IEEE

IEEE Robotics & Automation Society

ENGAGEMENT ESTIMATION DURING CHILD ROBOT INTERACTION USING DEEP CONVOLUTIONAL NETWORKS FOCUSING ON ASD CHILDREN

Dafni Anagnostopoulou[1], Niki Efthymiou[1], Christina Papailiou[2], Petros Maragos[1]

[1]School of ECE, National Technical University Athens, Greece
[2]Department of Early Childhood Education and Care, University of West Attica, Greece

Engagement Estimation During Child Robot Interaction

Social robots increasingly involved in daily lives:

- educational process of children
- therapeutic purposes for children with autism spectrum disorders

For qualitative interaction: robots **must** adapt their behavior to children cognitive state.

Key characteristic of human response to an interaction: **Engagement**

Engagement:
the level at which the child is both attentive and cooperative with their partner towards their common goal.

Engagement Estimation During Child Robot Interaction: Challenges

Engagement: **internal mental state**

However:

Observers have to resort to **external cues** like:

- vision
- speech/audio

to estimate its level.

Most of the time: middle engagement level

↓

Fully engaged or fully disengaged instances relatively rare - difficult to train models

Indicative cues like:

- eye gaze
- blinking
- head-pose

do **not** appear so connected with engagement in ASD children.

→ Engagement is easier to predict for TD children than for ASD children.

Engagement Estimation During Child Robot Interaction: Challenges

Engagement: **internal mental state**

However:

Observers have to resort to **external cues** like:

- vision
- speech/audio

to estimate its level.

Most of the time: middle engagement level

↓

Fully engaged or fully disengaged instances relatively rare - difficult to train models

Indicative cues like:

- eye gaze
- blinking
- head-pose

do **not** appear so connected with engagement in ASD children.

→ Engagement is easier to predict for TD children than for ASD children.

Engagement Estimation During Child Robot Interaction: Challenges

Engagement: **internal mental state**

However:

Observers have to resort to **external cues** like:

- vision
- speech/audio

to estimate its level.

Most of the time: middle engagement level

↓

Fully engaged or fully disengaged instances relatively rare - difficult to train models

Indicative cues like:

- eye gaze
- blinking
- head-pose

do **not** appear so connected with engagement in ASD children.

→ Engagement is easier to predict for TD children than for ASD children.

Contributions

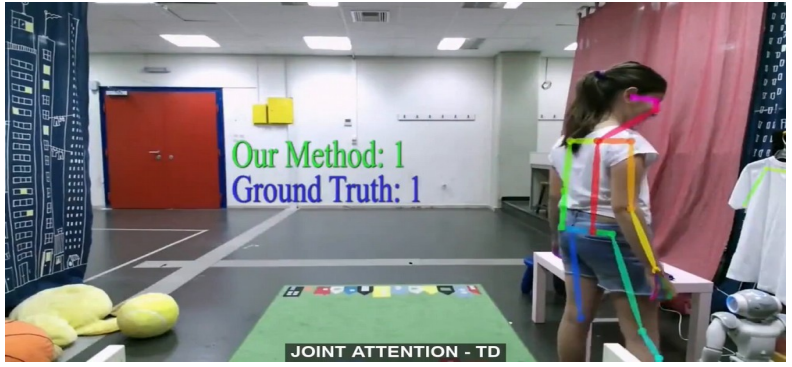
Main Goal: Estimate the engagement level for ASD children interacting with social robots.

- Tested many different architectures.
- Concluded to these that outperformed previous
- Test generalization

ASD & TD in same interaction with robot

ASD in many different interactions with robots

ASD in interaction with their mothers



Variety of participants, conditions and interactions

Data Set: ASD Games & ASD Joint Attention

- 7 sessions of 20 minutes
- 7 children facing autism spectrum disorder
- Especially adapted laboratory
- Robots: Nao and Furhat
- Five different games: Show me the Gesture, Express the Feeling, Pantomime, Guess the Object and Joint Attention
- *ASD-Games Data* & *ASD-Joint Attention Data*

Data Set: TD Joint Attention & BabyAffect

- 25 TD children
- Nao robot in joint attention task
- *TD-Joint Attention Data*

- 3 younger ASD children
- Playing with mothers in home
- *BabyAffect Data*

Joint attention with a robot: ASD spent twice as long time disengaged compared to the TD children (7.80% vs. 17.68%). This time is doubled in the human condition (34.62%).
ASD children spent approximately the same time cooperating with a robot almost in all structured conditions in the laboratory.

Data Annotation

Class 0: **disengaged**

- pays limited or no attention to the robot
- does not act towards their common goal

Class 1: partially **engaged**

- acts relatively to the common goal **or**
- pays attention to the robot

Class 1:**fully engaged**

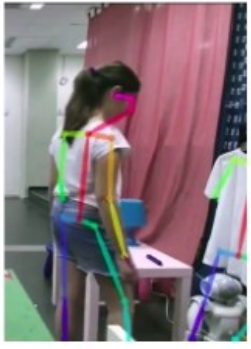
- actively cooperates with the robot
- to complete common goal

Data annotations by laboratory members according to a set of instructions containing groups of visual and acoustic cues that correspond to each engagement level provided by an expert psychologist.

Data Processing

OpenPose library to extract 18 body, 2 x 21 hands and 70 face 2D keypoints.

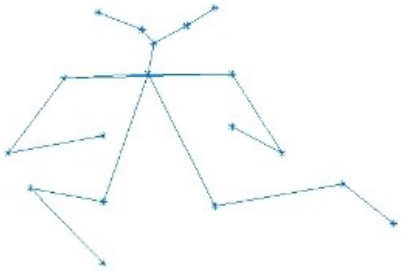
- missing values → linear interpolation
- depth & multiple views → 3D keypoints (joint attention)
- children interact with partners → calculate pose regarding partner
- interested in relation between pose parts → subtract left hip coordinates
- normalize feature values



Many skeletons detected (other people, items like dolls etc) → Must find children pose.
Use:

- previous poses
- torsos lengths

Data Processing



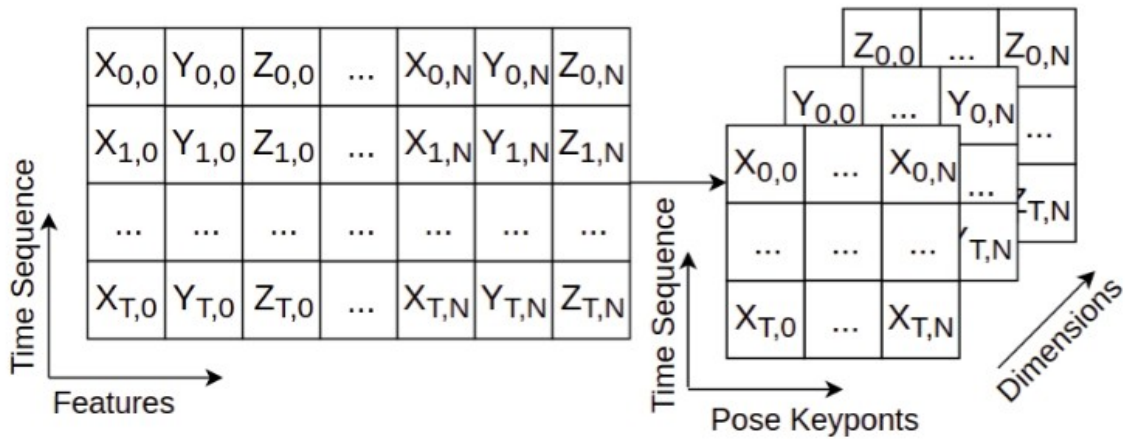
Skeleton example of babyaffect

Engagement depends on temporal information (use of LSTMs)

Pose features resemble images which are successfully processed by CNNs

Action recognition via pose: Rearrange feature vectors to resemble images

Horizontal axis	Skeleton Parts
Verical axis	Time
Image channels	Part Coordinates



D. C. Luvizon, D. Picard, and H. Tabia, “2d/3d pose estimation and action recognition using multitask deep learning,” in Proc. CVPR. IEEE, 2018.

Method

1

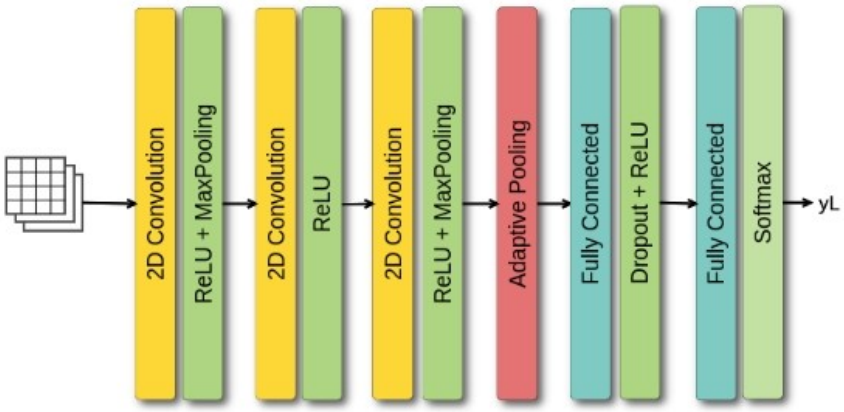
AlexNet Architecture

Convolutional layers with suitable to our inputs characteristics.

2

Simpler 2D CNN

For greater computational efficiency (time and space)



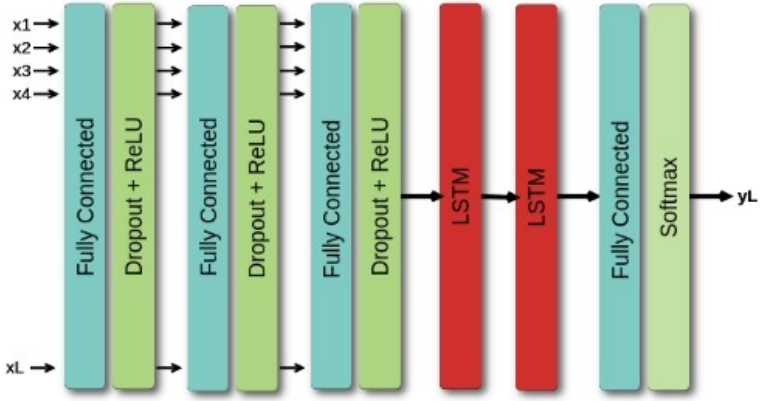
Number of channels
=
data dimensionality

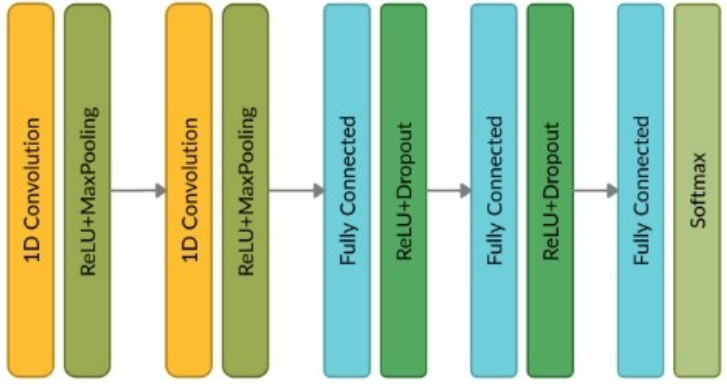
- Use PyTorch library
- Data augmentation: flip vertically & add small amount of Gaussian noise
- Batch size: 128.
- Learning rate of 0.0003.
- Sequence length of 200 frames (6 to 7 seconds).
- Optimization algorithm: Adam Optimizer.
- Scheduler: ReduceLROnPlateau.
- Loss function: Cross Entropy Loss

Method

Compare results with:

1. Recurrent neural network based on LSTM layers [1] & improved.
2. One dimensional multi-channel convolutional network [2] & improved.
3. Network based on ResNet-50 with RGB input [3].





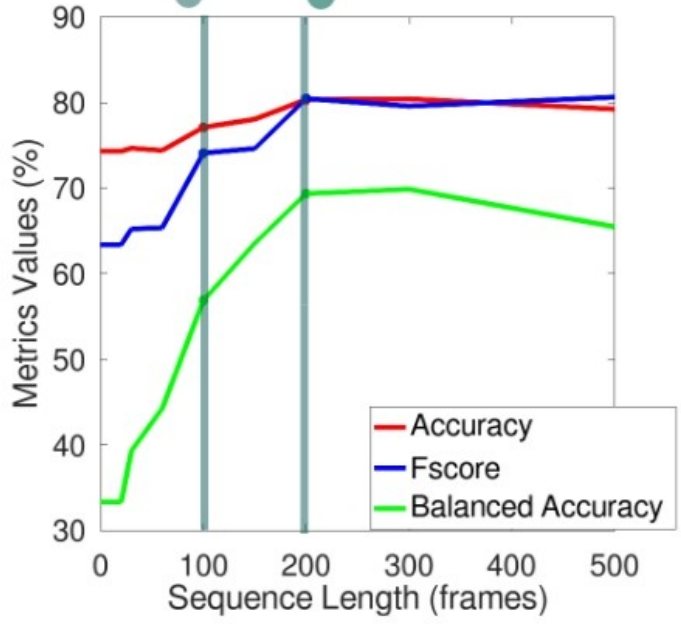
[1] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," in Proc. IROS. IEEE, 2019.

[2] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder–a personalized computational modeling approach," Frontiers in Robotics and AI, vol. 7, pp. 43, 2020.

[3] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in Proc. CVPR Workshop. IEEE, 2019.

Results: Sequence Length

Sequences **larger than 100 frames** (approximately 3 seconds) allow the network to train and estimate engagement.



Best results: sequences of **200 frames**, i.e. approximately 6 to 7 seconds

Sequence length corresponds to the time window that the neural network "sees" every time.

In accordance with corresponding psychologists conclusions. The time frame 3 to 6 seconds is considered fundamental to human perceptual functions.

- Change of evaluation metrics for different sequence lengths (fps 30 secs) given to the convolutional AlexNet network to estimate engagement.
- Similar results regardless of network architecture!

Results: Joint Attention Data

Network	Accuracy	F-score	W. Precision
majority class	74.32	63.38	55.25
1D CNN	72.98	60.15	56.32
ResNet50	74.52	64.52	63.285
LSTM (one layer)	76.23	74.15	74.35
1D CNN	77.44	75.15	76.30
2D CNN	78.93	76.46	77.43
LSTM	79.47	76.88	78.04
AlexNet	**80.36**	**80.48**	**80.71**

- TD pretrained networks
- Fast training
- Small amount of ASD joint attention data

We obtain networks that succeed in estimating ASD engagement.

TD Joint Attention

Accuracy

Fscore

ASD Joint Attention

Accuracy

Fscore

Results: ASD Games Data

Accuracy

Fscore

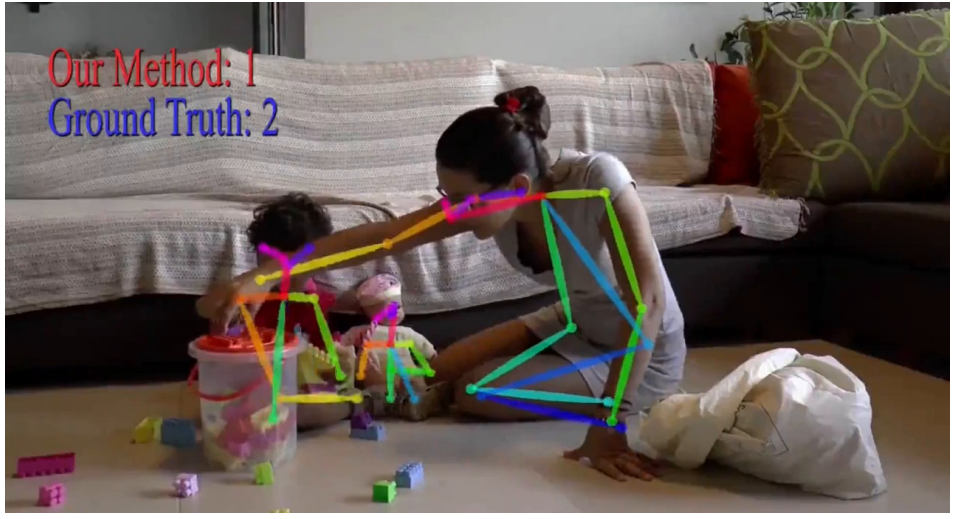
- Satisfying engagement estimations for these interactions
- Less accurate than Joint Attention data sets.
- Variety of interactions during which children are asked to talk, gesture, move around the room or play before a screen.

Results: BabyAffect

2D CNN

Accuracy

Fscore



- Different and difficult conditions, ASD children play with their mothers in their home environment.
- Both convolutional networks achieve high engagement estimation accuracy results.
- Can estimate engagement for children-adult interactions too.

Conclusion

- Focus on engagement estimation for children with autism during interaction with robots.
- Deep convolutional architectures trained with pose features.
- Extensive experiments showed the superiority of our method to previous ones.
- **Greater challenge:** create a model that can relatively easily adapt from its training conditions to different ones.
- **Future work:** test and generalize our method to different kind of interactions, such as interactions during which children are seated.

THANK YOU