MEDICAL ANOMALIES DETECTION USING HUMAN ACTION RECOGNITION IN PUBLIC AREAS - FG 2021 Submission

Anonymous FG2021 submission
- DO NOT DISTRIBUTE -

***Abstract*—Medical Anomaly detection in public areas is a crucial application to ensure public safety. Especially during the times of Covid-19 pandemic it becomes crucial to monitor visual symptoms in public areas in order to take necessary precautions. The recent advancements in the field of Computer Vision combined with the modern sequence classification models made the action recognition and action classification which demand high computation power possible to deploy in real time . The work analyzes the need for automated monitoring software to provide safety to the public. Most of the existing Human Activity monitoring tools use various types of human wearable sensors to get the data and interpret action based on the collected data. which is not a very highly available source of data thus making it a very non scalable solution but we use video data as a source thus we can scale our solution regardless of the circumstances that we face in public area. Firstly, we revisit the recent advancements done in this field by exploring the existing methods and we delve into all the scope of improvements. In this paper we propose an automated approach which combines the power and versatility of the 2 state-of-art architectures(HR-Net[5] and Transformer[2], [17]) to get a highly accurate system.**

***Keywords*—Human Action Recognition, Anomaly detection, Transformer, HR-Net**

I. Introduction

Ever since COVID-19 broke out, it has become necessary to monitor medical anomalies in public places as it can be beneficial in breaking the chain of spreading this disease. In general, Automated processes are advantageous as compared to existing primitive methods in terms of the amount of manual work required, speed and scalability.

In the existing primitive methods, a person has to manually call for medical assistance in case of an emergency in public areas. When a large scale of these events are considered there is a lot of manual work involved which can be reduced by using the automated process presented in this paper. Here we used Human Action Recognition in order to detect anomalies in public areas.

Human Action Recognition is one of the challenging tasks in computer vision which involves key-points identification of the human body and classifying the action based on those key points. The project is aimed to detect medical anomalies in public areas by employing advanced neural network architectures developed in the field of computer vision. Due to many critical real-life applications, the problem demands high speed and accuracy. Since the motion of a human can be described by the combined motion of his/her joints(or key points), [13] The major challenges are broken into 2 parts: key points identification and anomaly detection through classification from the continuous stream of data. The existing methods for keypoint prediction employ a variety of architectures a few of them have delivered state of art results in both high accuracy and speed maintaining the resolution. Of all those models hrnet[5] delivers state-of-the-art results because of which HR-Net is employed in our project. The classification for a continuous stream of data is best solved by recurrent neural network type of architecture. The success of recently discovered transformer architecture is known for its attention mechanism, because of which it is considered as a better replacement for most of the standard recurrent neural network architectures.

Motivated by the success of transformers in other fields [2],[17] we decided to employ transformers for the anomaly classification.

II. Related Work

a. Key point detection

Highly accurate Key point identification is generally achieved by many advanced neural network architectures which include CNN[8], R-CNN[6] Hour-glass[4], Resnet-50 HR-Net architectures, most of these architectures are held back by problems like loss of information (Resolution), Occulatoin, large training time, longer inference time.

Loss of information could lead to voids in key points which causes misprediction which can be costly in case of medical anomaly detection. HR-Net addresses this by maintaining high resolution throughout its pipeline.

Occultation (occultation is an event that occurs when one object is hidden by another object that passes between it and the camera) causes loss in key-points which may result in null values in data leading to misclassifications. It can be addressed by using the HR-Net model.

HRNet-W32[5],[1],[7],[10]: It is a state of the art model which achieves high performance by maintaining high resolution throughout the process with the help of parallel layers of high resolution and fusing the low resolution to maintain high resolution. It also uses a bottom up approach which is very favorable for real time detection as it is very fast.

b. Anomaly detection

Traditional methods of anomaly detection for serial Data involve using recurrent type neural networks like RNN and LSTM [16] date back to 1997. RNN and LSTM have the capability of understanding the time-series data but are limited by their large computational cost, small memory windows and slow training speed. Transformers[20] are introduced as the models developed as better replacement for the RNN's to provide better performance driven by their attention mechanism which have delivered state-of-art results in sequential data classification applications like [2], [3], [18],

III. TECHNICAL APPROACH

The proposed system of our approach is as follows:

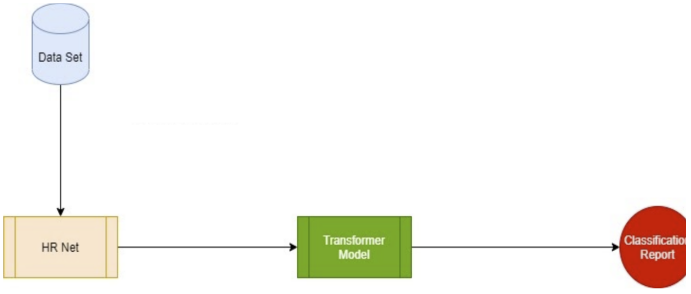

Fig. 1. Architecture of the proposed model

We use HR-Net for extracting key points of the people present in the dataset and the obtained key points are used to classify the action by using Transformer.

A. Keypoints detection

HR-Net helps in detecting keypoints of a person. HR-Net does this in a bottom-up approach. Since it is doing it in a bottom up approach it can parallelly detect keypoints for upto 30 people in a frame. This makes it an exceptionally fast model for detecting keypoints. HR-Net also overcomes an important problem that is found in most of the bottom-up approaches that we find in models nowadays, this problem is regarding the scale of the humans in a frame. HR-Net uses a scale aware approach. This Approach helps HR-Net to be able to detect keypoints accurately even if a person is far away from the observer or person is very close to the observer. So even in a mix of both the cases HR-Net can predict accurate keypoints. HR-Net overcomes voids in keypoint detected by maintaining a high resolution throughout its pipeline. As for tasks such as Human action recognition, having voids in keypoints makes it highly difficult to predict the action, the HR-Net helps in providing good data to the transformer model further down the pipeline. HR-Net also overcomes occlusion. When there are multiple people in a frame there is a high chance of people obstructing the view of the camera which hides some vital key points of other people, leading to voids in data. HR-Net was trained in a way to solve this problem .Thus we use HR-Net for keypoint detection.

Here the input images are passed through HR Net to get the key points of all the people present in the image.As Now the Output of the HR Net has a high resolution representation of our key points and lacks any voids in the data. This is then used as an input to the next Phase where we predict if any person is doing any anomalies action.

B. Action classification

Sequence classification tasks are better handled by the recurrent type neural networks. Many types of neural networks are evolved to handle the sequential data most famous of them include RNN, LSTM and GRU based architectures. With recent advancement in sequence to sequence translation in the field of Natural Language Processing using the novel Transformer architecture. Because of the state-of-art results and possibility to deploy in real time, the transformer is considered as the replacement for many existing models to handle the sequential data. Due to the attention mechanism designed in Transformer can derive the contextual information and handle the sequential data . In this paper we use the attention mechanism of the transformer to encode the sequential data and obtain the contextual information which is then passed to a simple deep neural network for classification of the data.

Transformer: Transformer architecture is introduced as an efficient neural network for sequence translation in natural language processing. The introduced Transformer is basically divided into 2 blocks: an encoder stack and a decoder stack. In this project we only use the encoder stack to obtain the contextual information of the sequence. Encoder stack contains attention heads which are responsible for generating the contextual information .

Attention Mechanism : Attention mechanism is implemented using attention heads which compute the relevance of every vector(each point in the sequence is given as a vector of numbers) in the sequence with every other vector. At first 3 learnable vecorts query(q), key(k) and value(v) vectors are computed by multiplying the each vectors of the sequence with weight matrices the self attention or relevance of each unit in sequence with other units will be calculated by computing scaled-dot-product attention as given below.

Encoder: each encoder contains many attention heads each of which will calculate different contextual information finally all the contextual information arrays are merged and final information is obtained.

A classification deep neural network is used to classify the contextual information obtained from the encoder into the action classes.

C. *Working of the model*

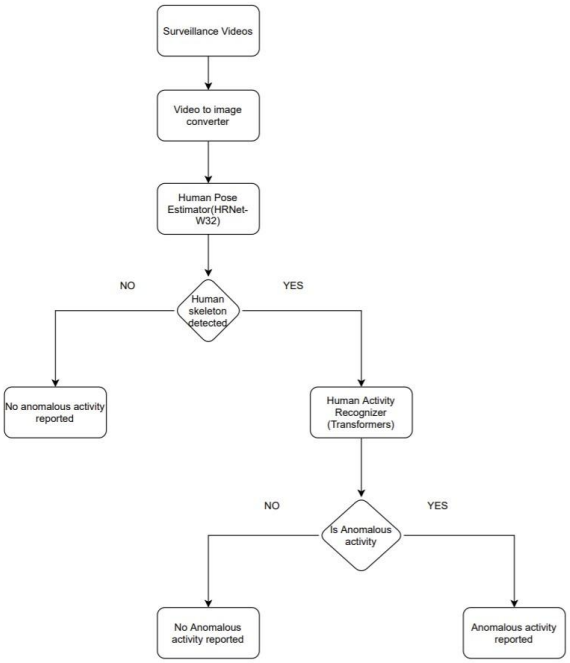

Fig. 2. Flow chart of working of the model

The videos from the dataset are converted into successive image frames and are passed onto the HRNet[5] model which extracts the key points of all humans present in each frame as explained previously. This intermediate data of key points then goes through the Human action recognizer which is implemented by the Transformer[2], [17] where the action is recognized based on the key points data. Then the actions are classified based on the preset anomalous activities and the anomalous activities are reported whenever detected.

D. *Optimizers and Training process*

The Training process consists of two stages. The first stage is getting the key points from the HR-Net. The next stage is training the Transformer model with the data acquired in the first stage.

The Transformer is trained to classify seven classes. In the input data corresponding to each output, Transformer takes in a vector of shape [no of framesX34] where no of frames rows correspond to 18 sequential frames of the input video and 34 columns correspond to co-ordinate of the key points. All the input data is scaled using Min-Max Scaler in sk-learn.

The loss function used is Categorical Cross Entropy and The metric used while training is Accuracy. The optimizer used is Adam optimizer with default parameters.

IV. Dataset and Experimentation

This work uses the Human activity recognition dataset from Rose labs. In this paper we chose Seven actions for the application of medical anomalous activity recognition. For training we used 1974 videos from the same dataset. The following actions were considered for this work:

- Sitting down
- Standing up
- Jumping
- Sneezing/coughing
- Staggering
- Falling down
- Nausea/ vomiting

The validation accuracy as high as 88% was achieved with a dataset consisting of 1974 videos by training on a Tesla T4 GPU for only 15min(for 83 epochs). Even higher accuracies can be achieved by increasing the dataset size and running for more number of epochs.

A. *Figures and Tables*

1) *Performance metrics*

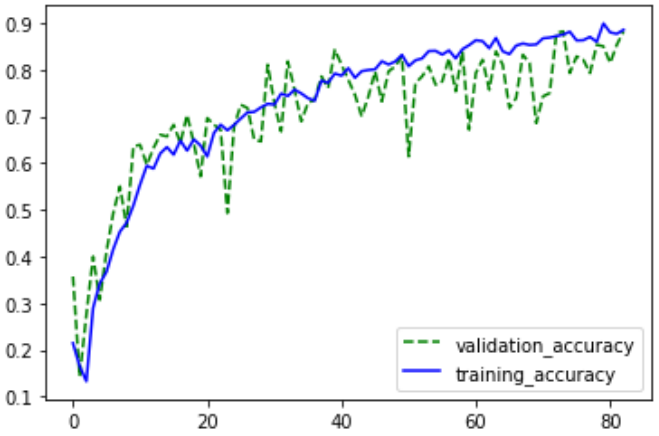

Fig. 3. Training and Validation accuracy

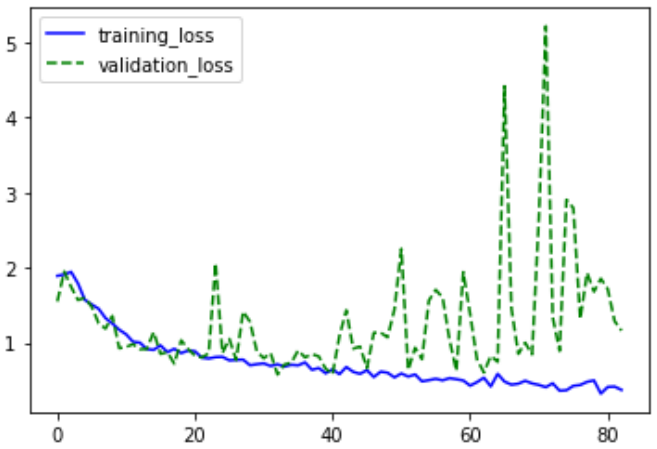

Fig. 4. Training and Validation losses

2) Evaluation metrics

	precision	recall	f1-score	support
0	0.93	0.93	0.93	40
1	0.91	0.97	0.94	40
2	0.76	0.70	0.73	40
3	0.59	0.75	0.66	40
4	0.83	0.72	0.77	40
5	1.00	1.00	1.00	40
6	0.76	0.65	0.70	40
accuracy			0.82	280
macro avg	0.82	0.82	0.82	280
weighted avg	0.82	0.82	0.82	280

Table 1. Precision, Recall, f1-score on the Rose Labs validation dataset

V. Conclusion and future scope

This work presents an automated solution for monitoring in public areas for any medical anomaly detection. The accuracy of the model is approximately ~80% on Rose Labs dataset . The novelty of our work lies in using Transformer architecture in the field of computer vision for sequence classification in a novel approach. This is a unique approach which no other work has put forward in the field of computer vision. This work is highly scalable in terms of expanding the set of action classes. Future scope for extending this work is to scale up this model for a large number of actions set and to implement in real-time applications.

REFERENCES

[1] Distribution-Aware Coordinate Representation for HumanPoseEstimation (Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, Ce Zhu)

[2] Epipolar Transformers (Yihui He, Rui Yan, Katerina Fragkiadak)

[3] Transformers in Vision: A Survey(Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir,Fahad Shahbaz Khan, and Mubarak Shah)

[4] Stacked Hourglass Networks for Human Pose Estimation Alejandro Newell, Kaiyu Yang, and Jia Deng University of Michigan, Ann Arbor

[5] Deep High-Resolution Representation Learning for Human Pose Estimation (Ke Sun Bin Xiao Dong Liu Jingdong Wang)

[6] Mask R-CNN (Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick)

[7] PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model (George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, Kevin Murphy)

[8] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In ICML, 2019.

[9] Rethinking on Multi-Stage Networks for Human Pose Estimation (Wenbo Li1, Zhicheng Wang Binyi Yin1 Qixiang Peng1 Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, Jian Sun)

[10] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In The IEEE International Conference on Computer Vision (ICCV), volume 2, 2017

[11] L. Ke, M.-C. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. arXiv preprint arXiv:1803.09894, 2018

[12] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017

[13] Skeleton-Based Gesture Recognition Using Several Fully Connected Layers with Path Signature Features and Temporal Transformer Module

[14] Human Pose Estimation via Improved ResNet-50 (Xiao Xiao, Wanggen Wand)

[15] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. CoRR, abs/1702.07432, 2017

[16] Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network Alex Sherstinsky

[17] End-to-End Object Detection with Transformers Nicolas Carion , Francisco Massa , Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko

[18] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale Alexey (Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby)