Child Activity Recognition using Deep Learning

Binjal Suthar, Bijal Gadhiya

Abstract: ***The human action recognition is the subject to predicting what an individual is performing based on a trace of their development exploiting a several strategies. Perceiving human activities is an ordinary region of eagerness in view of its various potential applications; though, it is still in start. It is a trending analysis area possessed by the range from dependable automation, medicinal services to developing the smart supervision system. In this work, we are trying to recognize the activity of the child from video dataset using deep learning techniques. The proposed system will help parent to take care of their baby during the job or from anywhere else to know what the baby is doing. This can also be useful to prevent the in-house accident falls of the child and for health monitoring. The activities can be performed by child include sleeping, walking, running, crawling, playing, eating, cruising, clapping, laughing, crying and many more. We are focusing on recognizing crawling, running, sleeping, and walking activities of the child in this study. The offered system gives the best result compared with the existing methods, which utilize sensor-based information. Experimental results proved that the offered deep learning model had accomplished 94.73% accuracy for recognizing the child activity.***

Keywords: ***CNN, Deep Learning, Child Activity Recognition.***

I. INTRODUCTION

Action recognition intends to recognize the activities and objectives from a series of observations and the natural conditions of at least one specialist. Robust human activity modeling and feature representation is a way to better human activity recognition. The presence of the human(s) in the picture space depicted through the feature representation in a video, as well as changes in appearance and posture are also extracted, but the two critical issues for the real situation are: interaction recognition and action detection [1]. From the perception of the data type, based on color (RGB) data and methods combining color and depth data (RGBD) [1], human activity recognition can be partitioned. This model extracts features from both the spatial and the temporal measurements, subsequently catching the motion data encoded in various adjacent casings by performing 3D convolutions. The 3D CNN model introduced in [2] extricates features by performing 3D convolutions on spatial and temporal measurements, so the movement information encoded in several contiguous frames can be caught. From input frames, it produces a few information channels, and from all channels, the entire feature representation joined as an output. The effectiveness model is evaluated on TRECVID and KTH, which gives an accuracy of 90.2%.Zhi Liu at el. has been proposed 3D based Deep CNN model that considers the straightforward position and edge data between skeleton joints and takes in spatial and temporal features and calculates component vector named JointVector. Also, the SVM is used for grouping the results from features, and then JointVector recognizes the actions by fusion [3]. By learning time-invariant and viewpoint invariant based feature representation and by using various proportions of training and testing sets for evaluation, the model gives a comparable performance on the UTKinect-Action3D and MSR-Action3D datasets with 95.5% of accuracy.

Two adaptive neural networks i.e., VA-RNN and VA-CNN [4] are presented by P. Zhang at el., based on RNN with Long Short-term Memory (LSTM) and CNN, which remove the impact of the viewpoints and enable the systems [4] for the learning of activity-specific features. The two-stream scheme named VA-Fusion provides the way to a final prediction by combining the scores of the both networks. The evaluation says that when CNNs are small, the model accomplishes large gains while sizable gains when CNNs are significant [4].

Vision-based methods are beneficial in comparison to sensor-based methods, as different camera types are there to provide more accurate data. The CNN architecture propose by H. D. Mehr and H. Polat comprises five convolutional layers, four pooling layers, and three fully connected layers [5] for the smart home activity recognition using the DMLSmartAction dataset. The softmax is used to predict the 12 classes of the dataset at the last fully connected layer [5], which gives 82.41% of the accuracy.

The temporal Convolutional Neural Networks (TCN) provides an approach to openly learn spatial and temporal representations by giving the interpretable inputs, for example, 3D skeletons for 3D human activity recognition [6]. It is used in updating the TCN in light of interpretability and how such attributes of the model are utilized to develop a ground-breaking 3D action recognition technique. It achieves comparable results on the NTU-RGBD dataset with 74.3% accuracy on CS and 83.1% on CV [6].

The method proposed in [7], which is multilayer maxout activation function based method for solving the challenges in deep neural network model training known as model parameter initialized method [7]. The action detection, spatial and temporal features encoding have been tracked using RBM, of the diverse parts of the body. These feature codes are incorporated into the global feature representation method by RBM neural network [7], and the action is recognized using SVM classifiers and achieves 92.1% accuracy.

The model proposed by Tomas and K. K. Biswas used CNN from Motion History Images (MHIs) of sampled RGB image frames to learn motion representations [8], however they used SAE for learning the different movements of skeletal joints. It can adapt low-level perceptions of joint motion orders and with the location of the image in every MHI frame, how motion changes.

Softmax function standardizes the class scores of every network in [0, 1] range, and performs late through taking a weighted mean of class scores, and gives 91.3% on MSR Daily Activity3D and 74.6% on MSR Action3D datasets [8].

Revised Manuscript Received on May 15, 2020.
* Correspondence Author

Binjal Suthar*, Department of Software Engineering from Government Engineering College, Gandhinagar, Gujarat.
Bijal Gadhiya, Assistant professor in Computer Engineering Department, at Government Engineering College, Gandhinagar.

Jayashree Onkar Nehete and D.G.Agrawal have recognized child activity of both genders up to the age of 16 to 29 months on leg, hand and waist to inhibit child accidents by 3-axis accelerometers [9]. The Coding is composed for GSM and RFID. By transmitting the ARM 7 board to transmitter end, the evaluation of body temperature, fall detection of children is done. They have identified seven types of activities. Sensor data with a synchronized video material has been recorded of almost 30 minutes. The design system presented in [9] resulted in an excellent performance of 97.8%±0.2%. For detection of activities, like head motion, body tilt, and hand motion, a waist-worn sensor can be failed [9].

Nam, Yunyoung, and Jung Wook Park have studied ten diverse subjects including 21, 23, 24, 26, 29 months-old baby girls and 16, 17, 20, 25, 27 months-old baby boys to recognize the child activities by using a barometric pressure sensor. Overall 1538 samples have been taken from one baby as training data and others are collected from ten babies used as a test set [10]. Six features are extracted from the preprocessed signals, including magnitude, mean, standard deviation, slope, energy, and correlation. The overall accuracy of the SVM is 86.2%, and of DT is 88.3% [10].

II. METHODOLOGIES AND DATASET

A. Convolutional Neural Networks (CNNs)

A convolution is a procedure which alters a function into somewhat different. The convolutions are performed to change the original function into a form that gives more data. Convolutions have been used to blur and sharpen images, and perform other operations like improve edges and stamp in image processing.

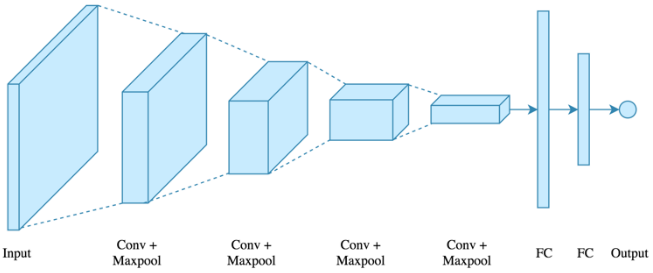

Fig.1.General Architecture of CNN

With the kernels, each layer links to the local section of the preceding layer. Furthermore, CNNs structure comprises a series of common layers. The convolutional layer is the first layer. By calculating weights that are known as kernels, every region containing feature maps are associated to the feature maps of a local section in the preceding layer. The summation of every local weight experiences non-linearity functions.

The pooling layer is the subsequent standard layer of CNNs. We can reduce the spatial measurement of the output of the convolutional layer without any modification in depth by using this layer. The advantage is that it stops the overfitting in the training procedure by decreasing computational operations such as, min, max and average. The max pooling layer has achieved reasonable results in most fields (Fig. 1).

The next is a fully connected layer in which every neuron is associated to all neurons of the preceding layer and determined scores of the classes of the datasets are set in this layer. Furthermore, to calculate the probability distribution using the labels of classes, the softmax function is utilized normally in the last most convolutional layers.

B. Dataset

Total 155 videos of crawling, running, sleeping, and walking activities of the children from YouTube, Instagram, and cameras are used for child action recognition. All videos are cropped to 360 x 360 pixels. The input to the algorithm is again resized to shape of having 224 x 224 sizes with 3 channels, which indicates RGB data is given. Out of that, we have taken 117 videos as the training dataset and other 38 videos as a testing set. We have partitioned dataset into 75% of training and 25% as validation sets. We split the data as train and test set with random state 42 and test size 0.2 while training the model. We have achieved 68,905 frames from all the training videos.

We have trained 2682 samples from all the training frames with input shape 224x224 with an RGB scale at the 30fps frame rate. While splitting the data as train and test set at 0.2 test sizes, we train the model on 2145 training samples and 537 validation samples with 10 epoch and a batch size of 128 with ADAM optimizer. The sample frames extracted from the training videos can be pictured from the images of Fig. 2. From that, it can be simply seen that we have taken almost probable scenarios for taking the videos as the dataset for the proposed work with a different objects and with a different child, age and a different positions, lighting appearance, all the situations with a different posture and gestures of each activity like sleeping in bed and sleeping in a cradle, running and walking at different speed and environment.

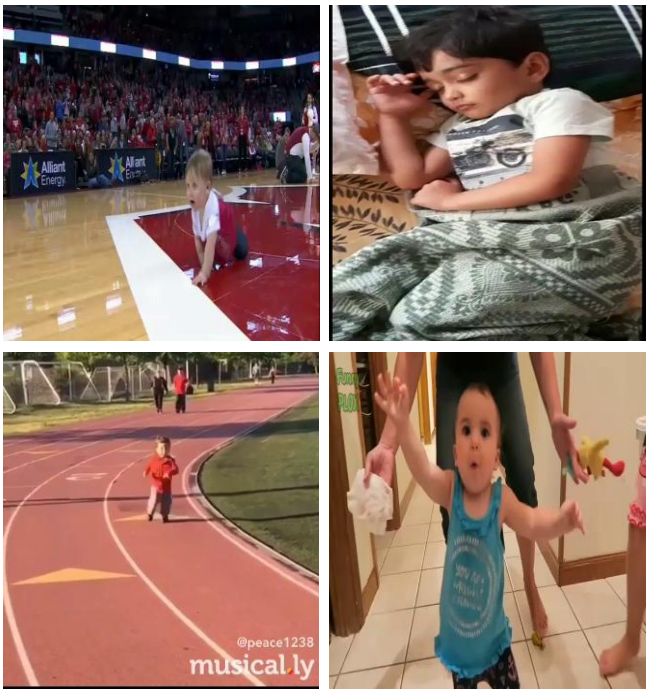

Fig. 2. Examples of child Activities

III. EXPERIMENTAL RESULTS

The Proposed architecture of a CNN comprises three convolutional layers, three pooling layers, and three fully connected layers to classify the child activity frames. In the final fully connected layer, the softmax was considered to calculate the probability of the 4 classes of activities, which includes Crawling, Running, Sleeping, and Walking. For non-linearity on the output of the convolutional and fully connected layers, The Relu activation function was utilized. 16 filters with 3×3 kernel size were executed on the inputs of size 224x224 pixels in the first convolutional layer. Furthermore, the Relu activation function is used as the conclusion of the output of the convolutional (Fig. 3). For reducing the dimensions of the output of the other convolutional layer, the max-pooling layer is applied with a 2×2 kernel size without any modification in depth size. This architecture is persistent until the softmax layer. The summary of all the layers of proposed CNN is demonstrated in the Table-I.

Table-I: Architecture Summary of the Proposed CNN

Type of Layer	No. of kernels	Kernel size	Output size
Convolutional	16	3x3	224x224x16
Max Pooling	-	2x2	112x112x16
Convolutional	32	3x3	112x112x32
Max Pooling	-	2x2	56x56x32
Convolutional	64	3x3	56x56x64
Max Pooling	-	2x2	28x28x64
Fully Connected	-	-	128x128x1
Fully Connected	-	-	32x1x1
FC with Softmax	-	-	4x1x1

To calculate the performance of the offered deep learning-based CNN architecture, the consequence of the proposed strategy was contrasted with the existing techniques which have utilized traditional machine learning strategies for recognizing child activity and deep learning methods for recognizing human actions. Table-II demonstrates the precision of the offered CNN for each activity.

Table-II: Activity Recognition of Proposed CNN for Each Activity

Activities	Precision
Crawling	1.00
Running	0.80
Sleeping	1.00
Walking	1.00

As it has appeared in Table-II, most of the activities were classified by fulfilling precision. The minimum accuracy of classification goes to running activity with 80%. The overall accuracy of the child activity classifier is 94.73%.

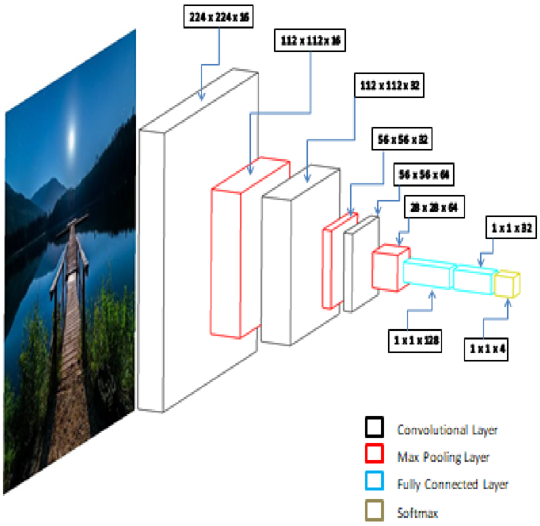

Fig.3. Visualized Concept of Proposed CNN

Table-III: Comparison with Existing Research Methods

No.	Method Used	Algorithm Used	Accuracy
1	**Based on Tri-axial Accelerometer and Barometric Pressure Sensor [8]**	SVM and DT	98.43% (SVM) 96.3% (DT)
2	**smart embedded sensor fusion and GSM technology Based [7]**	The output voltage from sensor data and RFID	97.8 %
3	**Vision based CNN method [4]**	2D CNN	82.41%
4	**Based on the waist and wrist sensor [9]**	RF, Logistic Regression, DT, K-nearest	63.8 % (waist) 41.2% (wrist)
5	**Pre-trained model VGG16 on proposed dataset**	VGG16	81.79%
6	**Proposed 2D CNN method**	2D CNN	94.73%

Furthermore, we have achieved 94.73% accuracy in the proposed method. In such manner, 95% of the supposed activity labels among all activities were classified and predicted accurately. However, almost 5% of supposed activities were wrongly classified.

With consideration of the accuracies of all existing studies which have used the traditional machine learning methods for classifying the human activities and child activities, including sensor-based data, RFID, and other classification methods on sensor data. The pre-trained model VGG16 is applied to the dataset created, and it achieves an 81.79% accuracy rate.

Though, our proposed CNN architecture proceeds advantage of spontaneous feature extraction of deep learning techniques, achieved the highest accuracy compared to existing research works on the sensor data and video data of child and human actions also (Table III).

IV. CONCLUSION

Child Activity Recognition is the process of identifying what the child is doing based on its movement. For human activity recognition, deep learning techniques are utilized broadly, but it has not used for child activity recognition that much. The existing researches have used the sensors installed on the baby's body for recognizing the child activities instead of deep learning strategies. The need for wearable devices can be eliminated using deep learning strategies like CNN with extracted features and video datasets for recognizing the child activity. So, in this work, Convolutional Neural Networks (CNNs) architecture has been proposed to recognize child activity as a deep learning model from a video dataset. We have recognized the child activities like sleeping, walking, running, crawling. The performance of the offered method has been compared with the existing methods, which have used traditional machine learning methods for recognizing child activity and deep learning methods for recognizing human actions. Experimental results showed that the offered deep learning model has accomplished 94.73% accuracy in the recognition of child activities. A fusion of various deep learning models can be proposed as future work, also the classifiers on an enormous precise child activity dataset.

REFERENCES

1. D. R. Beddiar and B. Nini, "Vision based abnormal human activities recognition: An overview," *2017 8th International Conference on Information Technology (ICIT)*, Amman, 2017, pp. 548-553.
2. S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013.
3. Zhi Liu, Chenyang Zhang, Yingli Tian. 3D-based Deep Convolutional Neural Network for Action Recognition with Depth Sequences, *Image and Vision Computing* (2016).
4. P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue and N. Zheng, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1963-1978, 1 Aug. 2019.
5. H. D. Mehr and H. Polat, "Human Activity Recognition in Smart Home with Deep Learning Approach," *2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*, Istanbul, Turkey, 2019, pp. 149-153.
6. T. S. Kim and A. Reiter, "Interpretable 3D Human Action Analysis with Temporal Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 1623-1631.
7. F. An, "Human Action Recognition Algorithm Based on Adaptive Initialization of Deep Learning Model Parameters and Support Vector Machine," in *IEEE Access*, vol. 6, pp. 59405-59421, 2018.
8. Wei, Li, and Shishir K. Shah. "Human Activity Recognition using Deep Neural Network with Contextual Information." *VISIGRAPP (5: VISAPP)*. 2017.
9. Tomas and K. K. Biswas, "Human activity recognition using combined deep architectures," *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, Singapore, 2017, pp. 41-45.
10. Jayashree Onkar Nehete, D.G.Agrawal.Real time Recognition and monitoring a Child Activity based on smart embedded sensor fusion and GSM technology , The International Journal Of Engineering And Science (IJES), Volume 4, Issue 7, PP.35-40, July – 2015.
11. Nam, Yunyoung, and Jung Wook Park. "Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor." IEEE journal of biomedical and health informatics 17.2 (2013): 420-426.
12. S. Gaglio, G. L. Re and M. Morana, "Human Activity Recognition Process Using 3-D Posture Data," in *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586-597, Oct. 2015.
13. B. Su, H. Wu, M. Sheng and C. Shen, "Accurate Hierarchical Human Actions Recognition From Kinect Skeleton Data," in *IEEE Access*, vol. 7, pp. 52532-52541, 2019.
14. S. Gaglio, G. L. Re and M. Morana, "Human Activity Recognition Process Using 3-D Posture Data," in *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586-597, Oct. 2015.
15. Max Wang & Ting-Chun Yeh. Human Action Recognition with CNN and BoW Methods, Stanford University, CS229 Machine Learning, 2016, pp.1-5.
16. B. Su, H. Wu, M. Sheng and C. Shen, "Accurate Hierarchical Human Actions Recognition From Kinect Skeleton Data," in *IEEE Access*, vol. 7, pp. 52532-52541, 2019.
17. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X. and Chen, D.S., 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors*, *19*(5), p.1005.
18. Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

AUTHORS PROFILE

Binjal Suthar has completed B.Tech in Computer Engineering from LDRP-ITR, Gandhinagar, Gujarat. She is pursuing last semester Master of Engineering in Software Engineering from Government Engineering College, Gandhinagar, Gujarat. Her major research field areas are Machine Learning and Data Mining.

Prof. Bijal Gadhia has completed M.Tech in Computer Science and technology from Ganpat University, Kherva, Gujarat. She is currently working as an assistant professor in Computer Engineering Department, at Government Engineering College, Gandhinagar. Her major research field areas are Image Processing and Machine Learning.