Constraints. In *CVPR*, 2018.

[37] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The "Moving Pose": An Efficient 3D Kinematics Descriptor for Low-Latency Action Detection and Recognition. In *International Conference on Computer Vision*, December 2013.

[38] X. Zhou, M. Zhu, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016.