Deep Learning Based Recognition of Meltdown in Autistic Kids

Sindhoor Preetham P.V, Feba Thankachan George, Kiran George, and Abhishek Verma.
Department of Computer Engineering
College of Engineering and Computer Science
California State University, Fullerton, California 92831
sindhoor@csu.fullerton.edu

***Abstract*— Children with autism often experience sudden meltdowns which not only makes the moment tough for the caretakers/parents but also make the children hurt themselves physically. Studies have discovered that children with autistic spectrum disorder exhibit certain actions through which we can anticipate mutilating meltdowns in them. The objective of our project is to build a system that can recognize such kind of actions using deep learning techniques thereby, notifying the caretakers/parents so that they can get the situation under control in lesser time. Using deep learning RCNNs, we can train the system faster yet reliable because unlike all the machine learning algorithms, deep learning algorithms are more efficient and have more scope into future. We have trained a classifier on images that are gathered from videos and reliable internet sources with most predictive gestures, through which we can detect the meltdowns more precisely. We have trained a model that validated the accuracy by ~93% which is accompanied by a loss/train classifier with a minimal 0.4% loss. Functional testing was done through feeding the deep neural network with chosen actions performed by five individuals that resulted in an accuracy of ~92% in all cases, which can assure the real-time usage of the system.**

***Index Terms*— Autistic spectrum disorder; deep learning; Convolution Neural Network; training classifiers; inference; Graphics Processing Unit.**

I. INTRODUCTION

Autism is a neurodevelopmental impairment prevails in the children affecting their ability to communicate, interact socially and cognitively. According to Autism Society of Minnesota, 1 in 68 individuals has Autism [1]. These people may experience frequent meltdowns. The sudden meltdown of autistic children is a nightmare for the parents and caretakers [2]. Researches show that many autistic children show some spasms of distress before a meltdown, sometimes called as a 'rumble stage' [3][4]. These individuals may begin to exhibit involuntary cues of anxiety before a meltdown such as rocking, wrist biting, self-scratching, kicking, banging their heads or even lashing out violently [5][6]. Thus, if we could recognize and notify such behaviors using deep learning before it relapses, these can become an aid for caregivers [7].

There are different techniques to detect these non-verbal behaviors using neural networks such as Multilayer Perceptron (MLP) or a Deep Convolution Neural Network (DCNN) which uses back propagation algorithm where the input image is repeatedly given to the network, and a loss is computed by comparing the desired output with the neural network output. This estimated loss is fed back to the network to minimize the error by adjusting the weights with each iteration, producing an intended output. This procedure is called as "training." With the support of multiple Graphics Processing Unit (GPU), training rate could be accelerated.

In this paper, we present a novel and efficient technique to recognize Autistic meltdowns in children using deep learning. In the past, many deep learning architectures have been used successfully for various computer vision applications such as face recognition, object recognition, audio recognition, etc. However, deep learning in the field of behavior identification is not entirely developed. The objective is to classify the meltdown behaviors expressed by the children in real-time environment captured using a camera. The architecture used for behavior classification is Convolution Neural Network (CNN) present in Caffe DIGITS®. The trained model retrieved from DIGITS® is deployed using mobile GPU for real-time recognition and classification. [8] A prototype is developed in the lab that could recognize an autistic kid's sudden behavior change. The recognized behavior including covering ears and covering face was taken as two instances for the detection of meltdowns in the autistic children using deep learning. The prototype developed has an alert mechanism using which the caretaker or parents of autistic children is notified using a buzzer in case of any meltdown symptoms occur.

This paper is organized as follows: Section II of the paper has some of the related works in this area; Section III elaborates the design and implementation of the system; Section IV comprises the analysis as well as the results provided in detail; and finally, Section V present conclusion and future works.

978-1-5090-4881-6/17 $31.00 © 2017 IEEE
DOI 10.1109/ICHI.2017.35

II. RELATED WORKS

Orrawan et al. [9] provided a concept of detecting the repetitive motion in autistic spectrum disorder by correlating two adjacent frames to identify the relation between the frames. Self-similarity between adjacent frames of videos is measured mainly for classifying repetitive gestures.

Nastaran et al. [10] use a combination of CNN and multi-sensor accelerometer signal to detect stereotypical motor movement thereby differentiating Autistic kids from healthy children.

Tanaya et al. [11] provide the concept of sparse representation using learned dictionaries for human action detection. For large data, sparse representation is a successful tool. It is done by creating a set of spatial-temporal descriptors of small dictionary elements.

Pavlo et al. [12] presented a three-dimensional recurrent convolution neural network that detects and classify the dynamic hand gestures simultaneously. They use connectionist temporal classification to recognize labels of classes from gestures in the unsegmented input.

Previous researches present many techniques for behavior recognition, but researches on autistic meltdown recognition are hardly explored.

III. METHODOLOGY

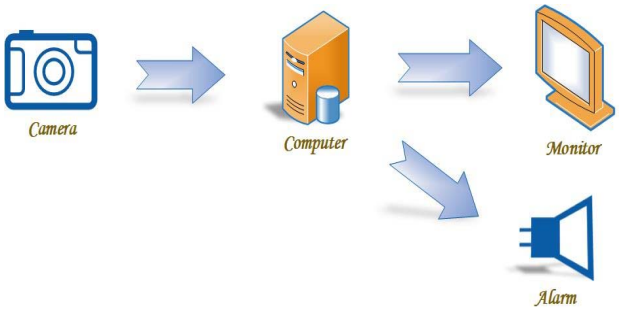

Fig. 1. System Block Diagram

Lucid requirements of the project made us choose one of the famous software development models which are 'waterfall' model. Process of building the system is described below

A. Requirements gathering

We have done our preliminary reading and research in the field of autism and its adverse effects in kids. Based on the previous work done on autism we have chosen certain behavioral gestures in children which can assure the happening of sudden meltdowns [13] [14].

B. Requirements analysis

A hard ground work is done on the collected gestures to refine them for a more reliable recognition and found that few of those gestures in their descending order of consistency are 'covering ears,' 'covering face,' 'biting hands' and 'flapping hands' [15] [16].

Transfer learning: Using the previously developed training knowledge in a different yet related problem is defined as transfer learning in machine learning semantics [17]. A small number of top layers in the neural network are trained using the transfer learning which induces the previous learning knowledge into those layers thereby giving the neural network essential features of problem-solving from the prior knowledge that is related to the current problem.

CNN: Convolutional Neural Network refers to a feed-forward artificial neural network which has its extensive use in image recognition and analysis. CNN uses multilayers for the purpose and reduces the preprocessing overhead which is inspired by regular neurons present in the human brain. CNN is mainly built on the convolution layer which has small receptive fields.

GoogLeNet: GoogLeNet is a convolutional neural network which can be used as a transfer learning supplement in our current scenario because it was trained on 'ImageNet' database which contains around 1000 classes of images that use a massive number of 14 million images for training [17]. The presence of the inception module makes it much faster yet compact for the image classification as it does the convolution and pooling at the same time and integrate them.

C. Design

We have first gone through the process of collecting raw pictures and videos that contained the desired actions. This data was acquired from various databases and reliable internet sources, which is followed by the data cleaning process that crucially involved filtering the database from deplorable images and frames thereby gaining a more feasible set of images for training which is further processed by cropping and annotating using various image processing tools.

Having a plausible database ready, selection of robust machine learning algorithms has been carried out with extensive research in the artificial intelligence area. Several machine learning algorithms were put into comparison for potential measures such as compactness, robustness, reliability, etc., Since deep RCNN design being very suitable for the context and very much future proof, has been chosen for training. RCNN is redesigned with an extra layer of 'loss/accuracy' classifier which compares the loss and accuracy at every layer of the neural network and adjusts the learning rate appropriately, resulting in a deep recurrent neural network that is almost 30% accurate more than the standard RCNNs.

Training was done using Caffe where the prepared database was fed to the redesigned RCNN which is backed up by 'GoogLeNet' as a pretrained model. NVIDIA® TX1 was used for training which has three instruction clock cycle as well as a 64-bit architecture that combinedly delivers extra processing power for deep learning neural network architectures which resulted in a typical duration of 30-60 minutes for successfully training of a classifier set. Neural network models were trained for all gestures with two classes, namely positive and negative per each gesture.

All the selected gestures were trained as separate models, each resulting in a validation accuracy of 93% and training loss of 0.43%. We also had models with potentially more than two gestures which resulted in more classes and less accuracy percentages.

Dataset for the gesture 'covering ears' with two classes(positive set and negative set) were taken, where we had approximately thousand pictures per class with a validation set of around 15-25%. Thus created data set is fed to the neural network model for training the system for the gesture 'covering ears' (statistical graph of which is shown in figure 2). We have observed a decreasing training loss which is accompanied by a validation accuracy of approximately 93%. The model was further developed by feeding the neural network with a pre-trained model and its corresponding weights. Thereby creating room for semi-supervised learning through which we can make the neural network more intelligent and robust.

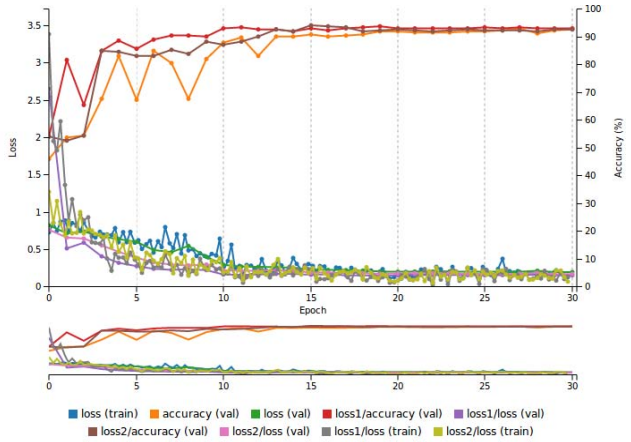

Fig. 2. Graphical data for covering ears training model

The figure 3 represents the graph of the gesture 'covering face' training model. This result could be further improved by training with more explicit images and by fine tuning the neural network models. Although the identification was successful, the challenges involved were clustered background, viewpoint blockages, and database collection, etc.

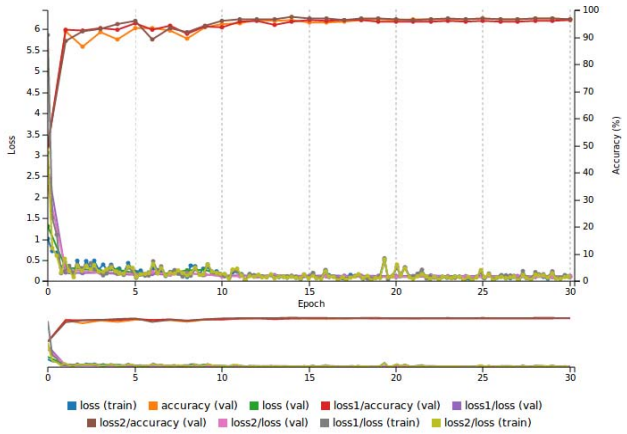

Fig. 3. Graphical data for covering face training model

We also have a model trained with all the three gestures, which were clustered into three separate datasets with 1000-1200 images per class. However, the validation accuracy percentage, in this case, is noticeably low when compared with the rest of two neural network models. Figure 4 shows the design process flow of an entire system from collecting images to the classification for the gesture 'covering ears'.

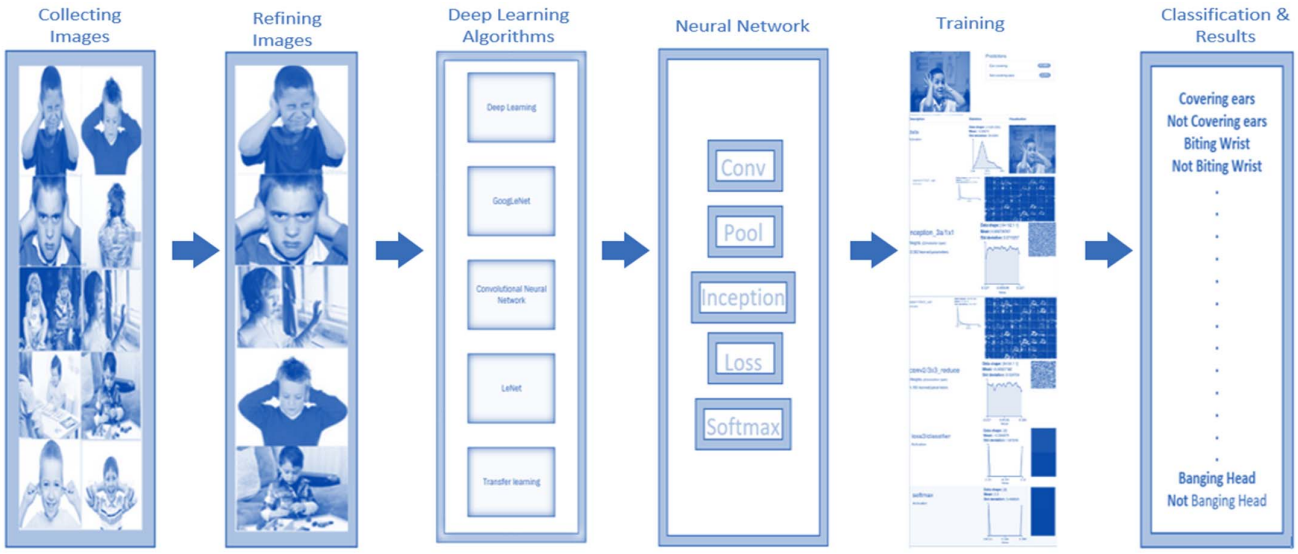

Fig. 4. Design process flow of the entire system

Further, the trained neural network was deployed through an API that runs on TensorRT which communicates with code modules that are developed in C++ and Python. The whole model libraries that contain the necessary files such as Caffe binaries, prototxts, labels, etc., were used for the deployment through the C++ code blocks. Usage of C++ coding allowed the parameter referencing and made class scopes more flexible. TensorRT being chosen as a library has increased the whole performance of the system by 100% (when compared to cuDNN). Any camera devices that produces RGB frames can be used as capturing device that streams content to the neural network model through TensorRT where each frame in every 30 will be classified into available label categories, and the result is continuously displayed on the screen along with the accuracy values.

Such trained model was tested on five individual human subjects who performed advised actions in front of the camera, in the case of which, for every gesture and person we have achieved an accuracy percentage of ~92 which makes the system fit for real time usage. However, it requires more training data at the beginning which is not only harder to acquire but also makes it hard to manage as a database, but hopefully, this issue should be addressed with futuristic approaches such as one-shot learning. One-shot learning transfers knowledge by model parameters, sharing features and contextual information through which very few images for training the deep neural networks are used which in turn reduces the size of training image data set and outraging usage of disk space in a significant amount.

An alarming mechanism in our case is also included as a part of the additional hardware to the system which alerts the authorized person of the escalating situation through a buzzer. As shown in figure 5 the buzzer is operated once the positive gesture is identified and it is triggered using General Purpose Input Output (GPIO) pins available on the system which can also be potentially extended to smart devices with an alert notification in an appropriate manner. Generally, these smart alert mechanisms may include notifications through mobile apps, etc.,

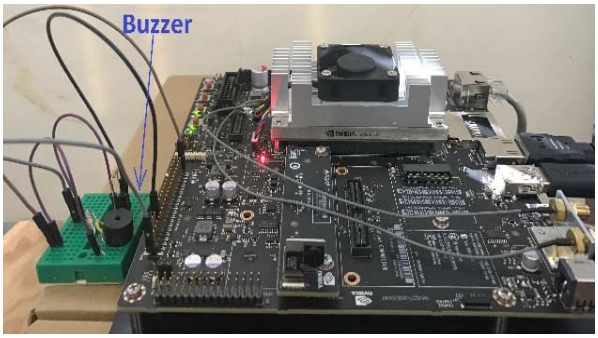

Fig. 5. Additional hardware for alerting mechanism.

Such trained and tested model is used as a pre-trained model for further training through which the input dataset can not only be minimized in a significant rate but also a better percentage of accuracy can be attained since the upcoming model's first layers are induced by the knowledge of this model.

IV. Experiments and Results

Around 2000 data images were collected for each class of behavior, and ~92% classification accuracy was achieved for training alone in NVIDIA DIGITS®. Table I represents the training classification results of five random images used to test the accuracy and the time taken to obtain the results. Training time taken by NVIDIA DIGITS® for 2000 images in class was approximately around 10-15 mins.

TABLE I. Training Results for Positive Class

Image	Time Taken (sec)	%Accuracy
1	8	89
2	7	92
3	4	91
4	3	93
5	2	96
Average	4.8	92.2

In figure 6, 'covering ear' gesture is tested on the custom-net and accuracy of approximately ~92% was achieved. The images were trained using 22 layered customized GoogLeNet consisting of a chain of convolution layer, pooling layer and response normalization layers with different weights performed parallelly and combined later. The data layer of the GoogLeNet has the data shape represented as [3,224,224] where 3 is the number of channels and 224x224 represent the size of the image. Convolution layer 7x7 gets the input from data and uses the one 7x7 convolution filter with a padding of 3 and stride of 2. In between the network, 1x1 convolution layer was used to decrease the size of input given to the convolution layer with larger weights [17]. Two max pool layers were used to minimize the spatial dimensions. After passing through all these layers, we will get the final classifier.

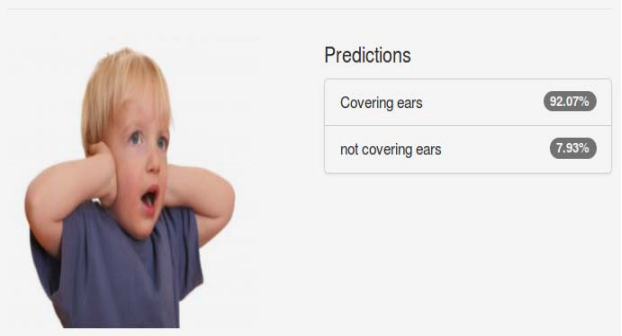

Fig. 6. The test result for positive input at training phase.

In the same way figure 7, represent the 'not covering ears' gesture which passed through all the layers in GoogLeNet to achieve an accuracy of approximately 92%. This custom model was used to classify the gestures given from a real-time

video capturing device and achieved an inference accuracy of ~92% shown in the figure 8.

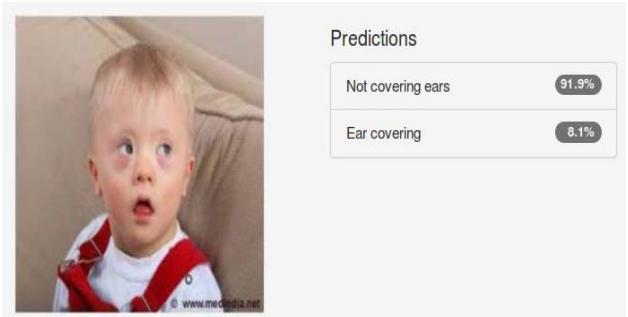

Fig. 7. The test result for negative input at training phase.

Experiment I.

For experiment I, 'covering ears' was taken as a test condition since it is one of the most commonly exhibiting behavior by children with autism. The figures 8 and 9 show ~92% test result when performed on a healthy subject. For this round of experiment, we used an HD camera with a recording frame rate of 30 FPS. The test was conducted on four healthy subjects for covering ears and covering face. The average time taken to correctly identify the behavior was ~5 sec with an accuracy of 91.34% for covering ears. Table II and III provide the results of inference for covering ears and covering face.

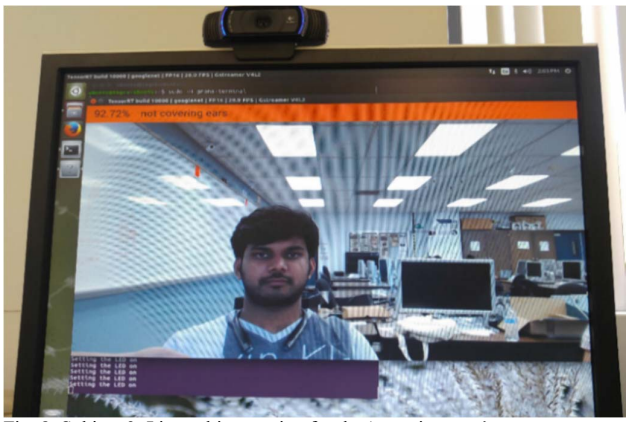

Fig. 8. Subject 3: Live subject testing for the 'covering ears' gesture.

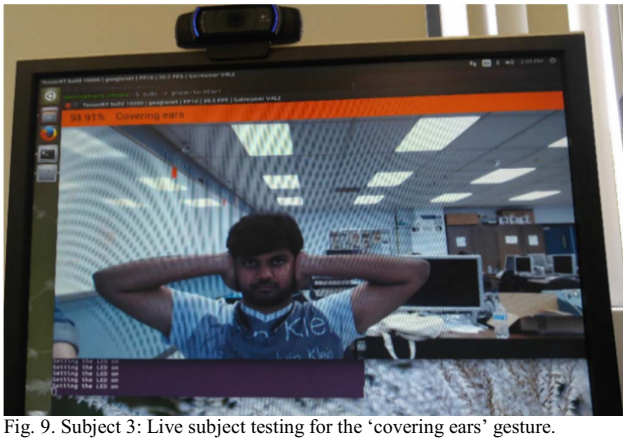

Fig. 9. Subject 3: Live subject testing for the 'covering ears' gesture.

TABLE II. INFERENCE TEST RESULTS FOR COVERING EARS

Subject	**%Accuracy**	**%Accuracy**
	Positive	***Negative***
1	91	92.33
2	89	91.67
3	91.45	93.23
4	93.91	92.72
Average	91.34	92.49

Experiment II.

For experiment II, 'covering face' behavior was taken as a test condition which was trained and tested. The figures 10 and 11 shows the test results for 'covering face' behavior performed by a healthy subject. As shown in Table III, an acceptable result was achieved for the first two healthy subjects. The average time taken to correctly identify the behavior was ~5 sec with an accuracy of 88.49% for covering face. However, comparatively less percentage was achieved for the last two subjects due to the lack of more robust dataset and clustered background.

Fig. 10. Subject 4: Live subject testing for the 'covering face' gesture.

Fig. 11. Subject 4: Live subject testing for the 'covering face' gesture.

TABLE III. INFERENCE TEST RESULTS FOR COVERING FACE

Subject	%Accuracy	%Accuracy
	Positive	***Negative***
1	93.63	93.34
2	92.72	90.65
3	80.91	92.34
4	86.72	93.63
Average	88.49	92.49

An alert mechanism, buzzer, was operated parallelly without any delay whenever a positive class is identified that indicates the meltdown behavior.

The present custom model can further be used for training another model in future where we can use this model as a pre-trained model. We are not only reducing the input craving of the neural network for training images but also increasing the accuracy of the future models in a noticeable way by injecting the prior knowledge into them. This process of transferring the knowledge from a model to another model with the related problem is referenced as transfer learning, [18] about which was talked in detail in the 'Transfer Learning' section above.

V. CONCLUSION

Using efficient image database and robust yet compact deep learning techniques we have achieved the recognition of severe autistic meltdowns through dependable non-verbal gestures with an optimum percentage accuracy of ~92. However, we have faced few challenges regarding the database creation and training the model with low loss classifiers; we have a brighter side with comparatively fast yet intelligent neural network developed. Nevertheless, we have emerging technologies such as one-shot learning which may address the problem of massive database collection. With some few further hardware additions to this system, it can be made mobile implying the usage of entire system on a single mobile board, which enables the system for various applications such as shoplifting detection, police body cameras, etc.,

In the future, we are planning to perform this on autistic children and continue training different gestures. Another area of potential enhancement would be creating a mobile application which allows parents or authorized people to get the live seed of the monitoring section and sending alerts.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation under grant no. ECCS-1150507.

REFERENCES

[1] Jon Baio, National Center on Birth Defects and Developmental Disabilities, CDC, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network," Morbidity and mortality weekly report. Surveillance summaries 63, Georgia, 2014.

[2] Elena Pattini, and Dolores Rollo, "Response to stress in the parents of children with autism spectrum disorder" Medical Measurements and Applications, IEEE International Symposium, Italy, 2016.

[3] W.Liu, Li Yi, and Xiaobing Zou, "Efficient Autism Spectrum Disorder Prediction with Eye Movement: A Machine Learning Framework," International Conference on Affective Computing and Intelligent Interaction (ACII), China, 2015.

[4] Erik Linstead, and Rene, "An Application of Neural Networks to Predicting Mastery of Learning Outcomes in the Treatment of Autism Spectrum Disorder," IEEE 14th International Conference on Machine Learning and Applications, California, 2015.

[5] Yun Jiao and Zuhong Lu, "Predictive Models for Autism Spectrum Disorder Based on Multiple Cortical Features," Eighth International Conference on Fuzzy Systems and Knowledge Discovery, China, 2011.

[6] D. Bone, M. Goodwin and S. Narayanan, "Applying machine learning to facilitate autism diagnostics: Pitfalls and promises" Journal of autism and developmental disorders, Maryland 2015.

[7] D. G. Amaral, and C. M. Schumann, "Neuroanatomy of autism," *Trends in Neurosciences*, vol. 31, no. 3, pp. 137–145, Maryland, 2008.

[8] Dorott-ee Frary, Daniel Polani, and Kerstin Dautenhahn, "On-line Behavior Classification and Adaptation to Human-Robot Interaction Styles" ACM, United Kingdoms, 2007.

[9] O. Kumdee, and P. Ritthipravat, "Repetitive Motion Detection for Human Behavior Understanding from Video Images" IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Thailand, 2015.

[10] N. Mohammadian, and Cesare Furlanello, "Applying Deep Learning to Stereotypical Motor Movement Detection in Autism Spectrum Disorders" IEEE 16th International Conference on Data Mining Workshops, Italy, 2016.

[11] Tanaya Guha, and Rabab K Ward, "Learning Sparse Representations for Human Action Recognition" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 8, Canada, 2012.

[12] Pavlo Molchanov et al., "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks" IEEE Conference on Computer Vision and Pattern Recognition, California, 2016.

[13] Steve Brown, Autism Spectrum Disorder & De-escalation Strategies, Book, USA, 2015.

[14] Chung Hyuk, Myounghoon Jeon, and Howard, "Robotic Framework with Multi-modal Perception for Physio-Musical Interactive Therapy for Children with Autism" 5th International Conference on Development and Learning and on Epigenetic Robotics, RI US, 2015.

[15] Feil-Seifer, and M. Mataric, "Robot-assisted therapy for children with autism spectrum disorders," ACM Proceedings of the 7th international conference on Interaction design and children, Chicago 2008.

[16] Suryani Ilias, Nooritawati Md Tahir, and Che Zawiyah, "Classification of Autism Children Gait Patterns using Neural Network and Support Vector Machine," Computer Applications & Industrial Electronics, IEEE Symposium, Malaysia, 2016.

[17] Olga Russakovsky et al., "ImageNet Large Scale Visual Recognition," New York, Jan 2015.

[18] Xiao Ling, Wenyuan Dai, Gui-Rong Xue, Qiang Yang and Yong Yu, "Spectral Domain Transfer Learning," China, 2008.