# On the Estimation of Children's Poses

Giuseppa Sciortino[2(✉)], Giovanni Maria Farinella[1], Sebastiano Battiato[1], Marco Leo[2], and Cosimo Distante[2]

[1] IPLAB, Department of Mathematics and Computer Science, University of Catania, Catania, Italy
{gfarinella,battiato}@dmi.unict.it
[2] ISASI, Institute of Applied Sciences and Intelligent Systems, C.N.R National Research Council, Lecce, Italy
giuseppa.sciortino@isasi.cnr.it, {marco.leo,cosimo.distante}@cnr.it
http://iplab.dmi.unict.it/
http://www.isasi.cnr.it/

**Abstract.** Deep Learning architectures have obtained significant results for human pose estimation in the last years. Studies of the state of the art usually focus their attention on the estimation of the human pose of adults people depicted in images. The estimation of the pose of child (infants, toddlers, children) is sparsely studied despite it can be very useful in different application domains, such as Assistive Computer Vision (e.g. for early detection of autism spectrum disorder). The monitoring of the pose of a child over time could reveal important information especially during clinical trials. Human pose estimation methods have been benchmarked on a variety of challenging conditions, but studies to highlight performance specifically on children's poses are still missing. Infants, toddlers and children are not only smaller than adults, but also significantly different in anatomical proportions. Also, in assistive context, the unusual poses assumed by children can be very challenging to infer. The objective of the study in this paper is to compare different state of art approaches for human pose estimation on a benchmark dataset useful to understand their performances when subjects are children. Results reveal that accuracy of the state of art methods drop significantly, opening new challenges for the research community.

**Keywords:** Human pose estimation · Deep learning methods

## 1 Introduction and Motivations

Human Pose Estimation has obtained remarkable interest by the community in the last decades [1]. Thank to the advancements of deep learning and the availability larger labeled datasets, a boost on the pose estimation accuracy has been achieved. The main aim of human pose estimation is focused on finding joints (usually called keypoints) and parts (connections between joints) of people present in image to infer the body pose (Fig. 1a).


S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 410–421, 2017.
https://doi.org/10.1007/978-3-319-68548-9_38

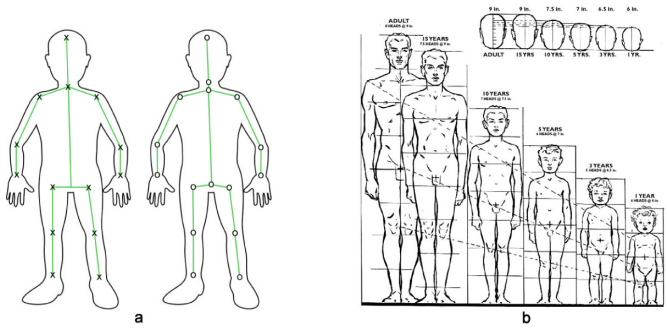


**Fig. 1.** a: Human skeleton, the human parts in green, the symbols "x" and "o" represent the joints, typically 14–16 joints are used to model the skeleton, 14 on the right and 16 on the left respectively. b: Age-related changes in human body proportions, image from book of Andrew Loomis "Figure Drawing For All It's Worth" (Color figure online)

The pose estimation is useful to analyze the human behaviors in many scenarios, such as in the context of Ambient Assisted Living (AAL) [2] for recognizing the user's activities of daily living, to the study and diagnosis of motor and psychological disorders, etc. The most of the available datasets used for human pose estimation [3–6] does not contain a sufficient number of images depicting children to evaluate the performances of the algorithms in this specific case. In the last years, MPII Human Pose Dataset [1] became a state of the art benchmark repository for the evaluation of human pose estimation algorithms. It contains 25K images related to 410 human activities performed in indoor and outdoor environments. Despite it is an excellent resource for human pose estimation algorithms, like other datasets does not contain enough images of children.

**Human Pose and Body Proportions in Children.** Although human pose estimation is a problem that has been studied for many years, most of the literature is focused on pose of adults, whereas it was sparsely studied in the case of children. It might seem that adults and children have the same body shape and therefore the same skeleton, but infants and children are not adults "miniatures", they are structurally different. The human body grows and develops continuously (not even uniformly) from birth through old age [7]. The proportion between body parts changes according to a predictable trend (Fig. 1b). For example, the length of head in adults is about one-seventh of the total body length, whereas in the infant is one-fourth. Proportionally, the trunk length is longer in children with respect to adults. The proportions of trunk and limbs change during the growth, and the lower limbs increase in length more rapidly than the upper limbs. The anatomy of the child's neck is one of the most particular aspects. The neck muscle strength increases with age, in children they are not generally properly developed and tend to appear as flattened due to the greater mass head perched on a thin neck. Indeed, in many images of children the neck joint is not visible. To the best of our knowledge there are no studies in the

literature focusing on the evaluation of the state of art algorithms for pose estimation which consider human at early ages (children). In [8] is described a tools for the non-invasive assessment of autism spectrum disorder (ASD) considering four behavioral markers: visual tracking, disengagement of attention, sharing interest, and atypical motor behavior. The last marker is evaluated using a pose estimation algorithm based on Object Cloud Model [9] to detect an asymmetrical position of the arms. In [10] a work to simulate babysitter's vision is presented. The main aim is the tracking of a child-object in an indoor and outdoor environment. The algorithm is able both to track whole child-object and to track body parts (head, hands, legs and feet) of the child. In [11] a method to estimate body pose of infants in depth images by using random ferns is introduced. A pixelwise body part classifier is proposed and joints are located computing the center of mass of the points belonging to each body part. To the best of our knowledge no attempt has been made so far to establish a more representative dataset aiming to cover children's images for human pose estimation. A focused work to study child behaviors is reported in [12], where a study of typical autistic behaviors is performed. The annotations examine a set of representative attributes of the behavior (as stimming behavior category and intensity) but the ground truth of the joints is not available for this dataset. Some of these videos are used in the present work to build our experimental dataset.

In this paper we present a benchmark dataset in which the subjects are related to children and toddlers. On this dataset we compare different state of the art methods [13–16] to estimate the human pose and by performing an indept evaluation. The remainder of this paper is organized as follows. In Sect. 2, we briefly discuss some relevant related works for human pose estimation. In Sect. 3 we introduce the methods used for the comparative evaluation and present the benchmark dataset describing the collection and annotation processes. In Sect. 4, we detail the experiments and results. Section 5 concludes the paper.

## 2 Related Work

**Human Pose Estimation.** In this section we briefly review state of art methods for human pose estimation exploiting convolutional neural networks, by highlighting their main peculiarities. Nowadays, many limitation of classic approaches have been overcome through the widely use of convolutional neural networks. In the literature there are many models to solve articulated pose estimation exploiting deep learning architectures. To solve 2D human pose estimation from single image, some methods regress image to build heatmap or confidence maps [14,15,17–20], where heatmap represents the probability that the joints appear in a particular position of the image. Differently, in [21] Cartesian coordinates of human joints are directly predicted. Regressing heatmaps are preferred because the framework can be multimodal, indicating the existence of multiple joints. In [13] is presented a hybrid method that regresses on both heatmaps and cartesian coordinates.

**Human Pose Estimation Multiperson.** In the case of multi-person pose estimation, most of the approaches require a preliminary step where the person is detected [22–24]. This makes the results dependent on the correct detection of people within the images. In [25] the interdependence from people detection is considered, but the method requires additional initial assumptions. The authors of [16] propose a bottom-up method, which detects the joints as first step and then associates the different joints in parts to build the full skeleton. On the contrary in [26] a top-down strategy is presented. In the first stage the method predicts and scale the bounding box containing persons, then it regresses the locations of joints and finally performs the pose predictions.

## 3 Methods and Dataset

In this section we briefly describe the four methods we have considered to benchmark the problem of human pose estimation on images depicting children [13–16]. We also describe the dataset collected to perform the comparative evaluation of the different methods.

**Recurrent Human Pose Estimation (RNN).** In [14] a recurrent neural network model to estimate the human pose has been proposed. The approach is top-down based (i.e., joints are detected first and then pose is inferred without the need detecting the person as first step). The proposed architecture is composed by two main modules: Feed-Forward Module and Recurrent Module. The aim of the first module, is the detection of the "body joints" by regressing heatmaps (one for each joint) without knowing the body configuration or the association between couples of joints. The second module takes in input the heatmaps from the feed-forward module and it infers the contextual information. The first layer of feed-forward module use small filters, whereas larger filters are used in deep layers to learn the structures of the body. The whole network can be trained in an end-to-end fashion and outputs 16 joints (Fig. 1a). The network was trained and tested on MPII Human Pose [1] dataset and it has been tested on extended LSP [27] and on MPII [1] datasets obtaining good performances.

**Human Pose Estimation with Iterative Error Feedback (IEF).** Human pose Estimation with Iterative Error Feedback [13] introduces an iterative convolutional neural network to predict a body pose from 2D images. The authors present an iterative self-correcting model by feeding back-error prediction. Given a preliminary guess solution (i.e., Cartesian coordinates representations of joints positions) for each iteration the method applies a "bounded correction". In this way it predicts a direction in which to move a final solution. This correction is used to update the joints positions, and then the process is iterated. The IEF method takes in input the coordinates of any point that belongs to the torso as additional information. The architecture of the deep convolutional network consists of a pre-trained network (i.e., Imagenet [28]) where the first and the last layers were appropriately modified to adapt at the problem. The network is tested on two datasets, MPII [1] and LSP [27], obtaining good performances.

**Convolutional Pose Machine (CPM).** In [15] the authors proposed a method, Convolutional Pose Machine, to estimate human pose from 2D image for single-person. The method consists of a series of sequential convolutional networks. At every stage a network takes as input the belief map supplied as output from the previous network. The Convolutional Pose Machine is trained to learn image features and image-dependent spatial models. Each step corresponds to a sequential refinement. In the first stage a convolutional network is applied to obtain belief map from local evidence by using small receptive field. Successive stage use multiple layers to reach large receptive field in order to capture complex and long-range correlations between parts. CPM model takes in input additional information: a bounding box position containing the subject on which to estimate the pose. The model can be trained in end-to-end mode from scratch and outputs 14 joints. In evaluation section of the original paper, different training schemes are analyzed on LSP dataset. The performance are evaluate on three datasets (MPII [1], LSP [27] and FLIC [29]) obtaining good performances.

**Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields (MPP).** In [16] a model to estimate 2D human pose for multi-person with bottom-up approach is presented. The method improves the performance with regard to computational time of the state of art. The proposed architecture consists of two-branch multi stage convolutional neural networks, and introduces the Part Affinity Field (PAF), namely a set of 2D vector fields related to body parts. The two branches work separately. The first predicts the confidence score maps for part detector and the second predicts PAFs for part association. Then their output are concatenated and the process is iterated. In the same manner, the model learns implicit spatial relationship between different people. The model is trained in end-to-end fashion and it outputs 14 joints, if all of them are detected. The performance are tested on the benchmark dataset MSCOCO [30] and on the MPII [1] outperforming the state-of-art. The authors provide an analysis of the computational time, the good performances allow have realtime method and to be applied for video sequence analysis.

### 3.1 Dataset

In this work we exploit a dataset of images of a particular age group: toddlers and children (approximately one-eleven years old). The images have been extracted from videos available on public domain websites and video portals. The dataset covers various activities typical of this age group as: 'learning to walk', 'sports' and 'play'. The collected images depict different variabilities, such as: indoor, outdoor, clothing type and include interaction with various objects and environments. The images are not splitted in activity categories since our purpose is not to recognize activities but to evaluate human pose estimation algorithms on child's images, to highlight and understand the differences in accuracy with regard to the state of art when used on adult's images. The available datasets in literature do not consider this sub-category and they contain very few images

to carry out a specific evaluation. Our dataset covers a wide variety of poses. Children often assume poses more articulate than adults. Often due to privacy issues, the datasets regarding to child are not easily available on the web, and for this reason the images of our dataset are chosen from free videos available on youtube. Amateur and professional videos allowed to obtain a wide variety of subjects and varieties, due to the different acquisition sources.

**Images Collection.** As mentioned above, the dataset in [12] provides a list of 75 video URLs. These videos regarding to children behavioral disorders, are divided into three categories of stimming behaviors: arm flapping, head banging and spinning, that are typical behaviors observed in ASD. The videos in [12] were the starting point for our collection and research process. Some of them were discarded because they were not suitable for our purposes. For example some videos depict interaction between more persons, or the poses show strong truncations, same videos that show low quality. Other videos were searched and selected on youtube, using keywords like 'children', 'toddler', 'walking', 'learn to walk', 'children', 'video', 'talent show', 'gymnast', 'dancing', 'play', 'autism', and their possible combinations. In this way 150 videos were collected. The selected videos have been posted in youtube mainly by relatives or talent show. Afterwards, for each collected video all frames were extracted, without post-processing, preserving the original natural setting and a subset has been manually selected from several non-consecutive frames, trying to ensure different poses. We obtained a dataset with 1176 images related to 104 different subjects. This dataset is available for the research community upon request to the authors.

**Images Annotation.** The collected images were annotated by using a tool available on line [31]. The annotation process is designed by clicking control points in the image recording the positions, labels and visibility for each selected keypoints. The tool was slightly modified in order to mark up to 22 visible/occluded labeled keypoint locations: head (forehead, chin) ears, eyes, nose, mouth, neck, arms (shoulder, elbow, wrist), torso, legs (hip, knee, ankle). The annotations in our dataset are person-centric (i.e., right/left corresponding at right/left body parts of person) and are saved in an xml file for each image.

## 4 Comparative Evaluation

In this section we evaluate the performances of the methods described above on our dataset[1]. For the evaluation we considered the PCK measure [32], one of the most commonly used in the literature to measure the accuracy of detected joints. Generally, it is used a modified PCK measure, denoted PCKh, that considers a localized joint as correct if the distance between the predicted joint and correspondent ground-truth is less of 50% of head segment length. In this way, the PCKh measure is independent from the size of the bounding box considered by the measure PCK. The compared methods output different number of joints,

[1] The methods have been exploited considering pre-trained models without re-training or fine-tuning.

16 in [13,14] and 14 in [15,16] (see Fig. 1a). In our analysis, the joints related to eyes, nose, ears and mouth are not considered. In particular, for [15,16] shall not be considered chin and torso joints in evaluations because the joint relative to the chin and torso are not covered by their output (bin number 8 'Torso' and 10 'D Head' are missing in Figs. 3 and 4).

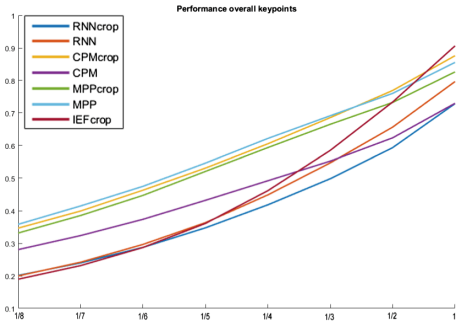


**Fig. 2.** Overall performance evaluation, for CPM and MPP methods considering 14 joint and 16 joints for IEF and RNN methods. With CPM crop, MPP crop and RNN crop we indicate the achieved performances on "cropped" images.

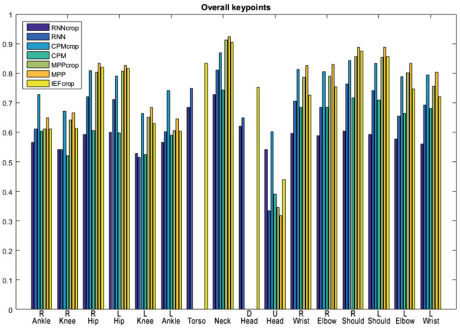


**Fig. 3.** Histogram performance evaluation per joint. In the x-axis the name of joints whereas in the y-axis the percentage of joints that are correctly detected by fixing the percentage to 50% of the head segment length for PCKh measure.

**Performance Evaluation.** Here we show the achieved performances for considered methods [13–16] on our dataset. Considering that CPM [15] and IEF [13] methods require additional input, we cropped every image on bounding box containing the person on which estimate the pose. The bounding box is roughly determined on ground truth annotation. Furthermore for IEF [13] method we considered that the center of the bounding box is the point belonging to torso. Figure 2 shows the performances of the considered methods depending on percentage of the head segment length taken into account in the evaluation measure. In the y-axis the percentage of joints that are correctly detected. For MPP, RNN and CPM methods are showed the performance achieved without input parameters. It is interesting to note that MPP method achieved the best performance when the full image is considered, without additional information about bounding box. In Fig. 3 are shown the performances for single joint. The performances are obtained by fixing the percentage to 50% of the head segment length for the evaluation measure PCKh.

In the analysis, the joints not present in the image (i.e., joints not present because the body part is not depicted in the image, namely truncations) are excluded from the evaluation for not penalizing the methods that output a fixed number of joints [13–15]. In Fig. 4 are presented the performances per single

joints, where we split the joints in visible and occluded sets based on ground truth label. The histograms in Figs. 3 and 4 reveal that better performances for the joints belonging to the trunk (neck, shoulder, hip) than joints belonging to the limbs (elbow, wrist, upper head, knee, ankle), it is most evident in the case of occlusions (see Fig. 4 on the right). This could be due to convolutional layers that implicitly encode a configuration model. On the other hand, the deep networks are trained on dataset where the subjects are mainly adults. Therefore the deep networks may have learned the anatomical proportions of adults, in agreement with what has been described in the introduction section.

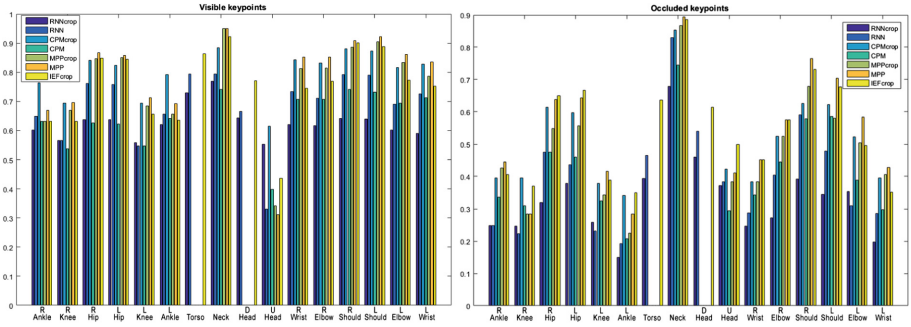


**Fig. 4.** Histogram performance evaluation on each joint 'visible' and 'occluded', on the right and on the left respectively. In the x-axis the name of joints, in the y-axis the percentage of joints that are correctly detected by fixing the percentage to 50% of the head segment length for 'PCKh' metric.

Occlusions and truncations can drastically affect performance. The best case is certainly represented by poses in which all the joints are visible and distinct. In this regard, we evaluate the performances of methods on subset of our datasets. We divide the dataset into four subsets:

**Case 1:** all the joint are present and visible in the image;
**Case 2:** all the joint are present but some are not visible;
**Case 3:** not all the joint are present in the image, those present are all visible;
**Case 4:** not all the joint are present in the image and some are not visible.

Based on these splitting we obtained four subsets. The first is about 20% of our dataset, the second 65%, the third 5%, and 10%. In Figs. 5 and 6 the performance are shown for each case depending on the head segment length taken into account.

To evaluate the accuracy of the entire pose for each image we calculated the average distance between each joint position and the respective ground truth joint position, if the average distance is less than 50% of head length segment the pose is considered as correct. In Table 1 the percentage of images that do not satisfy this relationship is showed. We observed that the highest performing

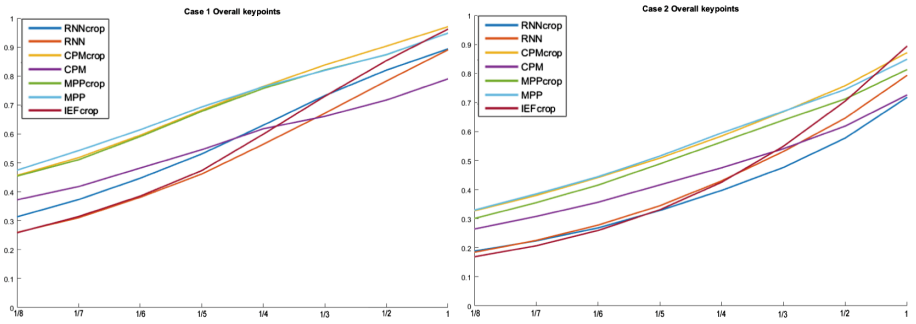


**Fig. 5.** Overall performance evaluation, Case 1: no occlusion and no truncation. Case 2: some occlusion but no truncations.

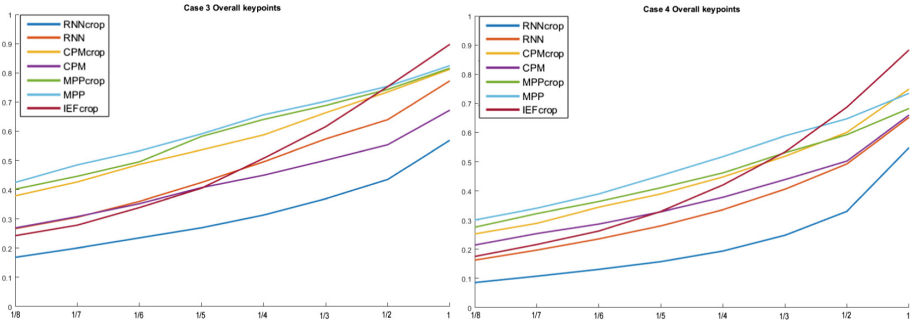


**Fig. 6.** Overall performance evaluation, Case 3: no occlusion but truncations are present, Case 4: occlusion and truncations are present.

**Table 1.** Percentage of worst pose based on average distance between all joints belonging to the pose and ground truth.

| | RNNcrop | RNN | CPMcrop | CPM | MMPcrop | MMP | IEF |
|---|---|---|---|---|---|---|---|
| Case 1 | 23% | 24% | 4% | 32% | 26% | 0% | 7% |
| Case 2 | 58% | 0.43% | 27% | 45% | 49 | 10% | 29% |
| Case 3 | 76% | 39% | 35% | 47% | 41% | 0% | 17% |
| Case 4 | 0.83% | 0.60% | 46% | 56% | 63 | 20% | 25% |

method is the one proposed in [16], although it is not the best in Case 1, it allows to have a good performances of the whole pose (despite occlusions and truncations), without requiring additional input parameters.

## 5 Conclusion

In this study we considered the problem of estimating pose of children from images. The aim is to verify whether the performance of state of the art methods on our dataset are comparable to those obtained on the benchmark datasets containing mainly images of adults. We have collected and annotated a dataset images containing children extracted from videos recorded in an uncontrolled environment and available on public domain websites and video portals. We have compared four well known methods on our dataset. Experiments point out that accuracy drops down for all methods in this application context. The results open new research challenges, especially for non-invasive assessment of behavioral or motor disorders of children for assistive technology. We expect that by retraining models the results can be improved. We plan to fine tune of the considered models and then extend our dataset in the next study.

## References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693 (2014)
2. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M.: Computer vision for assistive technologies. Comput. Vis. Image Underst. **154**, 1–15 (2017)
3. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int. J. Comput. Vis. **87**(1–2), 4–27 (2010)
4. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2014)
5. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: IEEE International Conference on Computer Vision, pp. 3192–3199 (2013)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
7. Huelke, D.F.: An overview of anatomical considerations of infants and children in the adult world of automobile safety design. Annu. Proc./Assoc. Adv. Automot. Med. **42**, 93–113 (1998)
8. Hashemi, J., Spina, T.V., Tepper, M., Esler, A., Morellas, V., Papanikolopoulos, N., Sapiro, G.: Computer vision tools for the non-invasive assessment of autism-related behavioral markers. arXiv preprint arXiv:1210.7014 (2012)
9. Miranda, P., Falcão, A., Udupa, J.: Cloud models: their construction and employment in automatic MRI segmentation of the brain (2010)
10. Aljuaid, H., Mohamad, D.: Child video dataset tool to develop object tracking simulates babysitter vision robot. J. Comput. Sci. **10**(2), 296–304 (2014)
11. Hesse, N., Stachowiak, G., Breuer, T., Arens, M.: Estimating body pose of infants in depth images using random ferns. In: IEEE International Conference on Computer Vision Workshop, pp. 427–435 (2015)

12. Rajagopalan, S., Dhall, A., Goecke, R.: Self-stimulatory behaviours in the wild for autism diagnosis. In: IEEE International Conference on Computer Vision Workshops, pp. 755–761 (2013)
13. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4733–4742 (2016)
14. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. arXiv preprint arXiv:1605.02914 (2016)
15. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
16. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. arXiv preprint arXiv:1611.08050 (2016)
17. Tompson, J.J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
18. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems, pp. 1799–1807 (2014)
19. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9911, pp. 717–732. Springer, Cham (2016). doi:10.1007/978-3-319-46478-7_44
20. Jain, A., Tompson, J., Andriluka, M., Taylor, G.W., Bregler, C.: Learning human pose estimation features with convolutional networks. arXiv preprint arXiv:1312.7302 (2013)
21. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1653–1660 (2014)
22. Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., Schiele, B.: Articulated people detection and pose estimation: reshaping the future. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3178–3185 (2012)
23. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3582–3589 (2014)
24. Iqbal, U., Gall, J.: Multi-person pose estimation with local joint-to-person associations. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 627–642. Springer, Cham (2016). doi:10.1007/978-3-319-48881-3_44
25. Eichner, M., Ferrari, V.: We are family: joint pose estimation of multiple persons. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 228–242. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15549-9_17
26. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. arXiv preprint arXiv:1701.01779 (2017)
27. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1465–1472 (2011)
28. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)

29. Sapp, B., Taskar, B.: MODEC: multimodal decomposable models for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3674–3681 (2013)
30. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). doi:10.1007/978-3-319-10602-1_48
31. Bourdev, L., Malik, J.: The Human Annotation Tool. https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/hat/
32. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**(12), 2878–2890 (2013)