ASD behaviour analysis with computer vision

Elaborated by :
Asma Abidalli
Sarra Hammami

Outline

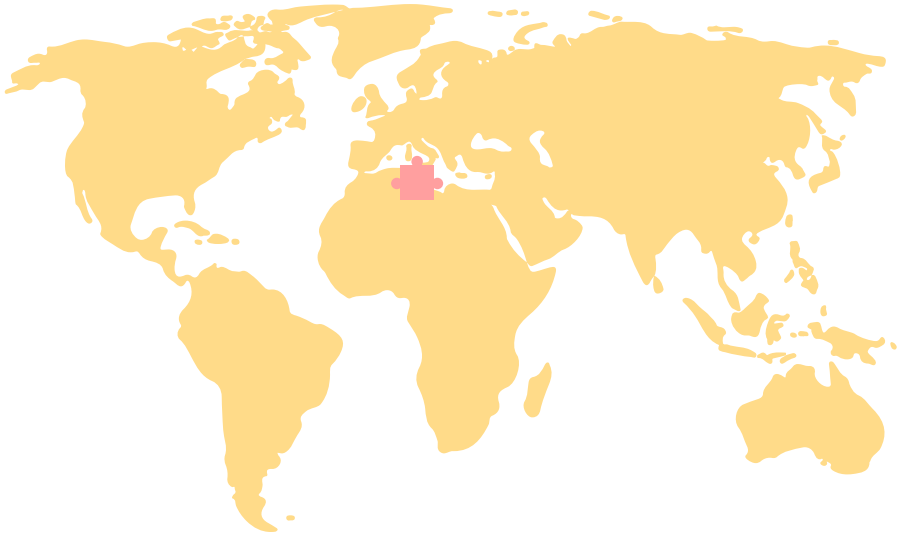
01

INTRODUCTION

Introduction

ASD is a neurodevelopmental disorder characterized by a set of social communication deficits, self-harm, or persistent repetition of actions. Moreover, this disorder often manifests in children during their early developmental stages and can have severe negative impacts on the quality of their life over a long time period.

Statistics



+200 000

Children affected by ASD in **Tunisia**

+75 million

People have autism spectral disorder around the world

02

Problematic

The diagnosis process is time-consuming as it requires long-term behavior observation.

There are two reasons for the long wait times for ASD diagnosis:

- The low availability of specialists
- There are no reliable biomarkers for ASD, and its diagnosis requires a long-term observation of stereotypical behaviors such as **headbanging**, **arm-flapping** or **spinning**.

Stimming behaviours are more common in autistic children and it is usually observed during children's regular daily activities.
These atypical behaviours, when observed early, can lead to an early intervention and diagnosis.

Arm flipping

Head banging

03

Solution

Solution

Building a region-based computer vision system that aims to help clinicians and parents analyze children's behaviors, and in particular help to identify behaviors associated with Autism Spectrum Disorder (ASD).

04

Dataset

Self-Stimulatory Behavior Dataset (SSBD) for ASD

The dataset was collected from **YouTube videos** recorded in uncontrolled environments

It includes three stereotypical behaviors:

- **Arm flapping**
- **Headbanging**
- **Spinning**

The original dataset contains **62 videos.**

We have added more videos collected from youtube .

[dataset link: https://rolandgoecke.net/research/datasets/ssbd/]

05

Approaches & results

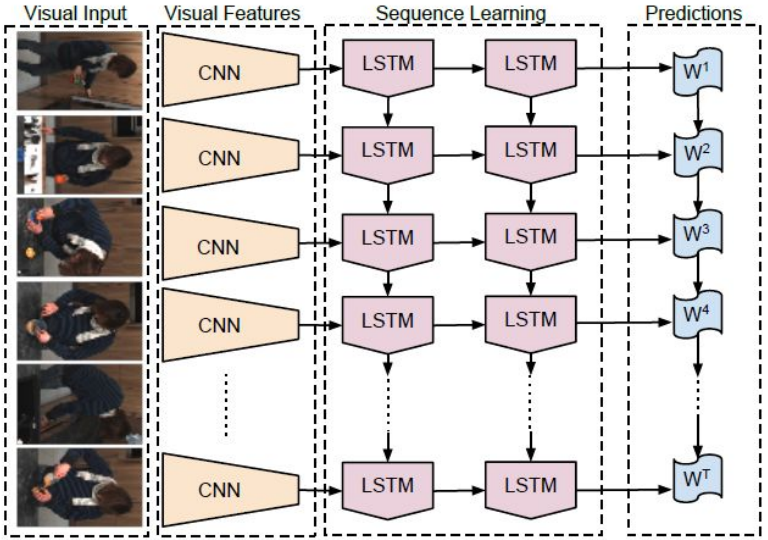
Preprocessing

- **Frame extraction**

- **Resizing and normalization**
- **Histogram equalization** to enhance the contrast of frames

LRCN Approach

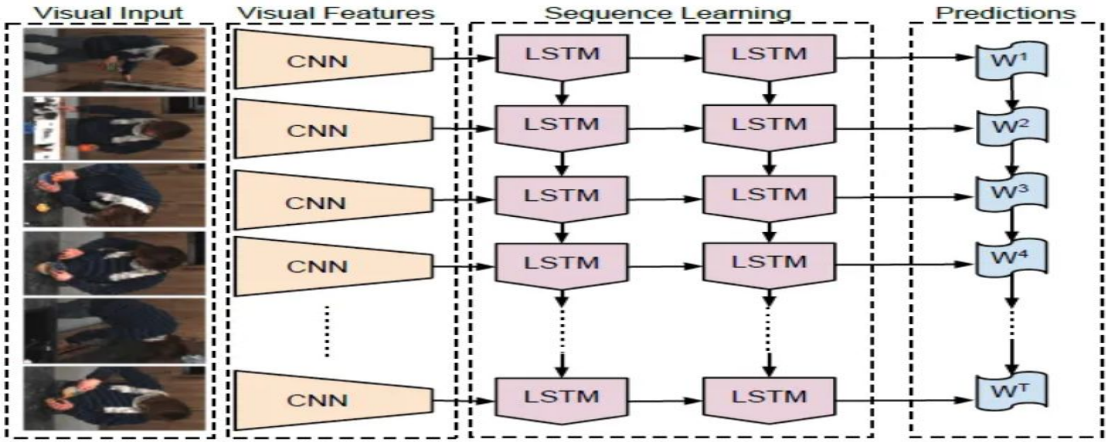
- The LRCN approach for action recognition is a deep learning architecture that combines Convolutional Neural Networks (CNNs) and LSTM networks in a single model.



[**LRCN Based Human Activity Recognition from Video Data** (Muhammad Sajib Uzzamana, Chandan Debnatha, Md Ashraf Uddina, Md. Manowarul Islama, Md. Alamin Talukdera, Shamima Parvezb)]

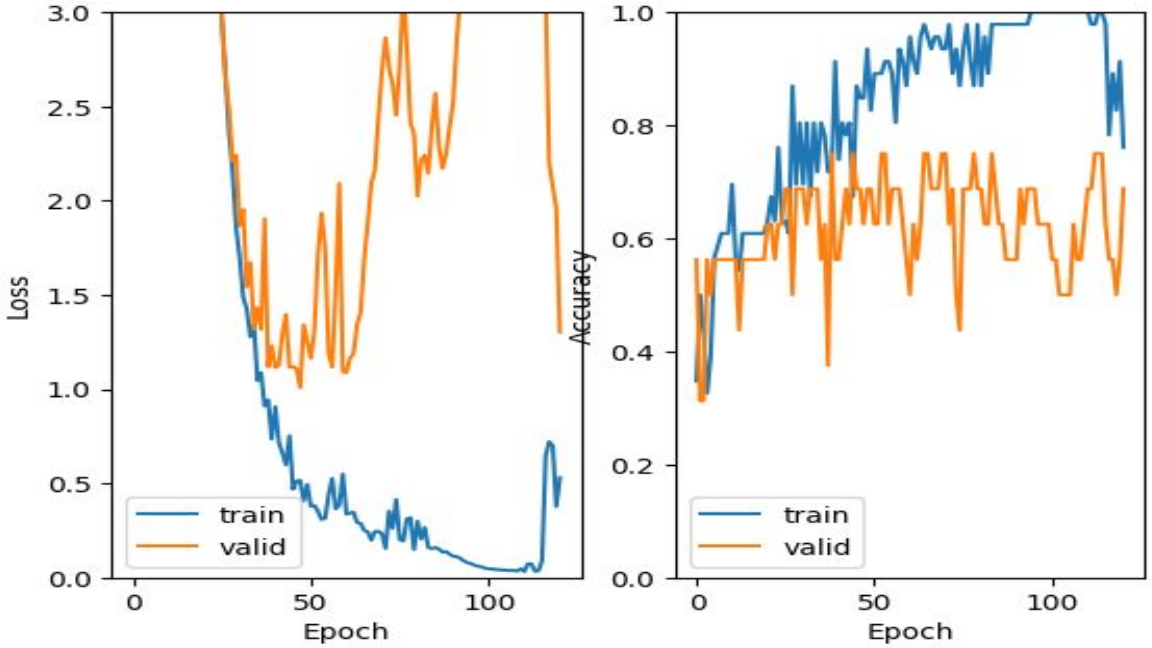
LRCN Approach

- The key innovation of LRCN is in combining the spatial feature extraction capabilities of CNNs with the sequential modeling capabilities of LSTMs. After extracting spatial features from individual frames using the CNN, the LSTM is used to model the temporal dependencies and relationships between these features over time.



Long-term Recurrent Convolutional Networks (LRCNs)

LRCN Results



Accuracy=0.68

3D CNN Approach

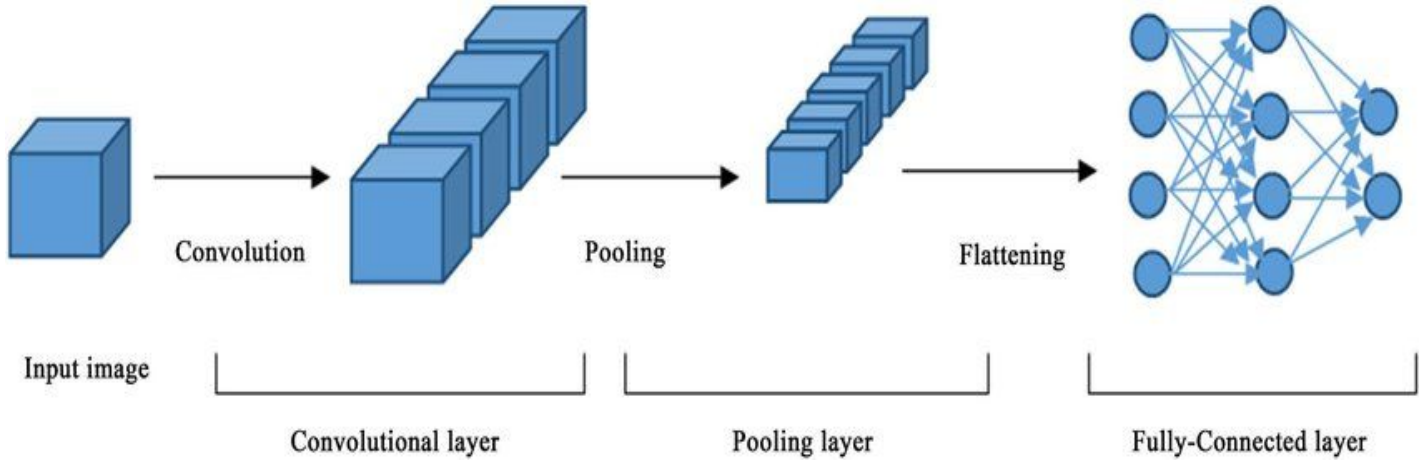
- 3D CNNs are designed to capture both **spatial** and **temporal features** simultaneously. They process video data as three-dimensional volumes, enabling them to learn intricate patterns within individual frames and temporal dynamics across multiple frames.
- It uses 3**D convolutional** layers to extract features from volumetric data. These layers slide a 3D kernel (a cube) over the input volume to detect patterns in all three dimensions.

[J. Arunnehru, G. Chamundeeswari, S. Prasanna Bharathi: Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos, 2018 (link: https://www.sciencedirect.com/science/article/pii/S1877050918310044)]

3D CNN Approach

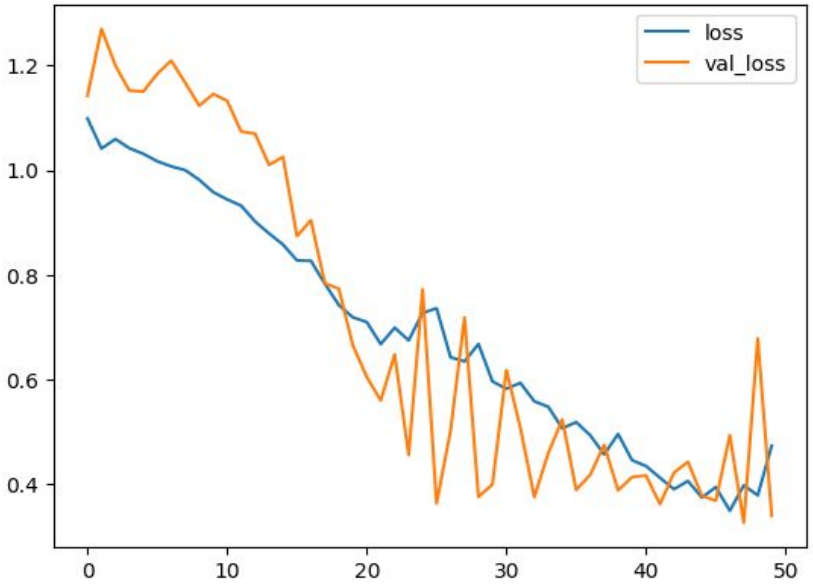
- The model employs 3D max-pooling layers and strides to downsample the spatial dimensions of the data, reducing the computational load.

  The final softmax layer produces probability scores for various action classes.

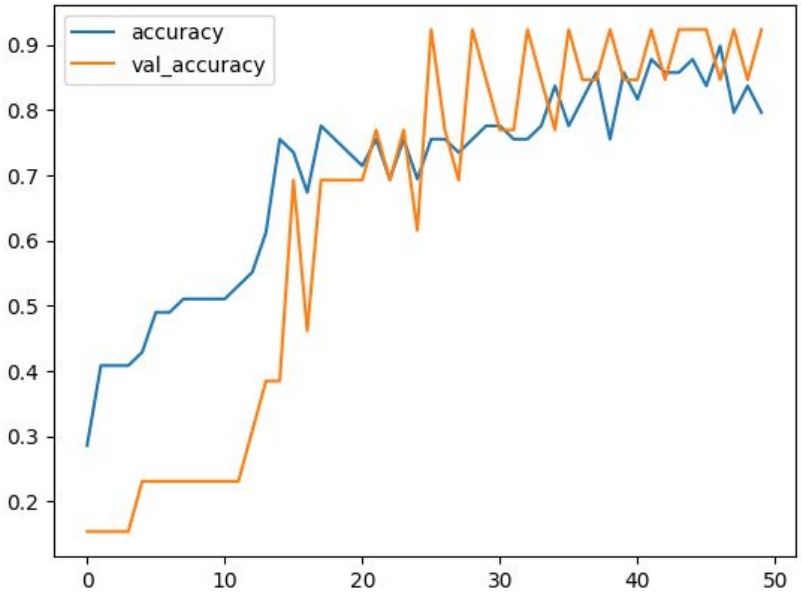


3D CNN- Results

Training and validation losses

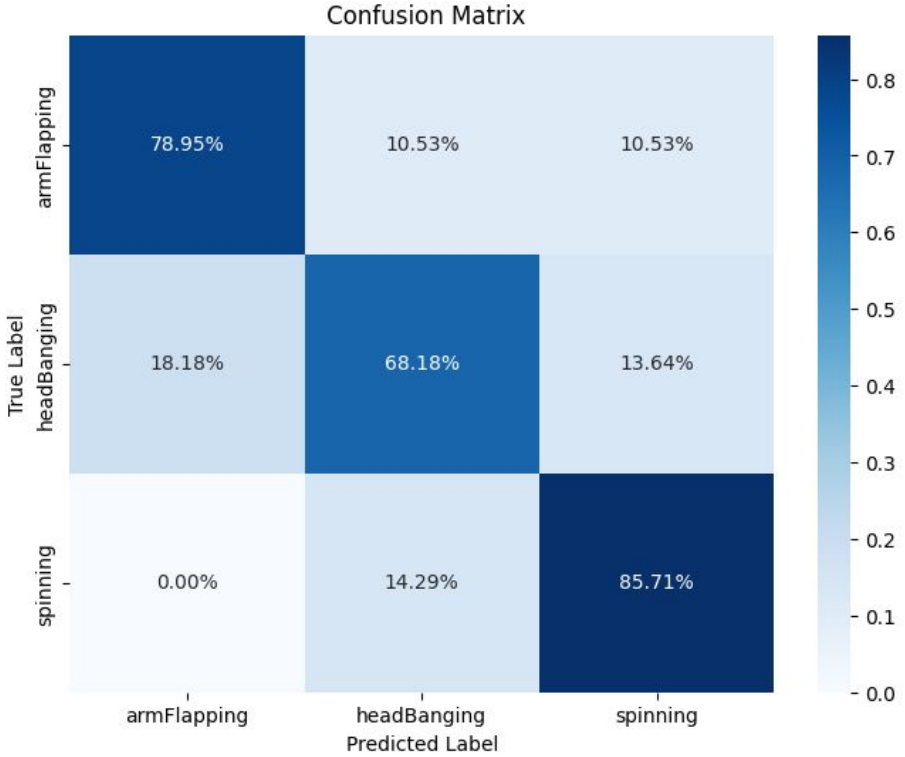


Training and validation accuracy




```
1/1 [==============================] - 5s 5s/step - loss: 0.3626 - accuracy: 0.9231
Validation Loss:0.3625984191894531 Validation Accuracy:0.9230769276618958
```

3D CNN- Results



Classification Report:

	precision	recall	f1-score	support
armFlapping	0.79	0.79	0.79	19
headBanging	0.75	0.68	0.71	22
spinning	0.78	0.86	0.82	21
accuracy			0.77	62
macro avg	0.77	0.78	0.77	62
weighted avg	0.77	0.77	0.77	62

05

Challenges &Demo

Challenges

- Lack of publicly available data.
- Objects in videos can be partially or fully occluded, making it difficult for the model to recognize and track them consistently
- Large frames requires significant computational resource.
- Videos arerecorded in uncontrolled environments, , resolution and quality are not good enough ...

Demo

We have developed with streamlit a simple app that integrates our model and predict the the behaviour class of a give video .

Streamlit

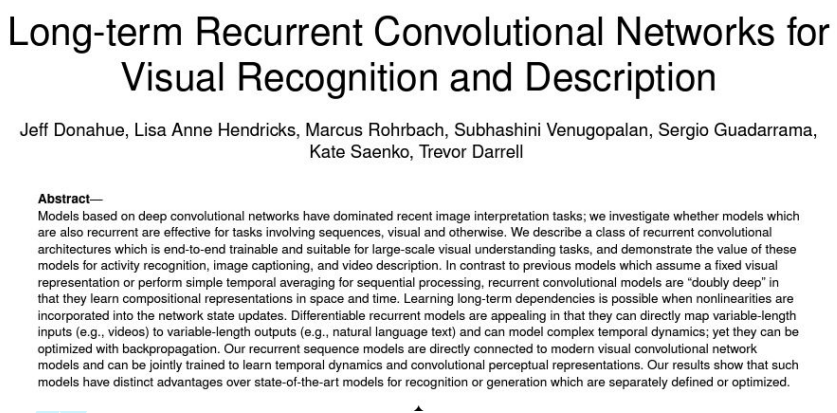
Papers adopted :

- **Vision-Based Activity Recognition in Children with Autism-Related Behaviors**

Link: https://arxiv.org/abs/2208.04206

- **Long-term Recurrent Convolutional Networks for Visual Recognition and Description**

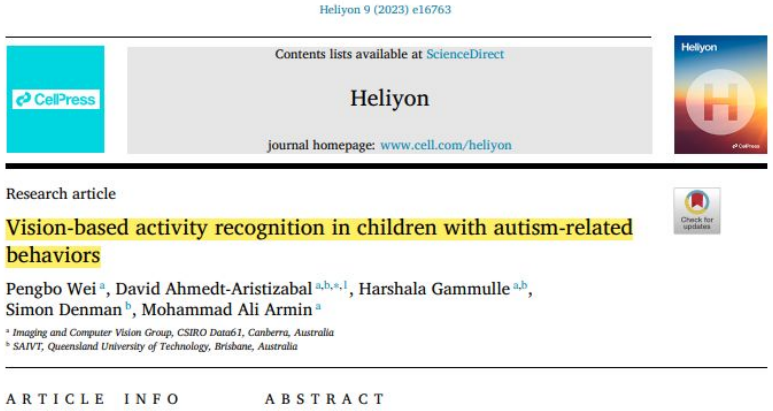
link:https://arxiv.org/abs/1411.4389?source=post_page&fbclid=IwAR0sJt1Nqpl4BqcpH0Jm_lO4uRlvYqa7dMl6d3o5PBHBadZqa0enZGsVEvc



Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, Trevor Darrell

Abstract—
Models based on deep convolutional networks have dominated recent image interpretation tasks; we investigate whether models which are also recurrent are effective for tasks involving sequences, visual and otherwise. We describe a class of recurrent convolutional architectures which is end-to-end trainable and suitable for large-scale visual understanding tasks, and demonstrate the value of these models for activity recognition, image captioning, and video description. In contrast to previous models which assume a fixed visual representation or perform simple temporal averaging for sequential processing, recurrent convolutional models are "doubly deep" in that they learn compositional representations in space and time. Learning long-term dependencies is possible when nonlinearities are incorporated into the network state updates. Differentiable recurrent models are appealing in that they can directly map variable-length inputs (e.g., videos) to variable-length outputs (e.g., natural language text) and can model complex temporal dynamics; yet they can be optimized with backpropagation. Our recurrent sequence models are directly connected to modern visual convolutional network models and can be jointly trained to learn temporal dynamics and convolutional perceptual representations. Our results show that such models have distinct advantages over state-of-the-art models for recognition or generation which are separately defined or optimized.



Heliyon 9 (2023) e16763

CellPress

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Vision-based activity recognition in children with autism-related behaviors

Pengbo Wei [a], David Ahmedt-Aristizabal [a,b,*,1], Harshala Gammulle [a,b], Simon Denman [b], Mohammad Ali Armin [a]

[a] Imaging and Computer Vision Group, CSIRO Data61, Canberra, Australia
[b] SAIVT, Queensland University of Technology, Brisbane, Australia

ARTICLE INFO ABSTRACT

THANK YOU FOR YOUR ATTENTION!