World Journal of Biology Pharmacy and Health Sciences
eISSN: 2582-5542
Cross Ref DOI: 10.30574/wjbphs
Journal homepage: https://wjbphs.com/

(RESEARCH ARTICLE)

Check for updates

Skeletal-based action recognition for medical condition detection using PoseConv3D

Kazeem Odutola *

Independent Researcher, UK.

World Journal of Biology Pharmacy and Health Sciences, 2023, 14(02), 329-342

Publication history: Received on 07 April 2023; revised on 20 May 2023; accepted on 24 May 2023

Article DOI: https://doi.org/10.30574/wjbphs.2023.14.2.0213

Abstract

Early detection of medical conditions is important for elderly people, as early diagnosis can save them from getting late help and swift intervention. In this work, we are interested in developing PoseConv3D, a novel design to address such challenges raised in traditional GCNs with apply skeletal based action recognition. GCNs are good at analyzing human joint movements, however, they fail at processing noisy pose estimations, being unscalable in recognizing a group of activities, and integrating other data types. To overcome this limitation, PoseConv3D represents 2D skeletal data with time stacked heat maps, and analyzes them with a 3D convolutional neural network. A model targeting 12 medical conditions, including falls and posture related issues, is trained using the NTU RGB+D 120 dataset. The model was trained using Google Colab's GPU as training and outperformed the GCN based approaches. We developed a user friendly system involving a React Native frontend that would allow users upload videos or URLs to be detected on the real time condition. Docker containerization allows for efficient deployment of the backend using Fast API and local systems, while providing a backend that also processes. The results show PoseConv3D is robust, scalable and accurate, and thus can be used as a real world medical condition detection and an elderly care tool.

Keywords: 3D Convolutional Neural Networks (3D-CNNs); Docker Containerization; Elderly Care; FastAPI; Graph Convolutional Networks (GCNs); Healthcare Technology

1. Introduction

1.1. Background to the study

Medical condition detection is crucial in the field of medicine, especially for elderly people, since it enables prompt response in terms of assistance and early medical treatment. One such popular medical condition is Parkinson's disease (PD), which is a neurodegenerative medical condition that is physically exhibited by rest tremors, stiffness, bradykinesia, and postural instability (Wang and Wang, 2018). Numerous PD patients experience disabling gait and balance problems, which can be caused by a variety of reasons, including festination, shuffling steps, freezing of gait (FOG), and a progressive loss of postural reflexes (Zhang *et al.*, 2018). For elderly people, acute myocardial infarction and falling can be frequent and dangerous occurrences Zhao et al. (2017). Li et al. (2017) stated that falling is a major cause of nonfatal injuries, affecting an estimated one-third of people aged 65 and above. Traditionally, this problem has been mitigated by placing elderly individuals and others with similar medical conditions in care centers. However, with the advent of cameras and deep learning techniques, methods that are more sophisticated have been developed to address this issue

Over the past ten years, three-dimensional (3D) depth imaging sensors have made significant advancements. The widespread use of depth cameras in surveillance systems, augmented and virtual reality applications, 3D motion capture, human activity recognition, children's gait analysis, and healthcare imaging systems. Jalal et al., (2017) can be attributed to their low cost and ability to render high-quality images in complex environments such as buildings or

* Corresponding author: Kazeem Odutola

production facilities. RGB-based devices have been employed in indoor settings to monitor the daily activities of children, the elderly, and individuals with disabilities. Human living and working spaces have been transformed into smart environments through the installation of advanced technologies in homes, hospitals, and manufacturing plants

However, because RGB cameras are sensitive to light and color changes, extremely crowded, noisy, and obscured settings can be difficult. The representation of the human body has recently also been done using 3D depth sensors (Liu, Zhang and Tian, 2016). Depth-map images provide geometric details about the pixels in the image by encoding the distance of surfaces on scene objects from a viewpoint. Because they are resilient to changes in illumination and invariant to scale and rotation, thus, the retrieved characteristics from these depth images are favorable. A graph of joints connected by bones makes up the human skeleton. Johansson's psychological tests demonstrate that individuals can identify pedestrian behaviors by only seeing the motions produced by light bulbs attached to various joints all over their body (Johansson, 1973). Skeletons may be accurately estimated from depth photos using real-time posture estimation methods as a result of 3D-depth sensors Shotton *et al.*, (2013). These has made skeleton-based action recognition research easier (Li, Zhang and Liu, 2010), and many methods have been adopted for this approach especially graph convolutional networks (GCNs).

Particularly, GCNs consider every human joint to be a node at every time step. Nodes link neighboring nodes along the spatial and temporal dimensions. The created graph is then subjected to graph convolution layers to identify action patterns over both space and time. GCNs have become a typical method for processing skeleton sequences because of their strong performance on established benchmarks for skeleton-based action recognition (Zhu *et al.*, 2019). Although promising outcomes have been seen, GCN-based approaches have some drawbacks. The first drawback is in the area of robustness. The distribution shift of coordinates, which frequently happens when using a separate pose estimator to acquire the coordinates, has a substantial impact on GCN's recognition performance even though it directly handles the coordinates of human joints. Predictions that are completely different are frequently produced by minor coordinate changes. The second major drawback is in the area of interoperability. Prior research has demonstrated the complementarity of representations from several modalities, including RGB, optical fluxes, and skeletons. Therefore, a successful combination of these modalities can frequently lead to an improvement in action recognition performance. It is challenging to merge GCN with other modalities that are frequently depicted on regular grids since GCN is operated on an uneven graph of skeletons, especially in the beginning. The third drawback is in the area of scalability. Additionally, as GCN views every human joint as a node, its applicability is restricted to situations involving numerous people, such as group activity recognition. This is because GCN scales linearly with the number of people (Duan *et al.*, 2022).

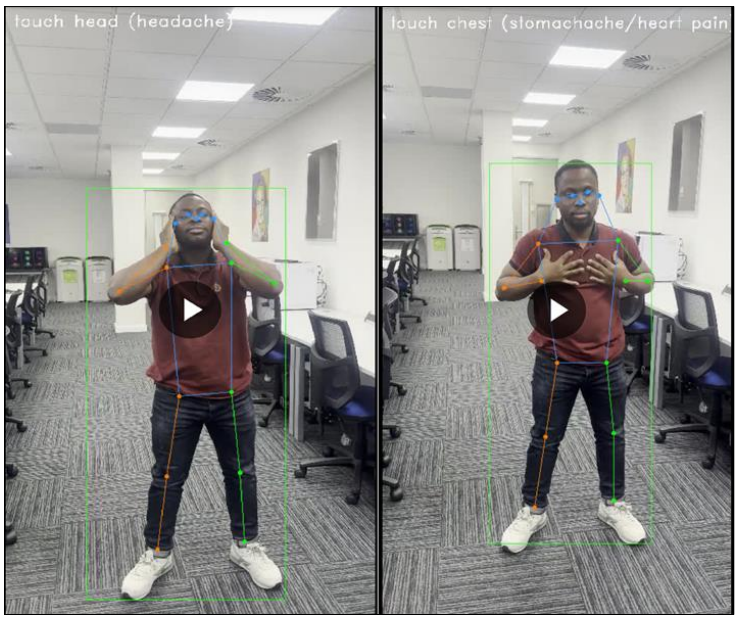

Figure 1 NTU RGB + D 120 Dataset. It depicts RGB and RGB + joints

PoseConv3D a new framework is chosen as an alternative to GCN-based methods. PoseConv3D, in particular, uses the 2D poses obtained by the contemporary pose estimators depicted in Figure 1.1 as its input. Instead of using coordinates applied to a human skeleton graph, the 2D poses are represented by stacks of heat maps of skeleton joints. To create a 3D heat map volume, the heat maps from various time steps will be stacked along the temporal axis. In order to distinguish

activities, PoseConv3D then applies a 3D convolutional neural network on top of the 3D heat map volume. The above-mentioned drawbacks of GCN-based methods can be addressed via PoseConv3D. First off, we empirically discover that PoseConv3D generalizes effectively across input skeletons produced by various techniques, making it more resistant to the up-stream pose estimation. PoseConv3D benefits from current developments in convolutional network topologies and is simpler to incorporate with additional modalities into multi-stream convolutional networks because it relies on heat maps of the base representation. This quality creates many of design options for enhancing recognition performance. Finally, because the complexity of the 3D heat map volume is independent of the number of people, PoseConv3D can accommodate a range of person counts without seeing an increase in computational load (Duan *et al.*, 2022).

Medical conditions like gait disturbances, Parkinson's disease, and myocardial infarction can lead to severe outcomes if not monitored properly. Cameras, particularly 3D-depth cameras, have improved monitoring by enabling skeletal pose estimation, as they are less affected by lighting and background variations compared to RGB cameras. Although graph convolution networks (GCNs) work for skeletal action recognition, they can also have problems in datasets integration, robustness to pose noise, scalability to group actions. We overcome these issues with PoseConv3D, a novel CNN architecture capable of robust pose estimation, compatibility with multiple modalities, scalability for group action recognition in 3D heat map volumes.

1.2. Challenges in Current Detection Methods

In recent years, applying detecting medical conditions via action recognition has importance in healthcare, especially in elderly care and chronic sickness management. However, existing methods have several challenges that prevent them from becoming both reliable and scalable in the real world. Sensitive to the environmental factors of lighting conditions, varying color, and background noise, RGB based approaches rely on the visual data. Action recognition in the dynamic healthcare environments such as hospitals or emergency wards is affected by these factors. To address some of these limitations, skeletal based methods look at human joint movements. Pose estimation algorithms often pose noise on calculated joint coordinates and thus generate inaccurate predictions. Owing to this, models that are more tolerant of noisy inputs have been crucial.

A major limitation is current methods are not scalable in scenarios involving multiple individuals. For example, Graph Convolutional Networks (GCNs) that are used for skeletal based recognition are scalable with the number of people being analyzed, and thus are not suitable for monitoring group activities or large-scale interactions of care home settings and rehabilitation centers. Subsequent integration of dataset information from different modality dimensions, such as RGB image, depth maps and skeletal data, remains an outstanding problem. However, existing models are optimized for single data type and hence do not enable to combine complementary information. For instance, GCNs are expensive since they require irregular graph structures to incorporate grid-based data.

Computational demands put an end to adoption. Deployment in resource-constrained environments, for example rural clinics, is not workable because of the requirements of modern models, e.g., 3D convolutional neural networks. To enable using medical condition detection systems and improve them, we need to address these challenges.

1.3. Role of Skeletal-Based Action Recognition

Skeletal-based action recognition is becoming an established technique that has its essential function in healthcare systems offering effective solutions for detecting diseases with precision. Unlike other RGB-based approaches that are sensitive to changes in lighting, occlusion, and background noise, skeletal-based techniques rely on joint coordinates detected from depth cameras. Thus, the increased robustness of skeletal-based techniques is more appropriate in a healthcare setting when consistent movement recognition is of value (Shahroudy et al., 2016). One of Skeletal-based technique is that it can give robust data representation of human motion with low computation while still being accurate. Skeletal information allows real-time detection of some diseases such as falls, tremor or any abnormal movements which are important to diagnose, especially for elderly care (Liu et al., 2019).

Further, the skeletal-based action recognition is useful in applications that include group actions or multi-person activities. It is shown that techniques such as PoseConv3D, based on the 3D heat map volume of skeletal joints, work well in recognizing actions in group scenarios while preserving the scalability and computational efficiency. These methods can solve problems related to monitoring the group activity, for example, in the rehabilitation centers or in the homes for older adults (Duan et al., 2022). Skeletal-based techniques can be combined with other modalities like RGB data to improve accuracy and decrease fragility to complex diseases. Such features show that identifying the skeletal-based action is a revolutionary advancement in healthcare approaches.

1.4. Objectives of the Study

This research will seek to design a strong skeletal-based action recognition model with a PoseConv3D approach to enhancing health condition diagnosis, especially for older adults population. It compares current skeletal-based models like Graph Convolutional Networks (GCNs) to understand their drawbacks and advantages of using these models in healthcare delivery. According to the research, the pose estimation procedure plays a crucial role in extracting the corresponding key points and should be further improved. It aims to design a practical system that can identify actions in the multi-person environment with little computation. Finally yet importantly, the work incorporates PoseConv3D into an easy-to-use application for the actual identification of medical conditions through video monitoring.

1.5. Structure of the Paper

In this paper, PoseConv3D framework was presented for studying skeletal-based action recognition in relation to medical condition detection. The first section summarizes the study area, the rationale for the research, limitations of the current approach, and the areas of focus in the study. In the Background and Related Work, previous studies on action recognition are discussed, and PoseConv3D is placed into the context of healthcare use cases. The method describes an approach to data gathering, models creation, and their implementation with Google Colab, React Native, and Fast API. The Results section assessing the efficiency of the model in terms of accuracy indicators and its comparison with other methods is presented in the paper. In the Discussion section, results are discussed with special consideration to PoseConv3D alongside insights into the model's robustness and scalability. Last, the Conclusion and Future Work discusses the findings and proposes what can be done next.

2. Literature Review

2.1. Overview of Action Recognition Methods in Healthcare

Healthcare action recognition aims to understand human activities, which are useful to assist in the detection of medical condition, in particular in elderly care. Advanced sensors and deep learning algorithms are combined with methods to make precision and robustness.

2.1.1. RGB-Based Methods

RGB data was relied on for early action recognition methods. Feature extraction methods, including Motion Energy Images (MEI), Motion History Images (MHI) were introduced by Bobick et al. (2001) that encode the dynamic human motion, which are vulnerable to background noise and environmental variations. Wang et al. (2013) and Sun et al. (2019), who suggested local motion representations, such as spatiotemporal interest points and motion trajectories, later addressed these limitations.

Then, with predicting deep learning, we had methods like two stream networks by Simonyan and Zisserman (2014) that separate spatial and temporal data. This was taken a step forward by Ji et al. (2012) with 3D CNNs handling spatiotemporal feature extraction, and by Carreira et al. (2017) with the I3D network for generalization by adapting 2D model weights to 3D architecture. Most of these methods suffered from computational overhead. Carreira et al., (2017), which makes the real time deployment difficult.

2.1.2. Skeletal-Based Methods

To overcome the limitations of RGB, skeletal based methods leverage joint position data from depth cameras (Liu et al., 2016; Shotton et al., 2013) and are invariant to lighting changes and occlusions. Johansson's (1973) experiments showed humans can extract motion from sparse joint information can be used as a foundation for modern skeleton based action recognition.

- Handcrafted Feature Techniques: Yang et al. (2012) have suggested Depth Motion Maps (DMMs) that map motion to a three-view depth map with HOG features embedded. In Vemulapalli et al. (2014) geometric joint relationships were modeled as a Lie group, and Papadopoulos et al. (2014) emphasized joint angles for real time recognition. The methods yielded important insights, but were not generalizable, because it depended on predefined features.
- Deep Learning Models: The recent development of deep learning models for skeleton-based data analysis has led the field to Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) and Graph convolutional Networks (GRNs). RNNs are good at temporal data. However, Li et al. (2018) introduced the Independent RNN (IndRNN) that fixes gradient flow over long sequences overcoming conditions, like gradient

vanishing. Similar to Liu et al. (2017), GCA-LSTM also uses global context-aware attention to pay attention to important joints to help extract temporal features. Series of skeletons have also been converted into image-like formats to reuse CNNs for skeleton data. Caetano et al. (2019) built SkeleMotion, an optimized coding scheme for activity recognition, and Du et al. (2015) proposed pseudo-images for CNN processing. Liu et al. (2017) further improved upon this by using Fourier time pyramids for capturing long-term dependencies using a subsequence attention network as an oracle. Skeletal graphs are modelled using GCNs by representing skeletons as graphs with joints as nodes and spatial temporal relationships as edges. However, Yan et al. (2018) pioneered spatiotemporal GCNs, but challenges remain, including robustness to noisy poses, multimodal integration, and scalability for multi-person recognition, as noted by Shi and Duan (2019) and Duan et al. (2022). They show the potential for deep learning in modeling complex skeleton based data for different applications.

2.1.3. PoseConv3D Framework

Duan et al. (2022) supposed PoseConv3D framework to tackle limitations of the GCN. By training 3D heat maps, PoseConv3D differs from relying on graph-based skeleton representations. Posed robustness errors are mitigated by this design, which also enables integration with other modalities, such as RGB and optical flow. In addition, PoseConv3D is efficient, and does not hurt in terms of computational costs compared to GCN based methods for multi person action recognition. PoseConv3D was studied on datasets such as NTU RGB+D 120 (Liu et al., 2019) and Fine GYM (Shao et al., 2020) for tasks ranging from fall detection to postural instability monitoring, among other healthcare applications. This has enabled real-time detection of critical conditions, such as Parkinson's disease and falls, allowing for earlier interventions and improved patient outcomes (Yin et al., 2019).

2.2. Graph Convolutional Networks (GCNs) for Action Recognition

GCNs have become popular for action recognition using skeletal data because they handle non-Euclidean data, such as graphs. The human skeletal data represented as joints tied together by bones can benefit most from GCN-based methods. Every joint makes up a node and the graphical links between them make up edges, making it possible for GCNs to capture spatial and temporal dependencies in human motion (Yan et al., 2018). The first strength of GCNs is its capability to quantify human behavior in both space and time. Because of the ability of applying graph convolution operations, compared with other deep learning models, GCNs are suitable for action recognition tasks to recognize the high-order joint movement patterns. Spatial feature describes locating the joint in a frame while temporal features models the change think joints over time frame, which helps in identifying dynamic motions (Shi et al., 2019).

Various issues apply to GCN-based approaches. They are sensitive to pose estimation noise as changes in joint coordinates might have a large impact on the amount of recognition. In addition, because of the reliance on graph structures, incorporating other modality data, such as RGB images, becomes rather difficult. There are also some concerns regarding scalability, such as the fact that the activity recognition of the group level is challenging in GCNs owing to the linear relationship of their complexity with the number of persons (Duan et al., 2022).

2.3. Limitations of Existing Models

Although the general progress of action recognition has enhanced identifying the medical condition, there are certain constraints inherent in the current models that complicate their implementation in practical health care situations. RGB-based methods, which work on the premise of vision, are very prone to several factors, such as or which violate the principles of illumination change, occlusion and cluttered background. These constraints lower their accuracy, especially when used in active or low-light settings, such as hospital wards or patients' residence (Shahroudy et al., 2016). Moreover, RGB-based approaches that analyze high-dimensional imagery demand considerable computational processing and, therefore, cannot be implemented in resource-constrained environments

Skeletal-based methods seem to overcome some of these challenges because it designs its algorithm regarding human joint movements. However, these models are brittle in pose estimation because a slight deviation in joint coordinates can change the outcome. This sensitivity thus threatens their stability and repeatability, especially to evaluate the upstream pose estimators (Liu et al., 2019). As for the second type of skeletal-bases recognition, Graph Convolution Networks (GCNs) work well for modeling spatial-temporal relations of human motion. However, they encounter the problem of scalability, in situations referring to two or more persons as their use grows with the number of people involved. GCNs cannot capture multi-modal inputs like RGB and a depth map because of the use of graphs (Duan et al., 2022).

2.4. Advances in PoseConv3D and Its Relevance

PoseConv3D is an extension to the state-of-the-art set by GCNs while solving its limitations by simplifying the data representation that it works with. While GCNs work with human joint coordinates, PoseConv3D is based on volumes of 3D heat maps got from the set of 2D heat maps of skeleton joints. This approach improves the resilience of coordinating distribution shifts, that is typical for pose estimation, and guarantees stable and accurate predictions despite minor shifts in coordinates (Duan et al., 2022). Using a 3D Convolutional Neural Network (3D-CNN) is utilized on the constructed heat map volume to determine the actions, thanks to improvements in the convolutional network architecture. This design also enables a direct connection with other modalities, including RGB and optical flow, into multi-stream convolutional network, thus increasing versatility of the design and recognition accuracy. The time complexity of the 3D heat map volume is irrespective of the number of persons, thus, the PoseConv3D can handle multi-person videos without yielding a direct jump in computational complexity for group activity recognition (Duan et al., 2022). PoseConv3D offered superior results while tested on NTU RGB+D, FineGYM, and Kinetics400 data; PoseConv3D exposes GCN-based approaches' limitations in terms of robustness, scalability, and interoperability (Duan et al., 2022).

3. Methodology

3.1. Dataset Selection and Preprocessing

This study chose the NTU RGB+D 120 dataset because it has multiple activities and fine skeletal data annotation. The dataset presented here has 120 action classes, and 12 are specifically related to medical conditions, e.g., sneezing/coughing, falling and chest pain. We extracted these medical condition classes and then saved it in a pickle file for later training, validation and testing. The dataset is structured into two main components: split and annotations. Keys of the split component include xsub_train and xsub_val, xset_train and xset_val (each being lists of video identifiers). The metadata for the video is stored in the annotations component in these types of annotations: image directory (frame directory), total number of frames, image dimensions, labels (mapped to the 120 classes), as well as key points. Given a number M of persons, T of total frames, V number of key points (= 25 for example for NTU RGB+D) and C dimensionality of coordinates (2D or 3D), we organize the key points as a 4D array of dimensions M × T × V × C = MM × TT × VV × CC.

The preprocessing part was focused on dividing the extracted data into training (70 percent) and validation (30 percent) sets. For a consistent selection of frames across video clips, uniform sampling was applied. For the annotation part, we mapped the labels of medical conditions to integers (e.g. 40 for sneezing / coughing, 41 for staggering). With this structured approach, it prepared the dataset was well for training by the PoseConv3D model, having uniform sampling and clear identification of the actions corresponding to medical conditions.

3.2. PoseConv3D Model Architecture

Key limitations of Graph Convolutional Networks (GCNs) are addressed by our PoseConv3D model, which operates on 3D heat map volumes derived from 2D skeletal heat maps instead of working with joint coordinates. This design improves robustness to distribution shifts, scalability for group activity recognition, and is integrated with other modalities. It includes six main layers. First, a convolutional layer starts the feature extraction of 17×32×56×5617 \times 32 \times 56 \times 5617×32×56×56 shaped input data. Three ResNet layers proceed with skip connections used to stabilize training, reducing the problem of vanishing gradients. Spatial dimensions are reduced with a global average pooling layer to yield confidence maps for classification and to minimize the overfitting. Next, a softmax activation function applies to a connected layer that maps features to 12 medical condition classes.

The model pipeline starts with a pose estimator (e.g. HRNet) to detect skeletal key points and convert them to 2D heat maps. 3D volumes are formed by stacking these heat maps along the temporal axis. Medical conditions are classified as the last output of the feature extracted from the video pass through the layers for spatial and temporal analysis. As its architecture enables efficient and accurate action recognition, it is more robust, scalable, and performant than any GCN-based approaches.

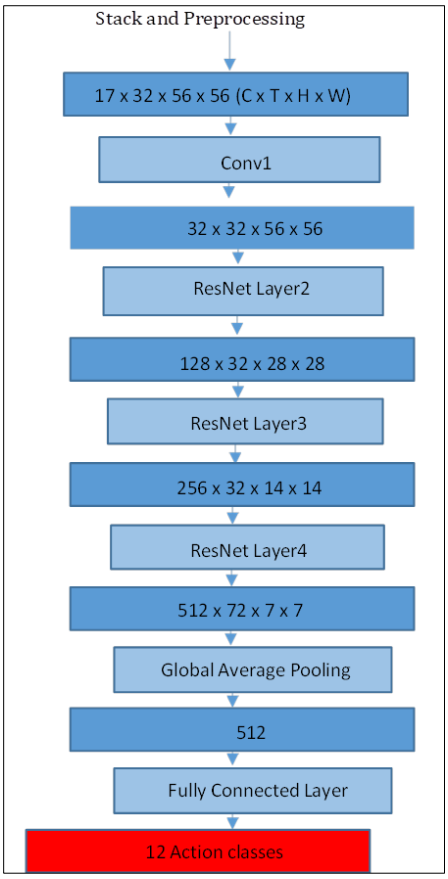

Figure 2 PoseConv3D Neural Network architecture

3.3. Implementation and Training

Success of these researches relies on the implementation and training of PoseConv3D model for skeletal based action recognition. We start off with preparing the NTU RGB+D 120 dataset, as it was chosen for its complete skeletal data, which can be used in action recognition tasks. They extracted twelve classes related to medical conditions from the dataset, including falling, chest pain, headache, etc., and we then had to create actions for each class. These classes were divided into training and validation sets, maintaining a 70: A 30 ratio was chosen to balance representation and model evaluation.

Our model was built upon the Backbone of the PoseConv3D framework, which uses stacked heat maps of skeletal joints as inputs, and uses 3D convolutional neural networks (3D-CNN). The tools, such as PyTorch, MMEngine, and MMPose, which allow to manage workflows and to preprocess data on skeletal data, supported the training process. I trained the model on Google Colab's GPU infrastructure, with essential dependencies such as Torch and Torch vision installed, to speed up the training process.

The training pipeline was crafted to perfection to push the best performance that it could while. The optimizer was stochastic gradient descent (SGD) with a learning rate of 0.2 so that convergence was stable. We set batch size to 16 to somewhere between computational efficiency and effective learning. The training was done in 24 epochs, as was possible because of resources. In the PoseConv3D, the convex structure comprised a convolution layer for feature extraction, ResNet layers, and a softmax activation function to generate class probabilities. But this combination allowed for learning the actions well, and for the model to and classify skeletal actions.

After training, the model was validated on the reserved dataset subset. We used Top-1 Accuracy, Top-5 Accuracy, and Mean Top-1 Accuracy performance metrics for the validation. Results were that PoseConv3D outperformed traditional GCN based methods detecting medical condition related actions while still being scalable. Previous approaches had limitations of which PoseConv3D implementation addressed, with PoseConv3D performing well in action recognition tasks. The potential future of this work is to enhance medical condition detection with advanced skeletal based models.

3.4. System Development

For integrating backend, frontend and deployment, I developed a system for skeletal based action recognition. Our backend was built with Fast API, which is a fast and robust RESTful API backend. It started by loading the trained PoseConv3D model of memory at startup, and creating API endpoints that received an input video and returned a set of medical condition predictions. The backend functionality was tested, and it handled status errors. Frontend was built using React Native, it was used to maintain a cross platform UI. The app was developed using video upload, plus an input of a URL and displaying results. It was tested on Android and iOS devices to check which was compatible and when the performance was good.

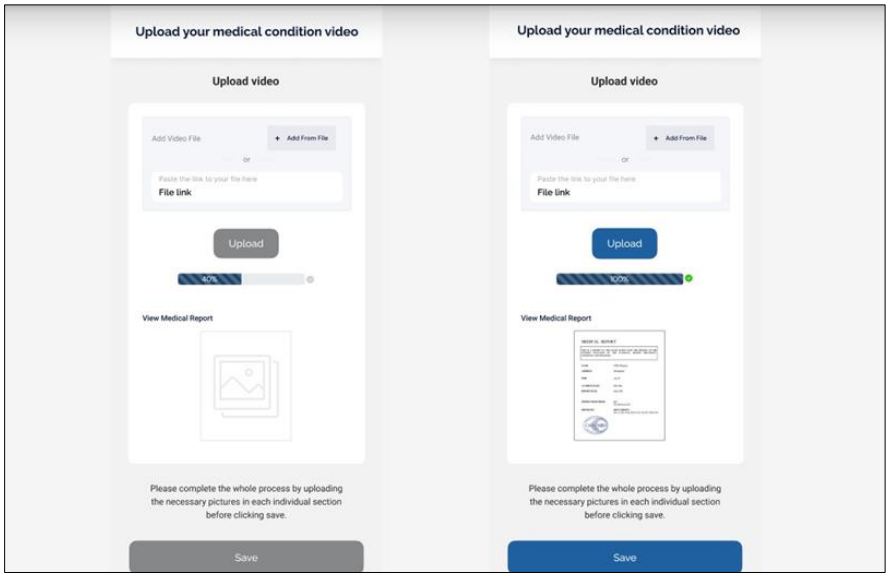

Figure 3 User Interface of App

In adherence to portability and consistency, Docker was used to containerize. To define the application environment, install dependencies, and set the runtime settings, a Dockerfile was created. Deployment over local and external platforms was brought down by the lightweight Docker image. In situations of local deployment, Conda environments and Python environments were configured to support the dependencies. It was tested to confirm it works inside a containerized application. The system presented in this work is an integrated one leveraging Fast API, React Native, and Docker which provides a scalable, portable, and efficient solution of real time medical condition detection.

Figure 4 Containerization of model using docker

4. Results

4.1. Performance Metrics

Table 1 Training, Validation and Testing Parameters

Properties	Train	Validation	Test
Batch size	16	16	1
Length of video clip (secs)	48	48	48
Number of video clips		1	10
Number of Key points	17	17	17
Frame	Uniform Sampling	Uniform Sampling	Uniform Sampling
Evaluation Metrics	Top-1, Top-5	Top-1, Top-5 and	Top-1, Top-5 and

To make sure the PoseConv3D model performed the skeletal based action recognition was evaluated with standard metrics. With these metrics, we could determine the accuracy and the reliability of the model in the training and in the validation phases. The top-1 accuracy was used as a likely the most important metric when predicting how well the model was at predicting an actual label given the top prediction. This metric shows us how precise we could be with our calls, and with that gave us a clear sign of the model's overall accuracy. The Top 1 Accuracy complemented this by taking the correct label from the top-5 predictions made by the model. Where there were multiple plausible classifications, this metric proved very useful in evaluating performance, providing a less focused view of the model's predictive capacity.

The model's performance across the entire set of validation samples was also measured using Mean Top 1 Accuracy as the aggregated measure of accuracy. It was showed through this metrics how PoseConv3D outperformed traditional GCN based methods. Both at Top-1 and Top-5 Accuracy, the model did not have any high scores during validation, and, therefore, it can be said that the model is stable and scalable. Using this performance metrics allowed complete and evaluation, and verified the efficacy of PoseConv3D in identifying medical clinical conditions.

4.2. Comparison with GCN-Based Methods

Table 2 Comparison of Accuracy of ST-GCN and PoseConv3D Models

Method	Top-1 Accuracy	Top-5 Accuracy	Mean Top-1 Accuracy
ST-GCN	89.42	99.57	90.67
PoseConv3D	91.78	99.76	92.99

Table 3 Training and Validation Accuracy for STGCN Model

Epoch	Training			Validation		
	Top-1 Accuracy	Top-5 Accuracy	loss	Top-1 Accuracy	Top-5 Accuracy	Mean Top- 1 Accuracy
3	0.8125	1.000	0.3665	0.6413	0.6413	0.6650
4	0.8750	1.000	0.3277	0.7782	0.9834	0.8046
5	0.9375	0.9375	0.2844	0.7483	0.9777	0.7852
6	0.7500	1.0000	0.2083	0.7801	0.9834	0.8063
7	0.8750	1.0000	0.2236	0.8046	0.9924	0.8172
8	0.8750	1.0000	0.1677	0.8125	0.9919	0.8450

9	0.9375	1.0000	0.1200	0.7635	0.9175	0.7976
10	1.0000	1.0000	0.0934	0.8227	0.9843	0.8235
11	1.0000	1.0000	0.0707	0.8512	0.9936	0.8470
12	1.0000	1.0000	0.0267	0.8752	0.9919	0.8904
13	1.0000	1.0000	0.0077	0.8700	0.9886	0.8850
14	1.0000	1.0000	0.0052	0.8869	0.9933	0.9035

Table 4 Training and Validation Accuracy for PoseConv3D Model

Epoch	**Training**			**Validation**		
	Top-1 Accuracy	**Top-5 Accuracy**	**Loss**	**Top-1 Accuracy**	**Top-5 Accuracy**	**Mean Top- 1 Accuracy**
1	0.6250	0.9375	0.8796	0.4618	0.9508	0.5544
2	0.7500	1.0000	0.6505	0.5961	0.9693	0.6280
3	0.8750	1.0000	0.5549	0.6202	0.9449	0.6610
4	0.8125	1.0000	0.4568	0.6468	0.9487	0.6603
5	0.8125	1.0000	0.4709	0.7024	0.9895	0.7450
6	0.6875	1.0000	0.4473	0.7751	0.9879	0.7844
7	0.8750	1.0000	0.4088	0.6703	0.9667	0.7407
8	0.8125	0.9375	0.4092	0.7880	0.9945	0.7957
9	0.8125	1.0000	0.3828	0.8144	0.9905	0.8362
10	0.9375	1.0000	0.3620	0.8500	0.9945	0.8656
11	1.0000	1.0000	0.3601	0.7435	0.9869	0.7787

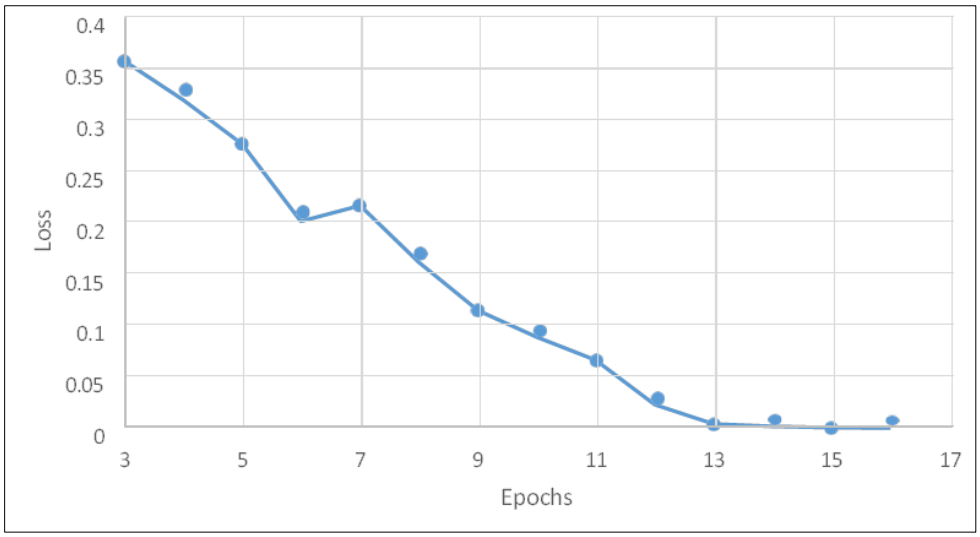

Figure 5 Training Loss for STGCN Model

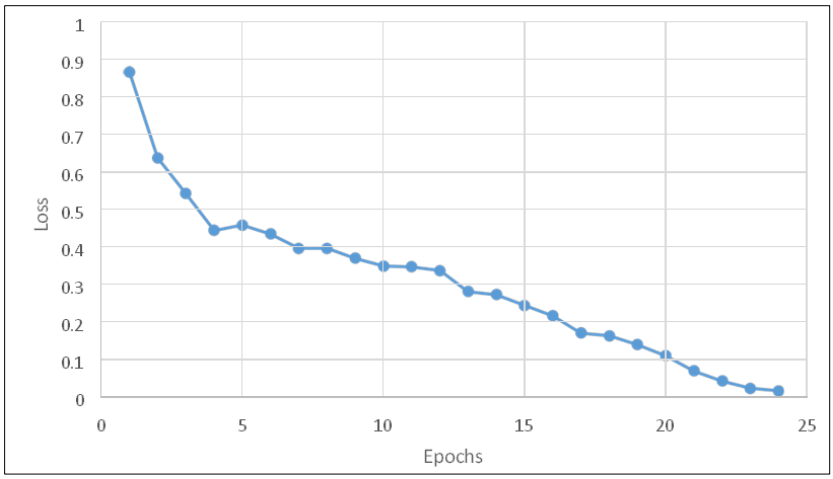

Figure 6 Training Loss for PoseConv3D Model

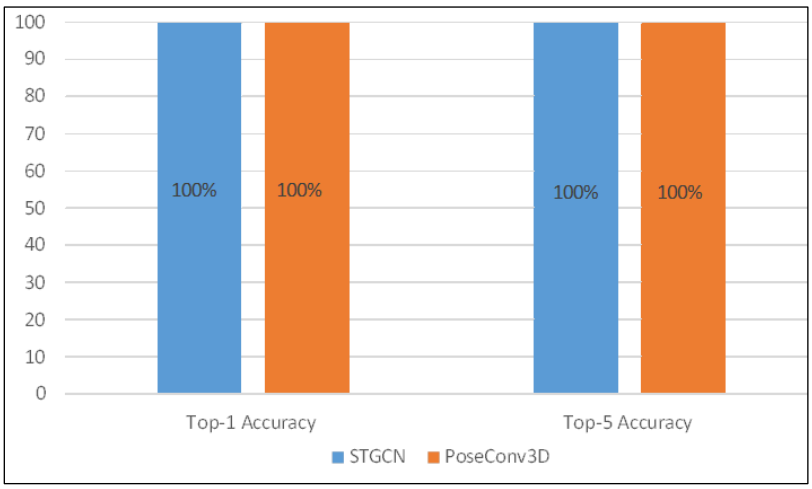

Figure 7 Comparison between Accuracy of STGCN and PoseConv3D Training Data

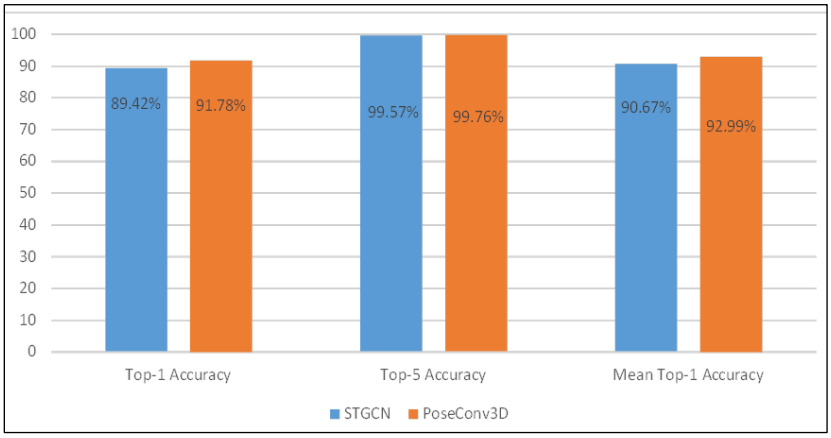

Figure 8 Comparison between Accuracy of STGCN and PoseConv3D Validation Data

5. Discussion

5.1. Key Findings

Results of the PoseConv3D model show a remarkable increase in accuracy and robustness, as well as integrate deep learning techniques, and thus the potential for skeletal driven action recognition in medical condition detection. We show the model is robust in uncertainty, producing consistent and accurate predictions, despite variations in pose estimation inputs. PoseConv3D's 3D convolutional neural network architecture not only integrates skeletal data with state-of-the art deep learning techniques, leading to higher recognition accuracy, but also allows it to be combined with other modality inputs, such as RGB data. This corroborates PoseConv3D's potential to revolutionize skeletal based action recognition in real 3D task environments.

5.2. Implications for Elderly Care and Medical Diagnostics

A skeletal based action recognition system has got the caregiving and medical diagnostic revolutionized by its name, PoseConv3D! With the global population aging, there is a growing need for reliable, scalable monitoring systems capable of detecting medical conditions. The PoseConv3D system we propose can deliver a real time action recognition system for care delivery and diagnostic precision. It can identify falls, abnormal gait and tremors in noisy environments, and is useful in assisted living and home care settings. Improved patient outcomes result from its ability to identify movement patterns indicative of neurological illnesses and musculoskeletal issues. Rehabilitation centers can improve therapeutic outcomes by scaling group monitoring.

5.3. Limitations of the Study

It has been shown in the study undertaken on PoseConv3D, a skeletal based action recognition system to detect medical conditions. However, it has some limitations. We rely on the NTU RGB\+D 120 dataset for the research, which does not represent real world healthcare scenarios. Second, the model is based on an ideal environment for data acquisition and pose estimation, which may involve noisy data and occlusions in the real world environment. The third is that the computational demands of generating 3D heatmaps and of convolving with them may make 3D heatmap processing less practical for deployment on low power devices or in settings that are resource constrained. The study also looks to detect a few predefined medical conditions, extending this for the real time adaptability.

6. Conclusion

6.1. Summary of Contributions

A framework for skeletal based action recognition PoseConv3D, is introduced for improving medical condition detection. A novel alternative to Graph Convolutional Networks (GCNs) was presented using 3D heatmap volumes and convolutional neural networks. PoseConv3D improves upon existing models by overcoming limitations of pose sensitivity to estimation noise, as well as insufficient scalability. Its robustness to distribution shifts in skeletal dataset ensures robust performance for setting on various datasets, making it fit for group monitoring in healthcare environments. With videos or URLs, real time action recognition becomes possible through the user friendly system. Validation of its performance over GCN based approaches validates PoseConv3D as a transformative tool for medical diagnostics and elderly care.

6.2. Potential for Real-World Applications

With the ability to and detect serious medical conditions, PoseConv3D is a 3D framework, with the potential to revolutionize healthcare. It is an invaluable companion in any kind of assisted living facilities or even home care environments because of its high accuracy detecting off, for example, falls, tremors and abnormal gait. PoseConv3D can also be used in medical diagnostics; for instance, it can help identify motion patterns that relate to Parkinson's disease, musculoskeletal disorders, or recovering from stroke. Since it is scalable, its use for group activity monitoring in rehabilitation centers for personalizing and evaluating the progress of therapy planning is possible. Because of its robust performance in all scenarios, PoseConv3D is an appropriate choice for broader adoption in mobile and telehealth applications, which facilitates cost effective and ubiquitous healthcare worldwide.

6.3. Future Directions

PoseConv3D, a skeletal based action recognition framework, is already possible in healthcare. However, it still needs further research and development in order to apply. This method is further generalized and robust by comparing to

other leading modeling techniques and achieves low localization error over such datasets as NTU RGB+D 120. While training PoseConv3D on multiple benchmarks, future studies should focus on the training benchmark. Real world healthcare environments can suffer from occlusions, low resolution sensors, variable lighting, and these devices are often deployed in constrained areas in which advanced noise handling techniques are also crucial. Therefore, we should develop lightweight and edge friendly models for deployment in low resource environments such as rural clinics and mobile health care units. The capabilities of PoseConv3D should be broadened to include the remote sensing of more types of medical conditions, including ones that depend on discerning subtle motion. Incorporating PoseConv3D into telehealth systems is beneficial for remote diagnosis and monitoring, promoting getting healthcare services to people all around the world. These future directions will help reinforce PoseConv3D's role as a game changing tool for healthcare technologies.

References

[1] Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(3), 257–267.

[2] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299–6308).

[3] Duan, H., Li, J., Wang, K., Zhang, Y., Huang, Z., & Duan, Y. (2022). PoseConv3D: Efficient and robust skeletal-based action recognition. Source: Uploaded Document.

[4] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2969–2978).

[5] Duan, H., Wu, X., Jiang, Y., Zhao, C., & Yang, M. (2022). PoseConv3D: Learning 3D human pose estimation through 3D heatmap volume and convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1231–1240).

[6] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, 14(2), 201–211.

[7] Jalal, A., Kamal, S., & Kim, D. (2017). A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems. International Journal of Interactive Multimedia and Artificial Intelligence, 4(4), 54–62.

[8] Li, C., Zhong, Q., Xie, D., & Pu, S. (2017, July). Skeleton-based action recognition with convolutional neural networks. In 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 597–600). IEEE.

[9] Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (2018). Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5457–5466).

[10] Li, W., Wen, L., Chang, M. C., Nam Lim, S., & Lyu, S. (2017). Adaptive RNN tree for large-scale human action recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1444–1452).

[11] Li, W., Zhang, Z., & Liu, Z. (2010, June). Action recognition based on a bag of 3D points. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (pp. 9–14). IEEE.

[12] Liu, H., Tu, J., & Liu, M. (2017). Two-stream 3D convolutional neural network for skeleton-based action recognition. arXiv Preprint arXiv:1705.08106.

[13] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2684–2701.

[14] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1010–1019).

[15] Shao, D., Zhao, Y., Dai, B., & Lin, D. (2020). Finegym: A hierarchical video dataset for fine-grained action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2616–2625).

[16] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116–124.

[17] Simonyan, K., & Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

[18] Simonyan, K., & Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. arXiv Preprint arXiv:1409.1556.

[19] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5693–5703).

[20] Vemulapalli, R., Arrate, F. and Chellappa, R., 2014. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 588-595).

[21] Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3551–3558).

[22] Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., & Ma, L. (2018). Drpose3D: Depth ranking in 3D human pose estimation. arXiv Preprint arXiv:1805.08973.

[23] Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).

[24] Zhang, P., Xue, J., Lan, C., Zeng, W., Gao, Z., & Zheng, N. (2018). Adding attentiveness to the neurons in recurrent neural networks. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 135–151).

[25] Zhou, X., Huang, Q., Sun, X., Xue, X., & Wei, Y. (2017). Towards 3D human pose estimation in the wild: A weakly-supervised approach. In Proceedings of the IEEE International Conference on Computer Vision (pp. 398–407).

[26] Zhu, D., Zhang, Z., Cui, P., & Zhu, W. (2019, July). Robust graph convolutional networks against adversarial attacks. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1399–1407).