A Review on Computer Vision-Based Techniques for Autism Symptoms Detection and Recognition

Esraa T.Sadek
Computer systems. Faculty of Computers and Information Sciences
Ain Shams University
Cairo, Egypt
esraa.sadek@cis.asu.edu.eg

Noha A. Seada[2] , Said Ghoniemy[3]
Computer systems. Faculty of Computers and Information Sciences
Ain Shams University
Cairo, Egypt
noha_sabour@cis.asu.edu.eg[2]
ghoniemy1@cis.asu.edu.eg[3]

2020 15th International Conference on Computer Engineering and Systems (ICCES) | 978-0-7381-0559-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCES51560.2020.9334560

***Abstract*— Autism is a mental disorder appearing in children as a delay in their social and communicational skills. ASD causes are still mysterious but scientists believe that they are caused by genetic defects. Diagnosing autism has been an exhaustive process that attracts several researchers' attentions. In this work, an overall description of autism, its classes, signs and diagnosing protocols are covered. As computational technologies helped in assisting almost every field, it added advantages in detecting and recognizing autism. In this work, deep investigation of computer vision-based protocols in cooperation with machine learning technologies are discussed to propose autism diagnosing solution.**

***Keywords*— Autism spectrum disorder, autism signs detection, autistic self-stimulatory behaviors, Human Activity Recognition and Classification**

I. INTRODUCTION

Autism is a mental syndrome represented by symptoms of defects in socialization, interaction, and emotional expressing. Autism and other mental disorders are usually paired with repetitive motor behaviours like hand waving, body rocking, spinning, ear covering, and repeating phrases. Autistic children also usually tend to avoid eye contacts and have strong interest to their possessions. Autistics population exceeds medicals who can diagnose this disorder. Based on the Centres for Disease Control and Prevention (CDC) one of every 110 children is diagnosed with autism [1], while one of every 59 children is expected to be autistic. Although autism is a lifelong disorder; early diagnosing and treatment can improve children's development stages. The average diagnosing age is from 2 to 5 years old [2]. The main cause of autism is still mysterious; however, several researchers have contributed to help in autism diagnosis in early stages. There are several autism diagnosing techniques. Traditional protocol of diagnosing autism relies on clinical observation sessions for an autistic child. Several techniques are utilized to assist diagnosing autism including Electroencephalography signals (EEG), Functional Magnetic Resonance Imaging (fMRI), Blood and genetics analysis, wearable sensors, and computer vision techniques. Computer vision-based techniques have many advantages over the other techniques in terms of simplicity and minimized costs.

II. AUTISM DIAGNOSING

A. Autism signs

Autism has three main classes which are classical disorder, Asperger Syndrome, Syndrome and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS). Autism has several signs categorised into four classes which are developmental, communication and social, visual, and repetitive motor signs. Repetitive behaviour is one of the shared signs among the three classes of autism [3]. Repetitive behaviour sign has five forms which are:

1. Self-stimulatory behaviours (Self-stimulatory): which are set of repeated body movements like hand flapping, ear covering, head banging, or body rolling.
2. Persistent behaviours: autistics usually put object in a special form, keep checking things, or keep washing their hands several times.
3. Uniformity: Autistics usually do not like to change their daily routine, the way they play, or even the atmosphere they used to be in [4]. Autistics usually repeat their behaviours several times without getting bored. Forcing autistic children to change any of their habits increases their feeling of anxiousness.
4. Constricted interests: Autistics have sharp attachment to their properties and interests.
5. Self-hurt behaviours: in severe cases of autism, autistics usually tend to unconsciously hurt themselves. Self-hurt behaviours have several forms like skin cutting, hand biting, head banging, and eye punching eye-jabbing, skin-picking, hand-gnawing, and head-banging.

Repetitive motor behaviours are noticeable activities that can be tracked using computer vision-based techniques. Computer vision techniques in cooperation with machine learning protocols can help to detect various patterns of repetitive motor behaviours. S. S. Rajagopalan *et al.* [3] proposed a dataset of autistic children performing self-stimulatory behaviours (SSBD). The average duration of videos in SSBD is 1.5 minutes categorised into three classes of repetitive activities, which are arm waving, head rocking and rolling. The proposed data set was published with benchmark results off vision-based techniques of diagnosing self-stimulatory behaviours.

B. Autism Diagnosing protocols

Clinicals diagnose autism by long medical observation sessions measuring special parameters that define thresholds of this disorder. Several medical references are used to evaluate the observed parameters of children like ADOS (Autism Diagnostic Observation Schedule) [5], AOSI (Autism Observation Scale for Infants) [6], and DSM (Diagnostic and Statistical Manual of Mental Disorders) [7]. Repetitive motor behavior is a common symptom of autism in all these medical references. Severe autism cases perform repetitive motor behaviour in high frequency rate. About 60% of autistics perform one or more rhythmic motor behaviors [8]. Also, there are several other protocols to diagnose autism including genetics [9] ,EEG-based investigation [10], fMRI [11], wearables & sensors [12] ,as well as manual and computerized vision techniques [13] . As stereotypies and repetitive motor behaviors become more noticeable, computer vision-based Incorporation with machine learning techniques are used to detect these symptoms.

III. COMPUTER VISION BASED AUTISM DETECTION

Computer vision-based protocols have proved its efficiency in several fields starting from entertainment, Security surveillance, up to medical assisting frameworks. Accordingly, several researchers have contributed to design computer vision-based solution assisting in diagnosing autism and other mental developmental disorders. Several researchers contributed to design solutions to detect special symptoms of autism like engagement of attention and sharing interest as well as eye contact [14]. The proposed solutions relied on AOSI as a reference and proved its efficiency of using computer vision-based techniques to detect these symptoms.

Relying on eye tracking can also be used as a way of diagnosing autism symptoms in infants and toddlers [15]. Researchers proposed machine learning vision-based framework to detect the likelihood of being autistic based on analysing recorded videos between infants and adults. The proposed solution was tested using Multimodal Dyadic Behaviour (MMDB) database about 160 videos of child-grown-up interaction proving the efficiency of relying on eye tracking as a way of expecting autism. As missing development milestones is a clear symptoms of autism, Prego *et al* . [16] aimed to detect missing a milestone of development in children using vision-based techniques in cooperation with machine learning classification algorithms.

A. Human Activity Recognition Phases

Stereotypical Motor Movements (SMM) in autistic children can be recognized from recorded videos. S. S. Rajagopalan *et al* . [3] have conducted and annotated Self-Stimulatory Behaviours dataset (SSBD) to be used in further research. The dataset includes three categories of videos based on the type of stereotypy behaviour which are arm flapping, head knocking and body rolling. Computer vision based self-stimulatory behaviours detection goes through the main three phases of computer vision based human activity recognition (HAR) stages. The three stages are image segmentation, feature extraction and human body description, and activities patterns classification .

1) Image Segmentation

Image segmentation is the procedure of splitting image into set of meaningful zones based on set of properties. Segmentation techniques fall into two classes based on used algorithms which are background detection segmentation and objects detection segmentation. Background detection techniques is the most known method of segmentation specially with static cameras. Temporal or spatio-temporal information are used to differentiate between objects and their background.

2) Feature Extraction and Human Body Descriptions

After segmentation phase, meaningful features are extracted from the image. Several types of features can be extracted from images like colour, texture, or gestures of human body. Feature extraction algorithms are categorized into three classes which are universal, regional, and semiotic based techniques. Universal features describe the whole image, however, regional describe small regions of pixels in the image. Global features proved their efficiency in several applications involving pedestrian detection [17]. Global features miss specific details in the objects which in turn cannot be used in understanding specific actions that require deep details. Local feature descriptors can extract tiny details in image. There are two types of local feature extractors; grid and point based extractors. Grid based feature extractors aims to split the image in grid form and extract the features from each local cell. Point-based feature extractors aim to find remarkable pixels in the image. Semantic-based features are used to describe more complicated images like images including human body posture. Semantic-based features involve several types of features like shape, appearance, pose, and motion-based features.

1 Shape features:

Shape features describe objects in the same way that human brain does. Human brain recognizes objects in terms of set of edges describing the overall shape [18]. Shape features are preferred by many researchers because of its simplicity and proved efficiency an object detection.

2 Appearance-based features:

Appearance-based features describe objects by defining their colours, texture, or material. Appearance-based strategies are utilized for tracking people in recordings. Although that appearance-based features competes shape-based features under occlusion, they are sensitive to illumination and cloth variation.

3 Pose-based features:

Pose-based features that describe human body location and direction respecting to scene. As activities are sequence of poses performed by person, activity can be distinguished by pose features. However, one activity can be performed in various styles making the identification of this activity more challenging. Human body pose estimation methodologies are divided into two subfamilies

which are supervised and unsupervised methodologies [19]. In machine learning, supervised learning relies on prior information to recognise new data. Supervised pose estimation models are divided into direct and indirect based models. Direct models-based approaches can expect several poses by applying geometry and kinematics on human body model. Indirect models rely on 3D static human body models to estimate some poses. Indirect models lack for human body details making pose estimation more challenging. In unsupervised pose or known as model free estimation no prior models are used. However, higher level features in form of indicators are utilised to expected human body pose. Model free pose estimation is more robust against recording orientation; however, understanding numerous poses is the main challenge which requires large amount of training data.

4 Motion features:

Motion features are the description way of objects transformation. Temporal variation and optical flow are utilized two describe motion features of objects. Temporal variation estimates motion by subtracting high level features from two successive frames as shown in equation 1. To design strong descriptor, mixture of feature extractor can be combined to extract different types of information from one scene.

$$|f_t(x,y) - f_{t-1}(x,y)| > T \ \&|f_{t+1}(x,y) - f_t(x,y)| > T \quad (1)$$

(As t, t-1, and t+1 are current, prior, and next time instant, while f (x, y) represents the features in frame.)

3) Activities pattern Classification

Activity recognition relies on using proper classification techniques to assign each extracted body or series of them to specific predefined class of actions or activities. HAR includes two main steps: activity recognition and activity pattern discovery. Activity recognition identifies human activities accurately according to a model that define this activity. While activity pattern discovery is more about discovering activity patterns. Activity pattern discovery is usually used in security surveillance applications as suspicious behaviour pattern should be detected accurately. Activity recognition and activity pattern discovery are used together to produce more robust human activity recognition frameworks. Activity performed by multiple subjects is more challenging to be recognised using human activity recognition systems. The most basic step in complicated human activity recognition systems is to understand action performed by one subject. Successful action understanding supports in more complicated scenarios. As activities are sequence of gestures or body poses, they can represent it in a form of action patterns, classification algorithms are required to distinguish between different these patterns. Machine learning, convolutional neural networks, and deep learning have proved their efficiency in differentiating between different activities patterns in several fields.

Computer vision-based physical activity recognition system was proposed in [20] by Lingfei Mo *et al* . Global features were extracted, and direct human body model was constructed in a skeleton form. Deep learning model was then constructed to differentiate between twelve physical activities in CAD-60 dataset achieving average accuracy of 81.8 %.

B. *Human Body Pose Estimation and Skeleton Representation*

Vision based pose estimation is the procedure of identifying the location, posture, and direction of human in a scene. Several researchers have contributed to design machine learning models able to identify human body poses and represent them in effective forms. Pose estimation has been used in several fields including medical assessment [21], entertainment [22] and sports. Although that human pose estimation and skeleton representation are considered single action recognition, but it helped in understanding complicated behaviours and activity patterns. Pose estimation is categorized into two classes based on the type of the expected output which are 2D and 3D. 2D methods intends to localize set of blobs on the human body indicating set of body locations in (x,y) coordinates. However, 3D pose estimation aims to construct 3D model of the human body in (x, y, z) coordinates. Another categorization of the pose estimation techniques is top down and bottom-up estimation. Top-down estimator detects the whole human body first then estimates the body parts location. While bottom-up estimator aims to detect the body parts first then calculate the overall body pose. Pose estimation is a complex process due to several challenges like the numerous positions, occlusion, and illumination. Convolutional neural networks have added great value to pose estimation techniques making the process more faster, accurate and robust .

C. *Pose Estimation Evaluation Metrics*

Several evaluation metrics can be applied to evaluate 2D human body pose estimation models. These evaluation metrics include, but not limited to:

1) *Percentage of Correctly Predicted Key-points-PCK* [23]:
human body part is accurately discovered if the difference among the discovered part and actual position of aimed key-point less than specific threshold. The threshold can either be 0.2 or 0.5 based on the used reference. The threshold is set to 0.5 if the reference is the head bone link; and is set to 0.2 if the reference is the torso diameter. The higher the PCK value the more accurate human body pose estimation model.

2) *Percentage of Correctly Detected Joints – PDJ*:
human body part is accurately discovered if the fraction among the discovered part's position and actual position of aimed part is less than 0.2 x aimed torso diameter. The higher the PDJ value the more accurate human body pose estimation model.

3) *Percentage of Correctly Predicted Parts – PCPP* [24]*:*
Human body part is accurately discovered if the difference amongst the discovered part's position and actual position is less than half length of aimed part. The higher the PCPP value the more accurate human body pose estimation model.

4) *Mean Average Precision of Key-points (MAP)* [23] *:*

MAP is an evaluation metric used when ground truth of joints is only locations with no bounding box specifying each person boundaries. APK considers predicted joint location to be true if it lies within a threshold of actual location of aimed joint.

D. Pose Estimation and Skeleton Representation Models

Neural network based human body pose estimation and skeleton representation model requires no prior explicit feature extraction and representation as neural network can capture the whole scene as an input. Accordingly, it accelerates the activity learning process, as they capture high-level features [25]. Next, set of recent pose estimation and skeleton representation frameworks will be covered.

1) *DeepPose* [26]

Toshev *et al* . [26] have proposed deep neural network framework named DeepPose for human body joins localization and Pose estimation. DeepPose is a regression based deep neural network that transfer labelled input images and to settle joints with X&Y coordinates. DeepPose model architecture consists of seven convolutional layers of AlexNet [27] plus input layer with size of 220x220. The total number trainable parameters of DeepPose is 40M with 0.0005 learning rate. DeepPose was evaluated using two datasets, which are Frames Labelled In Cinema (FLIC) [28] and Leeds Sports Dataset[29]. Human body in this dataset was labelled by fourteen joints. Two evaluation metrics were used which were (PCPP) and (PDJ) achieving best accuracy of 69%.

2) *Convolutional Neural Network Based-Efficient Object Localization* [30]

J. Tompson *et al* . [30] have proposed convolutional neural network model for body posture guesstimate and skeletal representation. J. Tompson's proposed model architecture was inspired from [31] with some extensions and modifications. The proposed model is multi-resolution convolutional neural network that intends to find the overlapping amongst the two inputs and produce heat-map of the human body. The model inputs is two samples of the same image with different resolutions; the first is 320x240 pixels while the second is 256x256 pixels. The proposed model architecture has two branches of layers dealing with each input size. J. Tompson *et al.* have used two datasets of labelled images to evaluate their proposed model which are FLIC [28] and MPII [32] datasets. J. Tompson *et al*. have applied PCK as their evaluation metric achieving better accuracy than regression-based neural network. The best accuracy achieved for the whole-body joint was 82%.

3) OpenPose [33]

OpenPose is an open-source CNN framework for human 2D pose estimation that can effectively detect multiple human body poses in real time. OpenPose model architecture comprises couple phases, the first phase compromises the initial ten layers of VGGNet and they are used to construct a feature vector of the input image. The second stage is dual branch CNN in which the first branch expects the part of affinities fields (PAF), while the second branch aims to calculate the confidence map of body joints. OpenPose is trained on MPII [32] and COCO [34] datasets. MPII dataset annotation of human body defining 25 key points, while COCO dataset defines 18 key points. The input to OpenPose model is an image, while the result is group of joints. Each joint consists of three values which are X and Y coordinates and certainty factor. OpenPose comprises three blocks of pose estimation which are; first: body and foot, second: hand, third: face detection. The evaluation metrics used is MAP achieving 79.7% on MPII dataset.

E. Human Activity Recognition Datasets for Autism Detection

Several publicly available datasets of autistic children were proposed for research aims. Researchers usually use these datasets to evaluate their proposed solution. Autism related publicly available datasets can radically change the way of diagnosing and treating this disorder. Some of these datasets are: A Dataset of Wild Self-Stimulatory activities of autistics [3], Autistic Children Screening Dataset [35] , A Dataset of eye tracking of autistic children [36], The Multimodal Dyadic Behaviour dataset [37], Autism Spectrum Disorder Detection Dataset [38], and The DE-ENIGMA Database . Although that these data sets were used in several research directions, but there is still a need for large population ASD dataset including differences in age gender and autism symptoms.

F. Human Activity Regonition Evaluation metrics

One metric of calculating the accuracy of classification algorithms is done by calculating factors of correctly detected human activity recognition with respect to the total labelled human activity patterns. Several calculations are used to assess the correctness of classification techniques. Some are expected to be maximized while others should be minimized. Here we will list some of these calculations. Metrics used to evaluate the accuracy of classification algorithms, include:

True Positive Ratio (TPR): is the proportion among true positive detection and positive labelled samples, it should be maximized.

$$TPR= TP/P= TP/(TP+FN) =1-FNR \quad (4)$$

True Negative Ratio (TNR): is the proportion among true negative detection and negative labelled samples, it should be maximized.

$$TNR= TN/N=TN/(TN+FP) =1-FPR \quad (5)$$

False Positive Ratio (FPR is the proportion among false positive detection and negative labelled samples, it should be minimized.

$$FPR= FP/(TN+FP) = FP/P=1-TNR \quad (6)$$

False Negative Ratio (FNR): is the proportion among false negative detection and positive labelled samples, it should be minimized.

$$FNR= FN/(TP+FN) = FN/P=1-TPR \quad (7)$$

False Discovery Ratio (FDR): is the proportion among false positive detection and the summation of true and false positive detection, it should be minimized.

$$FDR= FP/(FP+TP) =1-PPV \quad (8)$$

False Omission Ratio (FOR): is the proportion among false negative detection and the summation of true and false negative detection, it should be minimized.

$$FOR= FN/(FN+TN) =1-NPV \quad (9)$$

Positive Prediction Amount (PPA): is the proportion among true positive detection and the summation of true and false positive detection, it should be maximized.

$$PPV = TP/(TP+FP) = 1-FDR \quad (10)$$

Negative Prediction Amount (NPA): is the proportion among true negative detection and the summation of true and false negative detection, it should be maximized.

$$NPV = TN/(TN+FN) = 1-FOR \quad (11)$$

Accuracy (ACC): is the ration between the summation of true positive detection and true negative detection and the total of samples, it should be maximized.

$$ACC = (TN+TP)/(N+P) = (TN+TP)/(FN+FP+TN+TP) \quad (12)$$

G. Proposed Computer vision based autism signs detection framework

Based on the review presented in this paper and after observing all the techniques; a computer vision-based autism signs detection framework is proposed. The proposed framework will go through stages of supervised human activity recognition systems. As the proposed system is expected to be applied on recorded videos, set of pre-processing operations are required at the very beginning stages of this proposed solution. Expected pre-processing operations are noise removal, video stabilisation and features improving. Then, each frame will be treated as a single image that we aim to find the subject in. Human body features are expected to be extracted using neural network based human body poses estimator. OpenPose [33] is the proposed neural network pose estimator that extract human body joints and represent them in a skeleton form. Since activities are a sequence of postures or poses, behavioural patterns can be tracked from extracted human body joints. Finally, neural networks will be utilised to classify behavioural pattern into their corresponding classes.

IV. CONCLUSIONS

Autism is a complex disorder with many variants and different severity levels. Detecting autism using genetic analysis is still unclear. Detecting autism using EEG-based investigation or fMRI, needs special arrangements and conditions not comfortable for autistic children. Detecting autism through wearables encourages anxious feelings in autistic children and can lead to extra SMM behaviours, a fatal drawback. however, computer vision techniques can preliminarily help families to detect red flags of such mental and developmental disorders, which in turn will lead to faster diagnosis and treatment of these disorders. Designing a human activity recognition framework that can recognize multiple activities is much useful than detecting single behaviour. Adding to that, detecting SMM of autistic child within several people in the same scene will help in recognizing autism red flags even if there is no human intervention.

To detect and recognize autistic self-stimulatory behaviours, the expected computer vision-based approach is expected to go through three phases; extraction of autistic child body from input video, pose estimation and skeleton representation, and recognition of repetitive behaviour and state whether it is autistic symptom or not. Since several poses are expected to be found with various viewpoints, deep learning models are expected to produce satisfying outcomes in extracting autistic child's body. Although computer vision-based approaches are not supposed to supress medical professionals work, they allow capturing behavioural patterns in a non-intrusive and continuous way over time and has low implementation costs.

REFERENCES

[1] C. for D. C. and Prevention, "Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators; Prevalence of autism spectrum disorders—Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008," *MMWR Surveill Summ*, 2012.

[2] G. Lösche, "Sensorimotor and Action Development in Autistic Children from Infancy to Early Childhood," *J. Child Psychol. Psychiatry*, vol. 31, no. 5, pp. 749–761, 1990.

[3] S. S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 755–761, 2013.

[4] K. S. L. Lam and M. G. Aman, "The repetitive behavior scale-revised: Independent validation in individuals with autism spectrum disorders," *J. Autism Dev. Disord.*, vol. 37, no. 5, pp. 855–866, 2007.

[5] C. Lord *et al.*, "The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Dev. Disord.*, vol. 30, no. 3, pp. 205–223, 2000.

[6] S. E. Bryson, L. Zwaigenbaum, C. McDermott, V. Rombough, and J. Brian, "The autism observation scale for infants: Scale development and reliability data," *J. Autism Dev. Disord.*, vol. 38, no. 4, pp. 731–738, 2008.

[7] American Psychiatric Association (APA), *Diagnostic and statistical manual of mental disorders*, 5th ed. American Psychiatric Association (APA), 2013.

[8] A. A. Baumeister and R. Forehand, "Stereotyped Acts," *Int. Rev. Res. Ment. Retard.*, 1973.

[9] Y. Shen *et al.*, "Clinical Genetic Testing for Patients With Autism Spectrum Disorders," *Pediatrics*, 2010.

[10] F. Albinali, M. S. Goodwin, and S. S. Intille, "Recognizing Stereotypical Motor Movements in the Laboratory and Classroom : A Case Study with Children on the Autism Spectrum," *Proc. Int. Conf. Ubiquitous Comput.*, pp. 71–80, 2009.

[11] G. Chanel, S. Pichon, L. Conty, S. Berthoz, C. Chevallier, and J. Grèzes, "Classification of autistic individuals and controls using cross-task characterization of fMRI activity," *NeuroImage Clin.*, 2016.

[12] U. Großekathöfer *et al.*, "Automated Detection of Stereotypical Motor Movements in Autism Spectrum Disorder Using Recurrence Quantification Analysis," *Front. Neuroinform.*, vol. 11, no. February, 2017.

[13] J. Hashemi *et al.*, "Computer Vision Tools for Low-Cost and Noninvasive Measurement of Autism-Related Behaviors in Infants," *Autism Res. Treat.*, vol. 2014, pp. 1–12, 2014.

[14] D. Wen, C. Fang, X. Ding, and T. Zhang, "Development of recognition engine for baby faces," in *Proceedings - International Conference on Pattern Recognition*, 2010, pp. 3408–3411.

[15] J. M. Rehg *et al.*, "Decoding Children's Social Behavior," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3414–3421.

[16] P. Kohli, J. Rihan, M. Bray, and P. H. S. Torr, "Simultaneous segmentation and pose estimation of humans using dynamic graph cuts," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 285–298, 2008.

[17] R. Benenson, M. Mathias, T. Tuytelaars, and L. Van Gool, "Seeking the strongest rigid detector," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[18] J. De Winter and J. Wagemans, "Contour-based object identification and segmentation: Stimuli, norms and data, and software tools," *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 4, pp. 604–624, 2004.

[19] T. B. Moeslund, A. Hilton, and V. Kr??ger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, no. 2-3 SPEC. ISS., pp. 90–126, 2006.

[20] L. Mo, F. Li, Y. Zhu, and A. Huang, "Human physical activity recognition based on computer vision with deep learning model," in *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, 2016, vol. 2016-July.

[21] C. Doignon, F. Nageotte, B. Maurin, and A. Krupa, "Pose Estimation and Feature Tracking for Robot Assisted Surgery with Medical Imaging," *Lect.*

Notes Electr. Eng., vol. 8 LNEE, pp. 79–101, 2008.

[22] S. R. Ke, L. J. Zhu, J. N. Hwang, H. I. Pai, K. M. Lan, and C. P. Liao, "Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming," in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 489–496.

[23] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[24] V. Ferrari, M. Marín-Jiménez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008.

[25] A. Yao, J. Gall, G. Fanelli, and L. Van Gool, "Does Human Action Recognition Benefit from Pose Estimation?," *Procedings Br. Mach. Vis. Conf. 2011*, 2011.

[26] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May-2017.

[28] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.

[29] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *British Machine Vision Conference, BMVC 2010 - Proceedings*, 2010, no. i, pp. 1–11.

[30] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using Convolutional Networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 648–656.

[31] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in Neural Information Processing Systems*, 2014.

[32] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014.

[33] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.

[34] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[35] "Autistic Spectrum Disorder Screening Data for Children Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++.

[36] "A Dataset of Eye Movements for the Children with Autism Spectrum Disorder." [Online]. Available: https://zenodo.org/record/2647418#.XT8jDugvPDc.

[37] "The Multimodal Dyadic Behavior (MMDB) dataset." [Online]. Available: http://www.cbi.gatech.edu/mmdb/.

[38] "Autism Spectrum Disorder Detection Dataset." [Online]. Available: https://pavis.iit.it/datasets/autism-spectrum-disorder-detection-dataset.