4th International Conference on New Frontiers of Engineering, Management, Social Science and Humanities
(IEI, Chandigarh) Institution of Engineers, India, Chandigarh
12th-13th May 2019 www.conferenceworld.in
Conference World
ISBN : 978-93-87793-88-0

Deep learning based an ensembled approach for human activity recognition

Ankush Manocha
Lovely Professional University
Jalandhar-Delhi GT Road

Ramandeep Singh
Lovely Professional University
Jalandhar-Delhi GT Road

ABSTRACT— *Computer-vision is the advanced and innovative field for medicinal services, and intelligent monitoring systems are becoming part of healthcare system rapidly. The mechanization in the identification of ordinary or unpredictable activities of a patient can improve the wellbeing results and can likewise reduce the endeavors of manual checkups and monitoring. In this examination, the work centers around different health focused exercises of a patient to figure the range up to a variation from the norm scale. The proposed solution is legitimately reliant on the recordings acquired from the surrounding cameras to identify the physical developments. Present day techniques are totally reliant on the amount and nature of the information used to prepare the system to create an intense reaction to the recent development. To deal with these issues, a deep learning based classification strategy is proposed to address different physical variations from the norm which can prompt a reason for health affliction. A subset of NTU RGB+D dataset with health focused models are utilized to train the system. In the proposed study, 3D CNN model is utilized to extricate the highlights from the recordings and LSTM model is utilized to classify the activities. The proposed monitoring system has additionally used the qualities of different state-of-the-art models and the results are processed by computing the parameters of review, exactness, precision, and F-measure. Tests hold the proof to legitimize the utility of 3D CNN model for posture classification and LSTM model for activity prediction in our proposed framework.*

1. Introduction

Computer Vision is considered as one of the real headway in reconnaissance and gives numerous successful solutions for different domains like remote sensing, public surveillance, self-governing transportation, healthcare, individual assessment and numerous others. Deep learning strategies can likewise be utilized to expand the dynamic effect of computer vision technology continuously. In the area of appraisal, this developing technology has changed over the way of manual checking to autonomous monitoring by keeping up the sensitiveness of the domain through smart-decisions to provide medicinal or assistive-care based services.

2. Related Work

2.1 *Activity detection using handcraft techniques*: The major issue in surveillance systems is the real-time detection and recognition of abnormal events caused by the execution of irregular exercises. Dedeoglu et al. (2006) presented an approach that extricates outlines of the objects by utilizing an adaptive background-subtraction for spatial area identification of the object. Saykol et al. (2010) proposed a methodology which focuses around the labeled frames to classify the occasion depends on the performed activity. 34 proposed a system that performs situation-based query processing for the surveillance system. These models turned into a valuable system for offline movement examination.

2.2 *Activity detection using deep learning:* From 2014, numerous models have been proposed and the majority of the systems are directly taking images from raw data to extract features without performing any normalization (Ji et al. (2013), Ioffe and Szegedy (2015)). Deep Learning based Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) methods have appeared most to achieve the magnificent results for video-based data classification (LeCun et al. (2015)). Now, researchers have used the vitality of deep learning to recognize the human activities from the open or closed areas. Deep learning based approaches are edge-toedge trainable and can also perform their operations directly to the raw data (Lin et al. (2016)). Karpathy et al. (2014) clarify various edge level combinations in CNN for image classification. Peng and Schmid (2016) proposed a system named Two-stream R-CNNs model to recognize the activities, where Spatial Based Region Proposal Network and Movement-Based Region Proposal Network are utilized to predict feature level activity. This approach mainly used to connect the activities to characterize the spatial highlights of a video independently by two-stream CNN. Neverova et al. (2016a) utilized RNN and CNN method to capture the identity of a person from the activity patterns captured by the sensors like accelerometers and gyroscope from cell phones. A gesture localization based multi-modular deep learning strategy is also described by (Neverova et al. (2016b)). Du et al. (2015) introduced a novel architecture depends on the Tanh-Bidirectional Recurrent Neural Networks (Tanh-BRNN). In the architecture of Tanh-BRNN, LSTM module is joined with the fully-connected (FC) layer by utilizing the skeletons as sources of information. The primary proposition which utilized CNN+LSTM based architecture for movement recognition from a video was presented by Donahue et al. (2015). Ko and Sim (2018) introduced a combined model of CNN and LSTM to detect the abnormal behavior from an RGB image. Nu´nez et al. (2018) introduced a two-step training model to recognize the poses of the human body with hand gestures using skeleton based 3D data sequences. Khaire et al. (2018) presented a novel approach based on the CNN network to recognize human activities with the combination of multiple vision cues. Therefore, the

computer-vision based smart monitoring system is explained in further sections to handle the health adversity of the patient in real-time.

3. Proposed System

In smart monitoring, visual sensors can retrieve the physical exercises from the nearby condition of the patient. The architecture of the proposed framework likewise partitioned into stages. The major point of the proposed framework is to monitor patients consistently to give a top to bottom consideration utilizing computer vision technology. The proposed checking approach is divided into four stages.

3.1 Data Generation

Data Generation stage is capable to retrieve the information from visual sensors about the physical exercises of the patient straightforwardly. The action related information is captured continually from the wired visual sensors installed at the area of the patient. The way toward screening and transmitting the frames related to complex activities of the patients is done by visual sensors in a proficient way. Physiological information of the patient is gathered in the grouping of frames which characterizes the presence of the subject with the kind of an activity. Hence, the information is normalized into a satisfactory organization by performing different information expansion methods before sending it for further analysis.

3.2 Data Normalization

In stage 2, different information growth strategies and handcraft highlights are utilized to standardize the information before offering contribution to the proposed classification. Two primary reasons are considered for information handling and to remove the additional substance from the frame: First, to keep up the security of the patient which is considered as the most sensitive imperative in video based monitoring system and second to diminish the processing cost of the model. As we probably aware foundation likewise influences the exercises of an individual yet in our framework, we concentrate the patient in their limited observing region where background always stay static.

3.3 Abnormal Activity Classification

Abnormality Classification: The proposed methodology is structured by joining the few layers of Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) model for feature extraction and characterization of the activities in a constant way. For the most part, the more raised features are acquired from the last layers of the CNN network, and the lower-level features are created by the upper layers of the network (LeCun et al. (2015)). In the first place, we prepared the CNN architecture freely by associating with the two Fully Connected (FC) layers to characterize the body posture and after that the element vectors identified with body posture are exchanged to the LSTM model to decide the kind of a movement.

3.3.1 3D Convolutional Neural Network (3D CNN):

3D CNN (Ji et al. (2013)) architecture is considered as one of the best deep learning methods for spatial and temporal features extraction to classify the event and activity. In our system, the 3D-CNN is used to process the segment of frames to differentiate the body posture of the patient in two states, either normal or anomalous. To extract the features from the abnormal video templates, total 10 convolutional layers are used with the number of filters are 32, 64, 2 × 128, 2 × 256, 2 × 512 and 2 × 512 as shown in Fig.1. Total 6 pooling layers are used to reduce the size of the input sample and max-pooling function is applied to the region of the frames to down-sample the input representations. The size of the kernel in the first pooling layer is set to 1 × 2 × 2 and 2 × 2 × 2 for other pooling layers to extract the temporal feature from the series of frames. 3D defines the three dimensions of the data. The height and width of the frames are considered as two dimensions which are used to calculate the spatial feature and number of frames in a segment is considered as the third dimension which defines the temporal feature of the subject to provides a better accuracy in activity prediction. To generate feature maps from the segment of frames, 3D kernels are convolved to all the dimensions of the frames with stride = 1. To detect the health-afflictions, the subset of NTU RGB+D dataset is used to train 3D CNN model. The NTU RGB+D dataset is considered one of the largest activity dataset related to normal and abnormal activity templates. The dataset is further explained in Section 4. The size of each frame is converted from 1920×1080 to 320×240 and then the video is randomly clipped with the length of 30 frames. The pose vectors are generated in the form of sequence for each frame of the video. Each vector contains a series of numbers which represent the features of an activity with respect to its activity class. To analyze the motion of the subject, the sequence of feature vectors would be recognized in sequence. The Long Short-Term Memory (LSTM) has the capability to process the data in a sequential order.

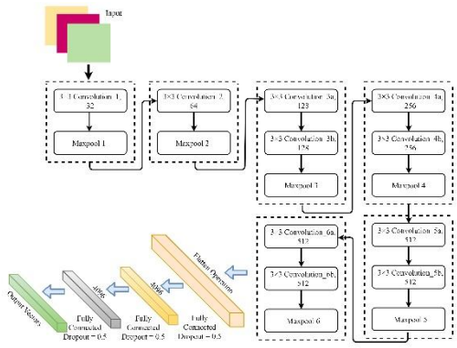

Fig 1: The 3-D-CNN model with 10 convolutional layers, total 6 max pool layers, 2 FC layers and 1 softmax layer to generate the output vectors

3.3.2 Long Short-Term Memory

RNN become the best solution for the video-based activity classification process, but a common issue of vanishing gradient usually occurs during the training process of the model. LSTM is essentially considered as the best solution to solve the issue of vanishing gradient in RNN. Gated cells of the LSTM are used to deal with the problem of vanishing gradient (Ijjina and Chalavadi (2017)). These gated cells are capable to hold some amount of resulted values in its neurons. More information about the LSTM can be found in (Hochreiter and Schmidhuber (1997)). Fig 2 describes the LSTM network architecture of the proposed methodology. In the LSTM network, the features maps generated by the CNN is transferred to the LSTM network at a particular time instance. Each LSTM cell generates the result by combining the state and the output vector of the cell. The cell state and output vectors are used to propagate the LSTM network at the next time instance. LSTM cell state operation is described in Fig. 3.

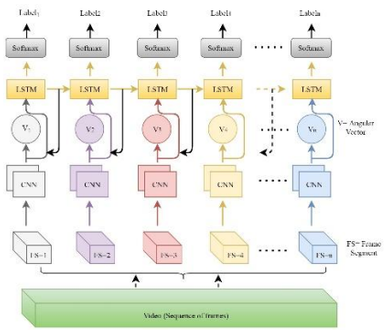

Fig 2: The architecture of LSTM network

1. The final feature vector from the 3D CNN is received by each cell of the LSTM network, and each vector is denoted as xt.
2. The previously calculated state of the cell Ct−1 is received at time t-1, and the current cell output is generated by ht.
3. The cell state updation process is based on the current inputs of the cell.

As described in figure 3, it = input gate, ft = forget gate, ot = output gate, ht−1 = hidden state and ct = cell state. The current input is selected by gt and update the memory of the cell. To generate the activity result, LSTM generates the activity labels related to its activity class. This estimation is generated based on the combination of the current input at a specific time instance with the previous cell state and previous cell output. The proposed method recognizes real-time activities of the patient from the given video template as the input and generates the estimation about the activity class without calculating the entire frame segment. Even the motions are observed at the input level which are completely uncertain, thus the proposed methodology filters all the uncertain motions by the combination of 3D-CNN and LSTM network.

3.3.3 Training Phase of the System

The training process of the neural network is quite complex because of the input variation in each layer during training period (Ioffe and Szegedy (2015)). The proposed model is evaluated by setting up the several parameters of the model. In the LSTM model, 250 hidden units are used with the learning rate of 0.001. L2-normalization is used for weight optimization. To train the model on the training dataset, the model takes around 80 to 150 minutes. The training session is divided into two stages. In the proposed classification model, we connect the last layer of 3D CNN with two fully-connected layers to determine the pose of the body. After finishing the training process of the CNN model, we transfer the feature vectors generated by the last layer of the CNN model to the LSTM network to train the model. By following this training procedure, we trained our entire classification model in two stages in an efficient manner and it also takes less time to train. The initial learning rate of the model is set to 0.001 which is

gradually reduced to 0.01 when no improvement has been observed. As we have less amount of data to train the system, it may prompt the problem of system overfitting. To handle the issue of overfitting, ”dropout” function with the size of 0.5 is applied to all non-recurrent connections.

3.4 Smart notification generation

The proposed activity classification procedure calculates the variation from the norm status of the patient. To ascertain the patient physical state, the anomalous action templates are utilized to train and test the system. If the value generated by the classification model lies to any activity class on which the system is trained, the health status of the patient is considered as unsteady and a prompt action needed to perform by the caretaker or doctor.

4. Experimental Setup and Performance Evaluation

The deployment of the system is motivated by the real-world scenarios to create an experimental environment. By taking the informal consent from the caretaker and doctor, the patient was monitored using the solitary visual sensor for two weeks for every sixty minutes. To capture various activity templates, distinctive activities are considered. To enhance the proficiency of the system we additionally utilized data augmentation technique (Dyk and Meng (2001)). Less amount of data can be a cause of overfitting problem. To augment the data multiple operations like rotation, flip, crop are performed on a single frame. Along with the captured activities of the patient, various health-oriented activities from the NTU RGB+D dataset is also selected to train the model. NTU RGB+D (Shahroudy et al. (2016)) dataset is considered as an efficient 3-D activity recognition dataset with both normal and abnormal activities. The dataset contains total 56,880 activity layouts. The activity templates are isolated into 60 distinctive activity classes. The dataset contains 60 diverse activity classes including 40 every day, 9 health-related, and 11 common activities. In this system, the health-oriented anomalous exercises are considered to train the proposed methodology. The total length for each activity which is used to train the model as described in Fig. 3. The presented system contains three noteworthy steps to evaluate the performance of the system.

1. Regulate the effectiveness of the proposed methodology for abnormality recognition.
2. Statistical determination of the effectiveness of the alert generation mechanism.
3. The complete performance analysis of the proposed system.

Fig 3: Type of Activities with combined length of the video templates (in minutes)

4.1 *Activity classification efficiency*: The efficiency of activity classification determines the effectiveness of the proposed methodology to classify the activities according to the trained activity classes. The procedure of activity classification is explained as follows:

1. In the first step, 3D CNN model extract the features to calculate the stance of the patient.
2. In the second step, LSTM calculate the feature vector in sequence to calculate the abnormality scale of the activity.
3. Complete classification methodology calculates the final activity probability. For different activity templates, each activity classification precision is more than 80%.

Table 1 demonstrates a confusion matrix resulted on 9 long-term abnormal activity classes. As shown in the confusion matrix, activities including falling, straggling, feeling warm and nausea display the best results with 100%, 93%, 97% and 94% respectively. For back pain (88%), neck pain (84%) and headache (85%), the model achieve less accurate results. The experimental outcomes demonstrate that sometimes the activities like a headache, neck pain are miss-classified. The main two reasons are considered for miss-classification of these activities, first, similar body motion (e.g., headache and neck pain have the almost same “head” motion) and sometimes even for humans, it is hard to distinguish whether the patient is suffering from a headache, neck pain or performing any other similar activity. Second, less amount of training samples. Different statistical estimations are also used to calculate the effectiveness of the classification of the proposed methodology such as Accuracy, F-measure, Precision and Recall.

Table 1: The activity prediction probability for each trained class. The content of the rows represents the ground truth, while the elements of the columns represents the predicted activity result.

	A	B	C	D	E	F	G	H	I
A	90	7	0	0	0	0	0	0	0
B	0	93	0	0	0	0	0	8	0
C	0	0	100	0	0	0	0	0	0
D	0	0	0	85	0	7	0	0	0
E	0	0	0	5	91	0	0	0	0
F	0	5	0	0	0	88	9	0	6
G	9	0	0	0	0	0	84	0	0
H	0	0	0	0	8	0	0	94	0
I	0	0	0	0	0	0	4	0	97

Table 2 portrays the statistical outcomes acquired from the proposed system over the whole monitoring method. The system is achieving better accuracy to recognize various physical adversities, numerating to 91.33%. In the case of precision, the system is efficient to generate the higher value of 92.52%. Similarly, the proposed deep neural model yields 89.28% and 91.27% value respectively for recall and F-measure.

Table 2: Comparison result of accurate detection of abnormal activities

Serial Number	Parameter	Values(in %)
1.	Accuracy	91.33
2.	Precision	92.52
3.	Recall	89.28
4.	F-measure	91.27
5.	False positive ratio	2.59

4.2 Temporal regulation for alarm generation

Temporal regulation manages the viability of the system to generate alarms for concerned caretakers and doctor to handle the critical events and circumstances. The deep learning based approach is used to decide such sort of proficiency. Delay in alarm generation to the concerned guardian is considered as an imperative parameter in this proposed system. This alert resulted in the immediate intervention of the concerned specialists or guardians for giving essential health-care services. Results are obtained by the doing the cross-examination of several video templates of datasets for a specific activity.

False alarm ratio: The fundamental objective of statistical analysis of the system is to calculate the overall performance of the system by determining the false positive (FP) alerts based on the total number of generated alerts. A proposed model displays a low rate of False Positive Alert, numerating to 2.59% (Table 2) demonstrating high exactness in activity classification and generation of an alarm.

4.3 State-of-the-art comparisons

This section deals with the activity classification efficiency of the proposed system by comparing the results of the stateof-the-art models on similar activities in a similar environment as described in table 3 so that the utility of the proposed system can be justified experimentally in real-time monitoring environment.

- ST-LSTM+Trust Gate (Liu et al. (2016)). A robust LSTM based gate scheme is proposed to learn the sequence of the input data.
- Two CNNs+Fuzzy (Ijjina and Chalavadi (2017)). This approach is used to recognize the motion sequence by using deep learning model from RGB-D videos.

Table 3: Comparison result of abnormality detection

Activity Name	ST-LSTM+TrustGate	TWO CNNs+Fuzzy	Proposed Model
Sneeze or cough	83%	91%	90%
Straggling	78%	90%	93%
Falling	82%	97%	100%
Headache	75%	86%	85%
Stomachache	86%	87%	91%
Back Pain	72%	84%	88%
Neck Pain	68%	78%	84%
Nausea	81%	89%	84%
Feeling Warm	82%	91%	97%
Mean	**78.56%**	**88.11%**	**91.33%**

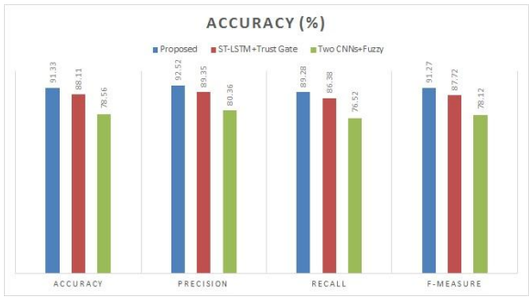

Fig 4: Comparative Accuracy Measurement

The results clearly illustrate that our system obtains the highest mean (91.33%), followed by Two CNNs+Fuzzy (88.11%) and ST-LSTM+Trust Gate (78.56%) for long-term activity recognition. More specifically, the system achieves the best recall for feeling warm (97%), falling (100%), straggling (93%) and stomachache (91%). By contrast, Two CNNs+Fuzzy model performs slightly better on some activities like a sneeze or a cough and headache as compared to our model by getting the result of (91% vs. 90%) and (86% vs. 85%). Different parameters (Fig.4) are also calculated to justify the utility of the proposed system in a real-time monitoring environment.

5. Conclusion

By the indulgence of artificial intelligence in computervision, the activity recognition process can be made more reliable and effective. The main aim of this article is to illustrate the proposed patient monitoring system based on the captured activity patterns by using the advanced deep learning methods. Moreover, this proposed system includes the key factor of the decision-making process by generating an alert in the form of notification to caretaker and doctor of the patient to handle the critical physical state of the patient. The temporal feature describes the most important factor for abnormality recognition. 3D CNN model incorporates to extract the features (spatial and temporal) from the video templates. To improvise the extraction and recognition the motion of the subject from the given input, 3D CNN model is also trained on the 9 health activities of NTU RGB+D dataset. The 3D feature maps generated from 3D CNN model is fed to LSTM layer to recognize the sequence of the anomalous event. By developing a patient monitoring system with an alert mechanism to deliver particular time instance based physical state increases the productivity and utility of the system. Statistical results are quite helpful to decide the performance and the efficiency of the system to handle the critical health state of the patient.

References

Dedeoglu, Y., T ˘ oreyin, B.U., G ¨ ud¨ ukbay, U., C¸ etin, A.E., 2006. Silhouette- ¨ based method for object classification and human action recognition in video, in: European Conference on Computer Vision, pp. 64–77.

Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

Du, Y., Wang, W., Wang, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1110–1118.

Dyk, D.A.V., Meng, X.L., 2001. The art of data augmentation. Journal of Computational and Graphical Statistics 10, 1–50.

Edition, F., American, P.A., et al., 2013. Diagnostic and statistical manual of mental disorders. Arlington: American Psychiatric Publishing .

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Ijjina, E.P., Chalavadi, K.M., 2017. Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognition 72, 504–516.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35, 221–231.

ISBN : 978-93-87793-88-0

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Khaire, P., Kumar, P., Imran, J., 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. Pattern Recognition Letters .

Ko, K.E., Sim, K.B., 2018. Deep convolutional framework for abnormal behavior detection in a smart surveillance system. Engineering Applications of Artificial Intelligence 67, 226–234.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

Lin, L., Wang, K., Zuo, W., Wang, M., Luo, J., Zhang, L., 2016. A deep structured model with radius-margin bound for 3d human activity recognition. International Journal of Computer Vision 118, 256–273.

Liu, J., Shahroudy, A., Xu, D., Wang, G., 2016. Spatio-temporal lstm with trust gates for 3d human action recognition, in: European Conference on Computer Vision, pp. 816–833.

Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., Taylor, G., 2016a. Learning human identity from motion patterns. IEEE Access 4, 1810–1820.

Neverova, N., Wolf, C., Taylor, G., Nebout, F., 2016b. Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 1692–1706.

Nu´nez, J.C., Cabido, R., Pantrigo, J.J., Montemayor, A.S., V ˜ elez, J.F., 2018. ´ Convolutional Neural Networks and Long Short-Term Memory for skeletonbased human activity and hand gesture recognition. Pattern Recognition 76, 80–94.

Peng, X., Schmid, C., 2016. Multi-region two-stream R-CNN for action detection, in: European Conference on Computer Vision, pp. 744–759.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252.

Saykol, E., Bastan, M., Ugur, G., Ulusoy, O., 2010. Keyframe labeling tech- ¨ nique for surveillance event classification. Optical engineering 49, 117203.

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, pp. 1010–1019.

Spasova, V.G., 2014. Experimental evaluation of keypoints detector and descriptor algorithms for indoors person localization. Annual J. Electronics 8, 85–87.

Ye, Y., Ci, S., Katsaggelos, A.K., Liu, Y., Qian, Y., 2013. Wireless Video Surveillance: A Survey. IEEE Access 1, 646–660.