Healthcare monitoring system based on Deep Learning

Rahul Singhal, Himanshu Joshi, Pranjal Chandel
Supervisor: Prof. Satish Chandra
Department of Computer Science and Engineering
Jaypee Institute of Information Technology
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, NOIDA, INDIA

INTRODUCTION

Nowadays, healthcare problems among elders have been increasing at an unprecedented rate and every year, more than a quarter of the elderly people face weakening injuries such as unexpected falls resulting in broken bones and serious injuries in some cases. Sometimes, these injuries may go unnoticed, and the resulting health consequences can have a considerable negative impact on quality of life. Constant surveillance by trained professionals is impossible owing to the expense, effort, and limits it imposes on an individual's living quarters. These concerns prompted the creation of a remote health monitoring system that addresses difficulties such as tracking patient vitals and detecting high-risk incidents.

Remote health monitoring allows healthcare institutions, such as hospitals or assisted living homes, to keep a check on patient health from afar by delivering pertinent data to healthcare practitioners automatically.

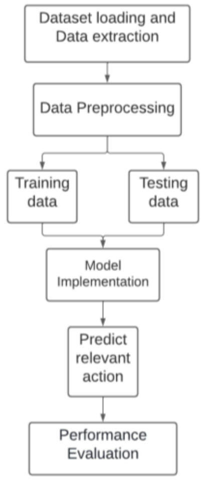
OBJECTIVE

To use nine medically relevant activities to train state of the art deep neural networks for action identification in patients. A video clip of a human action is fed into our neural network, and the output is a prediction of the action class (sneeze/cough, chest discomfort, falling down, back pain, headache, staggering and so on).

DATASET DESCRIPTION

The Rapid-Rich Object Research Lab (ROSE) at Nanyang Technological University, Singapore, granted us permission to utilize the NTU RGB+D dataset for the proposed healthcare application. This dataset contains 60 action classes and 56,880 video samples. We concentrate on a subset of 9 healthcare related action types. Sneezing/coughing, staggering, falling down, headache, chest pain, back pain, neck pain, nausea/vomiting, and fanning self are the action classes we focus on. We extracted videos according to our need from the NTU RGB+D dataset.

PROPOSED METHODOLOGY



Data pre-processing : All films are broken into individual jpg frames and scaled down from 1920x1080 to 427x240 pixels as part of the pre-processing. Each video sample is 3 channels (2 x 240 pixels x 427 pixels) x N frames, where N is the number of frames per movie, which might range from 33 to 222. The dataset is pre-processed in both geographical and temporal dimensions. To eliminate superfluous background information, all frames of the video are center-cropped to 240 × 240 pixels for spatial pre-processing. Cropping does not result in the elimination of the target topic, hence the scales are determined by manually evaluating 100 video samples. We use bilinear interpolation to spatially enlarge the generated pictures to 112 by 112 pixels after cropping. Finally, the input is normalized for each color channel using the mean and standard deviation values from the ActivityNet dataset, and each sample is horizontally flipped with 50% probability during training. After that, we use temporal pre-processing to minimize the number of frames from 33 to 16 in the next step. Since the action most distinctly occurs in the middle of the video, we pick 32 frames from the center of each video clip and then down-sample by 2 to generate the 16 frames. Looking at a 32-frame window is critical to capturing the full content of longer actions, such as staggering and falling down. The final size of the pre-processed video is 3 channels x 16 frames x 112 pixels x 112 pixels.

Model Implementation: The following state-of-the-art models were implemented:

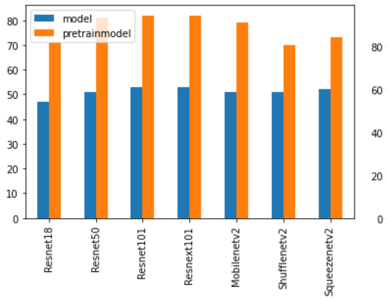
- C3D (based on 3D Convolution Neural networks)
- Resnet Models 3D
- Three dimensional Efficient CNNs (MobileNetV2, ShuffleNetV2 and SqueezeNetV2)

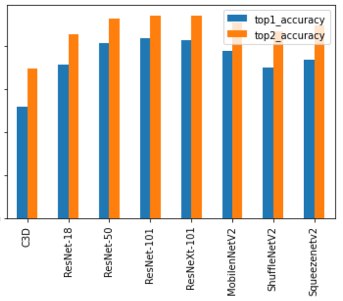
Training and Validation details: Categorical cross-entropy loss with stochastic gradient descent was employed for training the model. To optimize memory consumption and learning speed, we selected a batches containing sixty four videos. We limit the number of training epochs to 50 due to GPU constraints and the vast number of scheduled operations, although more epochs may increase performance. We played about with learning rates, settling on 0.001 for the initial forty epochs only until validation loss reaches the saturation point, then reducing it to 0.0001 for the final ten. We perform a center cropping to generate a squared picture, rescale the image to 112 × 112 pixels, normalize and spatially resize the images to retrieve the 32 center frames.

Risk Analysis and Mitigation: Because the method handles data files of enormous size, it takes a long time to run, and it's not unreasonable to label it sluggish. To operate efficiently, the model demands powerful GPUs. The collecting of data files is an issue that arises since there are many different datasets accessible, each with its own set of outcomes, making it necessary to test for the optimal use case dataset. Data pre-processing and duplicate data removal are two key issues that arise while doing action recognition. Some of the nine actions are quite comparable and have slight differences. Categories of headache, chest problem, and neck pain change merely in hand location. The three most important standards on the NTU-RGB dataset mix pose or raw depth and RGB photos to increase performance. Since such pose data is not always available, we restrict our methods to depend on a single modality. This places further constraints on training and deployment.

RESULTS

In the figure below, we see that when pre-trained algorithms are used, the results show a significant improvement. It is observed that ResNet50's accuracy has increased by 30%, while ResNet101's efficiency has increased by 29% which is a substantial gain. The second graph shows the top 1 and top 2 accuracies produced by all models. When the model depth grows from Resnet18 to Resnet101, the accuracy of ResNet 3D rises by 12.5 percent, demonstrating that deeper Convolutional Neural networks fare better. The highest accuracy was reached by Resnet 101 (84%) and Resnext 101(83%), whereas C3D had poor accuracy owing to the absence of pre-training. In contrast to ShuffleNetV2 and SqueezeNet, MobileNetV2 contains reversed residual blocks, which thrive at collecting dynamic movements, resulting in greater efficiency. In comparison to ResNet-101, the analysis reveals that MobileNetV2 and SqueezeNet are superior candidates for programs that require a light-weight neural network since both deliver reduced complexity for a six percent drop in performance.





We used Precision, Recall and Confusion Matrix as performance evaluation metrics. Neck pain and headache activities are quite similar, and as can be seen from the table below, the model mistakes the two and hence provides the least accurate results in both. Falling down, staggering, and fanning one's self, on the other hand, have unique characteristics, and the model seldom classifies them incorrectly. In situations like these, recall is crucial since the cost of a false negative is expensive and might result in the medical issue not being recognized. In order to prevent ignoring an unpleasant health-related incident, having a low recall is also important. Falling down/staggering, which is more concerning, has a 99 percent recall rate, but chest discomfort has the lowest (72 percent).

Activity	Precision	Recall	F1-Score
Sneeze/Cough	0.79	0.76	0.78
Staggering	0.98	0.98	0.98
Falling down	0.99	0.99	0.99
Headache	0.70	0.64	0.67
Chest Pain	0.76	0.72	0.74
Back Pain	0.80	0.79	0.79
Neck pain	0.66	0.81	0.73
Nausea/ Vomiting	0.90	0.86	0.88
Fan Self	0.98	1.00	0.99

CONCLUSION

In this project, we employed Activity Recognition to aid elder folks in making their lives easier and simpler. Since there is only a slight change in some of the activity videos, we planned to produce good results using state of the art deep learning architectures such as three dimensional CNNs, 3D Resnet architectures and three dimensional Efficient 3D CNNs. Significant data preprocessing was done to extract maximum information possible from videos. It was observed that Resnet101 gave the best overall Top 1 accuracy of 84.3% and Top 2 accuracy of 94.5%. It also gave the best precision, recall and F1-Score in all action categories in comparison to other deep learning models. Our proposed solution will act as a healthcare monitoring system as keeping a trained caretaker may be costly and requires extra human effort. This will help us in detecting healthcare problems which affect elderly people beforehand. Current work in this area has a limitation that detection gets complicated when a person conducts many activities at the same time. Future study in this field might involve extracting additional features from videos utilizing real-time data rather than static datasets and systems with great processing capacity. To capture additional data, several sensors might be fitted while collecting data. Magnetic induction may also be used in conjunction with deep learning to recognize a wide spectrum of emotions. This might aid in improving our system's performance.

REFERENCES

[1] Cho, S., Lim, S., Byun, H., Park, H., & Kwak, S. (2011). Human interaction recognition in YouTube videos. 2011 8th International Conference on Information, Communications & Signal Processing, 1-5.

[2] Ke, S., Hoang, L.U., Lee, Y., Hwang, J., Yoo, J., & Choi, K. (2013). A Review on Video-Based Human Activity Recognition. Comput., 2, 88-131.

[3] Cheema, S., Eweiwi, A., & Bauckhage, C. (2012). Who is doing what? Simultaneous recognition of actions and actors. 2012 19th IEEE International Conference on Image Processing, 749-752.

[4] Kong, Y., & Jia, Y. (2012). A Hierarchical Model for Human Interaction Recognition. 2012 IEEE International Conference on Multimedia and Expo, 1-6.

[5] Kobayashi, I., & Noumi, M. (2010). A study on verbalizing human behaviors in a video. 2010 World Automation Congress, 1-6.

[6] Ihaddadene, N., & Djeraba, C. (2008). Real-time crowd motion analysis. 2008 19th International Conference on Pattern Recognition, 1-4.

[7] Singh, V.K., & Nevatia, R. (2011). Simultaneous tracking and action recognition for single actor human actions. The Visual Computer, 27, 1115-1123.

[8] Ryoo, M.S. (2011). Human activity prediction: Early recognition of ongoing activities from streaming videos. 2011 International Conference on Computer Vision, 1036-1043.

[9] Ke, S., Hoang, L.U., Lee, Y., Hwang, J., Yoo, J., & Choi, K. (2013). A Review on Video-Based Human Activity Recognition. Comput., 2, 88-131.

[10] Köpüklü, O., Kose, N., Gunduz, A., & Rigoll, G. (2019). Resource Efficient 3D Convolutional Neural Networks. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 1910-1919.

[11] Hara, K., Kataoka, H., & Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6546-6555.

[12] Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 4489-4497.

[13] https://rose1.ntu.edu.sg/dataset/actionRecognition/